Abstract

Deep neural networks (DNNs) have set state of the art results in many machine learning and NLP tasks. However, we do not have a strong understanding of what DNN models learn. In this paper, we examine learning in DNNs through analysis of their outputs. We compare DNN performance directly to a human population, and use characteristics of individual data points such as difficulty to see how well models perform on easy and hard examples. We investigate how training size and the incorporation of noise affect a DNN’s ability to generalize and learn. Our experiments show that unlike traditional machine learning models (e.g., Naive Bayes, Decision Trees), DNNs exhibit human-like learning properties. As they are trained with more data, they are more able to distinguish between easy and difficult items, and performance on easy items improves at a higher rate than difficult items. We find that different DNN models exhibit different strengths in learning and are robust to noise in training data.

1 Introduction

Artificial neural networks were inspired by a mathematical model of neurons in the human brain (McCulloch and Pitts, 1943). With recent advancements and rapid improvements of DNNs, the number of tasks where DNN models outperform humans continues to grow (Campbell et al., 2002; Ferrucci et al., 2010; Silver et al., 2016; Schaul et al., 2016; Stadie et al., 2015). However, studies have shown that DNNs have the ability to "memorize" training data and therefore easily overfit (Zhang et al., 2016). We do not yet have a good understanding of what underlying information a DNN learns as it is trained, or whether these DNNs mimic human learning behavior.

Because of this, interpretability in DNNs has emerged as an important research area (Kim et al., 2016; Li et al., 2016). If these systems replace human decision makers, there needs to be insight into how decisions are made by these systems. Understanding how and what a network learns can eliminate bias and overfitting in models and help us understand why certain predictions have been made (Lei et al., 2016; Bolukbasi et al., 2016).

Traditionally progress and performance in artificial intelligence is measured at the aggregate scale. Performance improvements on a large data set are used as the benchmark for progress in a particular task. An aggregate score does not consider the characteristics of the data set, or the difficulty of the examples in the data set. Evaluation does not consider which questions are answered correctly and what that says about a model’s ability to generalize and remember what it has learned (Lake et al., 2016). It has recently been shown that traditional measures of learning in Machine Learning (ML) cannot accurately explain the success of DNN models (Zhang et al., 2016). In this paper we compare the performance of different DNN and traditional ML models to the intelligence of a human population and study whether and how they learn “intelligence.”

We use Item Response Theory (IRT) to compare the performance of DNN and non-DNN NLP systems to that of a human population in the task of Recognizing Textual Entailment (RTE). IRT was introduced as a more descriptive metric for evaluating NLP systems (Lalor et al., 2016). IRT models characteristics of individual examples (called “items”) such as difficulty and discriminatory ability to estimate ability as a function of correctly answered items. IRT estimated ability relies on which items in a test set are answered correctly, not just how many, and compares NLP models to
human ability directly. Our results show that although the IRT score is highly correlated with the accuracy score, it has much more discriminatory power in separating NLP performance by different training sets and level of noise. Using the difficulty parameters of items from IRT, we show that as a model is trained with larger training sets, it begins to exhibit these patterns of learning. A regression on the model outputs as a function of training size and item difficulty shows that for very small training sets, response probabilities are random with respect to an item’s difficulty. As training size increases, both easy and difficult items are more likely to be answered correctly. Moreover, as training size increases, an item’s difficulty begins to have a negative effect on the probability of answering an item correctly (that is, easy items become easier than hard items).

To test whether DNNs are simply memorizing, we examine how they learn from noisy data. Noise introduces confusion and leads to performance degradation, which our experiments confirm for the LSTM model. However, NSE and CNN performance is stable up to a substantial amount of noise (35%), supporting the idea that the models are learning, not simply memorizing.

Our contributions are as follows: (i) We first show that IRT is highly correlated with standard evaluation metrics (e.g. accuracy) while providing insight as to how well an NLP model performs as compared to a human population, (ii) we analyze “learning” in the context of NLP models by evaluating model responses qualitatively (with human-generated item groups) and quantitatively (using IRT-estimated item difficulty values), and (iii) we demonstrate the existence of human-like learning patterns in DNNs by analyzing the effect of training data size and noise on performance.

2 Background and Related Work
2.1 Item Response Theory
IRT is a psychometric methodology for scale construction and evaluation. It is widely used in educational testing in the construction, evaluation, or scoring of standardized tests. It is used both for designing tests and analyzing human responses (graded as right or wrong) to a set of questions (called “items”). IRT jointly models an individual’s ability and item characteristics to predict performance (Baker and Kim [2004]).

IRT models make the following assumptions: (i) individuals differ from each other on an unobserved latent trait dimension (called “ability” or “factor”), (ii) the probability of correctly answering an item is a function of the person’s ability, (iii) responses to different items are independent of each other for a given ability level of the person, and (iv) responses from different individuals are independent of each other (Lalor et al. [2016]).

A common IRT model for estimating a single latent trait is the three-parameter logistic model (3PL). For an individual \( j \), item \( i \), and latent ability of individual \( j \theta_j \), then the probability that individual \( j \) answers item \( i \) correctly can be modeled as:

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

where \( a_i, b_i, \) and \( c_i \) are item parameters: \( a_i \) (the slope or discrimination parameter) is related to the steepness of the curve, \( b_i \) (the difficulty parameter) is the level of ability that produces a chance of correct response equal to the average of the upper and lower asymptotes, and \( c_i \) (the guessing parameter) is the lower asymptote of the curve and the probability of guessing correctly. This curve is referred to as the Item Characteristic Curve (ICC) for the given item. A two-parameter logistic (2PL) IRT model assumes that the guessing parameter \( c_i = 0 \).

The ICC for a good item looks like a sigmoid plot, and should have a relatively steep increasing slope between ability levels −3 and 3, where most individuals are located, in order to have appropriate power to differentiate different levels of ability.

2.1.1 Interpreting IRT Results
The IRT ability estimate is with respect to the human population whose responses were used to estimate item parameters. For example, an individual’s (or NLP model’s) estimated ability of 1.2 is interpreted as being 1.2 standard deviations above the mean ability in the population. The traditional total number of correct responses generally does not have such quantitative meaning. This score can be converted into a percentile to indicate what percentage of the population a person’s ability is higher than. For this work we convert all IRT ability estimates to their percentile representations.

2.2 Recognizing Textual Entailment
RTE attempts to classify semantic relationships between two sentences (Dagan et al. [2006]). For text \( T \) and the hypothesis \( H \) sentence pair, \( T \) entails \( H \) if a human that has read \( T \) would infer that
Table 1: Examples of sentence pairs from the IRT data sets, their corresponding label, category, and difficulty as measured by IRT. Refer to §3.3 for descriptions of the sentence categories, and §2.1.1 for how to interpret difficulty values.

| Premise                                                                 | Hypothesis                                      | Label               | Category | Difficulty |
|-------------------------------------------------------------------------|-------------------------------------------------|---------------------|----------|------------|
| People were watching the tournament in the stadium                      | The people are sitting outside on the grass     | Contradiction       | A, D, M  | 0.51       |
| Two girls on a bridge dancing with the city skyline in the background   | The girls are sisters.                          | Neutral             | B, G, M  | -1.92      |
| A little girl eating a sucker                                           | A child eating candy                            | Entailment          | C, D, M  | -2.74      |
| Nine men wearing tuxedos sing                                            | Nine women wearing dresses sing                 | Contradiction       | A, D, L  | 0.08       |
| A man running in front of a vehicle                                     | There is a man running                          | Entailment          | B, J     | -2.45      |
| A young boy in a sweatshirt is doodling on a piece of paper              | The class pictures are on display               | Contradiction       | A, F, M  | 0.78       |
| A couple plays frisbee in a green field with trees in the background    | A couple fixes dinner in their kitchen          | Contradiction       | A, G, M  | -0.82      |
| Two girls on a bridge dancing with the city skyline in the background   | The girl is going to find out what is under the wrapping paper | Entailment          | A, M     | -2.69      |
| A man is popping balloons inside a cubicle using a homemade tool        | A man is popping balloons inside a cubicle using a handmade knife | Neutral            | A, H, L  | -0.43      |
| A scene of snow and water                                               | A snow and water scene at sunset                | Neutral             | C, I, L  | -1.01      |
| Bike rider jumps off of gray stone rise, does trick in air              | A person doing a trick                          | Entailment          | B, K     | -2.07      |

H is true. If a human would infer that H is false, then H contradicts T. If the two sentences are unrelated, then the pair are said to be neutral. Table 1 shows examples of T-H pairs and their respective classifications. In terms of IRT, each sentence pair is considered a single item.

Recently deep learning models have set and surpassed state of the art results in RTE (Rocktäschel et al., 2016; Munkhdalai and Yu, 2016), in part due to the availability of the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015). SNLI consists of over 550K human-annotated sentence pairs and is significantly larger than previously used data sets.

### 2.3 Related Work

Lalor et al. (2016) introduced the idea of applying IRT evaluation to NLP tasks. They built a set of scales using IRT for RTE and evaluated a single LSTM neural network to demonstrate the effectiveness of the evaluation, but did not evaluate other NLP models or tasks.

Martínez-Plumed et al. (2016) consider IRT in the context of evaluating ML models, but they do not use a human population to calibrate the models. They attempt to fit IRT models from the ML models without identifying good items initially from a human population, and obtain ICCs with negative slopes that are difficult to interpret using the existing IRT assumptions.

Bruce and Wiebe (1999) modeled latent traits of data points to identify a correct label. There has also been work in modeling individuals to identify poor annotators Hovy et al. (2013), but neither jointly model the ability of individuals and data points, or apply the resulting metrics to NLP models. Passonneau and Carpenter (2014) model the probability a label is correct along with the probability of an annotator to label an item correctly according to the Dawid and Skene (1979) model, but do not consider difficulty or discriminatory ability of the data points.

### 3 Experiments

#### 3.1 Model Selection

For our experiments we test four traditional ML models: Naive Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), and three representative DNN models. For the traditional ML models, we used a baseline implementation from a popular ML framework for consistency Pedregosa et al. (2011). As our goal is not to extend the state of the art but rather to compare the learning behavior, we used default parameters in all cases.

We tested DNN networks that have previously been shown to perform well on the SNLI data set or other NLP tasks: A Convolutional neural network (CNN) for sentence classification (Kim 2014); The LSTM neural network model released with the SNLI data set (Bowman et al. 2015); Neural Semantic Encoder (NSE), a memory-augmented RNN (Munkhdalai and Yu 2016).

1For model details please refer to the original papers.
To test whether DNNs exhibit traditional characteristics of “learning,” we trained the models on a number of variations of the SNLI training set. Our motivation is to observe how performance changes when more or fewer training examples are provided, or when noise is incorporated into the data set. Each model was trained on several modifications of the SNLI training set:

- The original 550K SNLI training set.
- A noisy version of the training set, where a random selection of sentence pairs had the correct label replaced with an incorrect label (5%, 20%, 35%, and 50% noise).
- A randomly selected subset of the training set (100, 1000, 2000, 10,000, 50,000, 100,000, and 200,000 examples).

These trained models were tested both on the original 10K example SNLI test set and on the IRT test sets in order to compare performance.

### 3.2 IRT Models and Test Sets

In order to model individual item characteristics and evaluate our models using IRT we fit a set of IRT models. For data we obtained the human response patterns and IRT models from Lalor et al. (2016). The data consists of approximately 1000 human annotator responses for a selection of the SNLI data set (Bowman et al., 2015). 5 IRT models (and subsequent test sets) were fit by Lalor et al. (2016), based on the inter-annotator agreement from the original SNLI data set: (i) entailment items with 5 of 5 agreement (5E), (ii) contradiction items with 5 of 5 agreement (5C), (iii) neutral items with 5 of 5 agreement (5N), (iv) contradiction items with 4 of 5 agreement (4C), and (v) neutral items with 4 of 5 agreement (4N).

The IRT models estimate difficulty, discriminating power, and guessing attributes for each item in the data set, and allow us to model “ability” (i.e. the performance of a model or individual on the task of classifying a particular label). The items used to fit each IRT model act as a test set to measure ability for that group. For NLP models that have been trained on a particular training set, the model output when tested on the IRT items is treated as a new response pattern. This response pattern is used to generate a theta score for the model, which we convert to percentiles (§2.1.1) to represent the percentage of the human population the model outperformed.

### 3.3 Dataset Characteristics

For a qualitative analysis of performance, we attempted to understand how the models perform with regards to certain semantic and syntactic characteristics observed in the data. We manually classified the items in the IRT test sets to determine if performance varied across items based on the properties of the sentence pairs. We classified the IRT test set items into the following groups:

- A) P & H both are sentences
- B) One sentence and one noun phrase
- C) P & H are both noun phrases

and one or more of the following 9 groups:

- D) Aligned lexical entailment or contradiction
- E) The difference between H and P is the omission of a word or phrase
- F) H is not relevant at all to P
- G) P & H have related subjects, unrelated objects
- H) Restrictive adjective or adverb is added to H
- I) Restrictive prepositional phrase is added to H
- J) H is a simplification of P
- K) H is a summarization of P
- L) P & H are describing semantically different objects
- M) Comparing P & H requires higher-level semantic inference

Table 1 shows examples from the data set for each group. Table 1 also includes the difficulty parameter for the examples. Most examples are “easy” in that an individual with mean ability (e.g. theta score of 0) would have a probability greater than 0.5 of answering the question correctly, which is appropriate for analyzing NLP models (Lalor et al., 2016).

### 4 Results

#### 4.1 Model Performance

Table 2 shows the results of evaluating the models listed above when trained on the SNLI training set. We report theta percentiles on the IRT test sets and accuracy on the SNLI test set as a baseline comparison of models. Table 2 shows that the DNN
models, which had the highest accuracy comparing to other ML models, also score higher in terms of IRT. Our results show that the RNN models (LSTM and NSE) outperform the CNN model as in both accuracy and IRT. However the IRT scores are much more sensitive in terms of identifying high performing RNN models. The NSE model, which had high overall accuracy (84.06%), scores lower when comparing with the LSTM model.

Because IRT considers the individual items that are answered correctly, it may be the case that the specific response pattern for the NSE output is associated with a lower ability score than the LSTM response pattern. The overall accuracy score does not consider the difficulty and discriminatory characteristics of individual items, and therefore cannot distinguish models on the basis of "ability." When you consider responses to the IRT test set items, the LSTM and NSE models answer a similar number of items correctly (in fact, the NSE model answers one more question correctly), but the specific items that are answered correctly affect the ability scores from the IRT models. In addition, there is a larger gap between IRT scores than accuracy scores, which helps when identifying high-performing models and distinguishing between randomness in performance scores.

To confirm the validity of IRT as a metric we also calculated the correlations between accuracy and the IRT scores to see how consistent the scores are. The IRT percentile scores are highly correlated with accuracy (above 0.9 for 5C, 4C, and 4N, above 0.7 for 5E and 5N). In addition, each IRT score is highly correlated with each other (above 0.89 for all except 4C, above 0.75 for 4C). The strong positive correlation between IRT scores and accuracy shows that IRT as an evaluation metric is consistent with existing metrics. In addition IRT provides more information in terms of comparing the NLP models against human performance.

### 4.2 Analysis of Learning in NLP Models

We next conduct an analysis of the performance of the different NLP models when trained with different training sets. We first analyze performance qualitatively, according to the groups described in §3.3. We then perform more quantitative analyses of model performance: (i) performance as a function of noise in the training set, (ii) performance as a function of training set size, and (iii) performance on easy and difficult test set items.

#### 4.2.1 Learning From Noisy Data

We attempt to understand learning by analyzing the output of models that are trained with different levels of noise incorporated into the training set. To do this, we randomly select a section of our training set (5%, 20%, 35%, or 50%) and replace the correct label with an incorrect one. Figure 1 shows how the DNN and non-DNN models perform when trained with noisy data. As expected, when noise is added to the training data overall accuracy drops in most DNN and non-DNN NLP models.

Because IRT scores are low to begin with for the traditional ML models they are not as sensitive to noise. Accuracy scores are close to random chance, so adding noise will not degrade performance further, both in terms of accuracy and IRT. In terms of the DNN models, the LSTM model is more significantly affected by noise than the CNN and NSE models. With only 5% of the training set replaced by noise, performance in IRT drops sharply. The noise may affect the types of long-term dependencies that are being learned by the model, which would explain the sharp drop in IRT scores, whereas the accuracy score change is more linear. In particular, entailment and neutral scores drop much more quickly than scores for contradiction for the LSTM model. It may be the case that IRT scores for entailment drop so quickly because the entailment test set is very easy (Lalor et al., 2016). Therefore any additional missed questions can affect the score significantly.

Although initial performance for the NSE and CNN models is lower than the LSTM, the NSE scores are less sensitive to noise. In particular, the NSE model outperforms both LSTM and CNN models in all test sets when 20% or more of the training data is noise. The results suggest that NSE is not simply memorizing but is able to learn representations of the data that are robust to noise.
4.2.2 Learning with More Data

We next consider model performance as training set size changes. We expect that by increasing the training set size when training a model, performance will also increase (Halevy et al., 2009).

Figure 2 plots performance metrics as a function of the training set size used to train the models. Our results (Figure 2) show that if measured by accuracy, all DNNs’ performances increase almost linearly with training size. In contrast, IRT is more sensitive to training size, demonstrating the benefit of using IRT as the evaluation metric.

For small training sets, the network cannot learn, and IRT scores reflect this. As the training set increases from 5000, performance in terms of IRT begin to improve, and continue to improve as more data is used for training. Moreover, for the DNN models there is a nonlinear trend in performance as training size increases. In human learning, a popular teaching methodology is a nonlinear teaching pedagogy, where curricula are flexible and adapted according to students’ learning styles and speeds (Chow et al., 2007). The nonlinear pedagogy seems to suggest a nonlinear learning process in humans (Thelen and Smith, 1996), which is comparable to the nonlinear trend found in the DNN models as shown by our analyses.

Our results also show that although NSE outperforms both the LSTM and CNN models in all training sizes according to accuracy, its performance varies when measured by the IRT scores. Specifically when training size is relatively small, NSE is the best DNN model. Previous work has shown that memory networks are able to learn faster than LSTMs (Graves et al., 2014). However, when trained with the full SNLI training set the LSTM model outperforms in each IRT group. Our earlier results demonstrate that LSTM is more noise intolerant (§ 4.2.1), suggesting that LSTM is the best DNN model for RTE if it is trained on a large, high quality data set. Due to NSE’s noise resilience, our results show that NSE can generalize better than other DNNs, and can perform better when small training data is available.

4.2.3 Learning Easy and Difficult Items

We next examine how the performance of the models change with item difficulty in the IRT test sets. Using the model outputs when trained on different amounts of data (§ 4.2.2), we performed logistic regression to predict whether a model would label an item correctly using the training size of the model, item difficulty and their interaction as predictors. We performed the regression for each IRT item group to identify per-group effects. For parameter interpretation of this analysis, we used the log of the training size and transformed the training size so that the largest value is 0. We normalized the item difficulties so that the mean item difficulty in the group is set to 0. For each model, we examined each regression to identify statistically significant ($p < 0.05$) coefficients.

For the traditional ML models, in most cases there were no significant coefficients for training set size, item difficulty, or their interactions. For a small number of subsets (e.g. KNN group 4C), there was a negative coefficient for difficulty, which suggests that more difficult items are less likely to be answered correctly. In other cases (e.g. DT group 5E), there was a positive coefficient for training size, meaning that more data is associated with a higher likelihood of answering correctly.

For the DNN models, there were several sub-
sets of items with interesting results. Each DNN model (LSTM, CNN, NSE) had at least one subset regression with a negative difficulty coefficient, a positive training set size coefficient, and a negative interaction coefficient. For an item with difficulty 0, as more training data is used, performance increases. The negative coefficient of interaction means that the negative slope associated with difficulty is flatter when training size is smaller (e.g. difficult has a smaller effect on performance). In addition, this interaction parameter tells us that the positive association with training size is steeper for easier items (items with difficulty less than 0) and flatter for harder items. In other words, easier items are easier to learn than the difficult items. This means that the slope of performance (in terms of log-odds) with respect to log of training size decreases with item difficulty. In other words, more difficult items are harder to learn and have a flatter learning curve. This negative interaction also means that the slope of performance with respect to item difficulty decreases with training size.

In all DNN regressions but one (CNN, group 4C), training set size had a significant positive coefficient. This indicates that increasing the training set size increases the likelihood of answering an item correctly, which is not surprising. Interestingly, for the LSTM group 5E regression we found significant positive coefficients for training set size, item difficulty, and their interaction. Interpreting this result as before tells us that difficulty has a larger effect on performance when there is less training data, and the more difficult items are easier to learn. We must consider the specific data subset for this result, as the 5E subset is composed of very easy items (Lalor et al., 2016). It could be that once the LSTM model is trained with the full training set, difficulty is less relevant because all of the items are easy.

With more training data, the models move away from treating each question equally regardless of difficulty to a structure more consistent with that of human learning, where the probability of answering an easy question correctly is higher than that of answering a difficult question correctly. This aligns with the expectation that there is a higher probability of answering an easy question than a harder one. As this interaction is evident...
Table 3: Count of correct sentence pairs per category for models trained on the full SNLI training set (Change from models trained on 100-example training set in parentheses). Largest count and largest change are boldfaced for emphasis.

| Category | A | B | C | D | E | F | G | H | I | J | K | L | M |
|----------|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Count    | 76| 38| 10| 18| 12| 17| 24| 30| 18| 12| 16| 30| 66|
| DT       | (13)| (11)| (11)| (6)| (6)| (8)| (6)| (8)| (6)| (5)| (7)| (10)| (5)| (10)| (5)| (23)| (0)|
| KNN      | (30)| (5)| (2)| (5)| (1)| (4)| (5)| (3)| (1)| (4)| (10)| (3)| (1)| (1)| (1)| (7)| (1)| (22)| (0)|
| LR       | (30)| (3)| (5)| (5)| (5)| (4)| (3)| (1)| (4)| (4)| (4)| (4)| (10)| (5)| (1)| (1)| (1)| (15)| (10)|
| LSTM     | (30)| (3)| (5)| (5)| (5)| (4)| (1)| (1)| (1)| (1)| (1)| (1)| (1)| (1)| (1)| (1)| (1)| (1)| (1)|
| CNN      | (51)| (24)| (8)| (5)| (5)| (3)| (2)| (1)| (4)| (4)| (4)| (4)| (10)| (5)| (1)| (1)| (1)| (1)| (1)|
| NSE      | (62)| (31)| (9)| (6)| (17)| (13)| (12)| (12)| (22)| (14)| (9)| (15)| (7)| (5)| (1)| (1)| (1)| (1)|

Table 3: Count of correct sentence pairs per category for models trained on the full SNLI training set (Change from models trained on 100-example training set in parentheses). Largest count and largest change are boldfaced for emphasis.

in each DNN model that we tested, we can say that by increasing training size for an NLP model, not only does the expected overall performance increase (Halevy et al., 2009), but the models exhibit a more human-like learning capability with respect to the difficulty of the test set items.

4.2.4 Interpreting Patterns Learned

We can examine how well the models perform on subsets of data to determine if certain groups are more or less likely to be learned as training data size is increased. Table 3 shows the total number of sentence pairs answered correctly for each category by the fully trained models, as well as the change from the number of correctly answered sentence pairs for the models trained with only 100 examples. Categories in Table 3 align with the labels in §3.3.

As Table 3 shows, the DNN models both outperform the traditional ML models when trained on the full training set, but also show consistent improvement with the increased training size. While performance drops in some subsets for the traditional models, performance for each subset is as good or better with a full training set for the DNN models. Our results show that both LSTM and NSE in most cases have comparable performance. The LSTM model performs slightly better when dealing with restrictive words or phrases (H, I), summarization (K), and inference (M). On the other hand, the NSE model performs slightly better on pairs that align closely with small changes (D) and pairs with simplified hypotheses and semantic differences (J, L).

5 Discussion and Future Work

Evaluating progress in NLP requires effective metrics to measure algorithms output on test sets. In this paper we demonstrate the reliability of IRT as a metric for RTE. IRT scores are consistent in that they separate models based on performance in a similar manner as accuracy, while providing more information with regards to the items answered correctly by each model. Our experiments have shown that the IRT metric is consistent with the standard accuracy results in identifying high-performing and low-performing models.

We also examine the output of a number of models for RTE, and demonstrate the gains in performance in DNN models as training size increases. Not only does raw performance improve, but performance in semantic subsets of the test set systematically improves. As the DNN models are trained with more data, we see learning patterns emerge that mirror those of humans. Whereas with little training, easy and difficult items have similar likelihood of being answered correctly, with larger training set sizes easy items have a higher likelihood of being answered correctly.

We demonstrate the certain DNN models, specifically NSE, are robust to the presence of noise in a data set and can still learn with as much as 35% of noisy training data.

Our results are dependent on the difficulties estimated from a human population of AMT annotators. Therefore it is possible that certain subsets of questions had a greater influence on the IRT models than others. A larger set of items in the IRT models could reduce the implicit weighting of certain questions and have a more appropriate distribution of ability levels. This is difficult due to the need for human annotators, but automating response pattern generation would be an interesting direction for future work.

Future work can explore additional models using more specialized features to attempt to improve performance. Ensemble models that consider the output of multiple DNNs (e.g. CNNs, RNNs, and memory networks) can take advantage of the high performance of different categories of sentence pair to further improve performance. In addition, IRT should be examined for other tasks to confirm the validity of the metric across NLP.
References

Frank B. Baker and Seok-Ho Kim. 2004. Item Response Theory: Parameter Estimation Techniques, Second Edition. CRC Press.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, Curran Associates, Inc., pages 4349–4357. http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and D. Christopher Manning. 2015. A large annotated corpus for learning natural language inference
In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pages 632–642. https://doi.org/10.18653/v1/D15-1075

Rebecca F. Bruce and Janyce M. Wiebe. 1999. Recognizing subjectivity: A case study in manual tagging
Nat. Lang. Eng. 5(2):187–205. https://doi.org/10.1017/S1351324999002181

Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. 2002. Deep blue
Artificial Intelligence 134(1):57 – 83. https://doi.org/10.1016/S0004-3702(01)00129-1

Jia Yi Chow, Keith Davids, Chris Button, Rick Shuttleworth, Ian Renshaw, and Duarte Arajo. 2007. The role of nonlinear pedagogy in physical education
Review of Educational Research 77(3):251–278. https://doi.org/10.3102/003465430730305615

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge
In Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment, Springer, pages 177–190. https://doi.org/10.1007/11736790_9

A. Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm
Journal of the Royal Statistical Society. Series C (Applied Statistics) 28(1):20–28. https://doi.org/10.2307/2346806

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building watson: An overview of the deepqa project. AI magazine 31(3):59–79.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines
CoRR abs/1410.5401. http://arxiv.org/abs/1410.5401

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data
IEEE Intelligent Systems 24(2):8–12. https://doi.org/10.1109/MIS.2009.36

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mac
In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pages 1120–1130. http://aclweb.org/anthology/N13-1132

Been Kim, Dmitry M. Malioutov, and Kush R. Varshney. 2016. Proceedings of the 2016 ICML workshop on human interpretability in machine learning (WHI 2016) CoRR abs/1607.02531. http://arxiv.org/abs/1607.02531

Yoon Kim. 2014. Convolutional neural networks for sentence classification
In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pages 1746–1751. https://doi.org/10.3115/v1/D14-1181

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2016. Building machines that learn and think like people
Behavioral and Brain Sciences pages 1–101. https://doi.org/10.1017/S0140525X16001837

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. Association for Computational Linguistics, pages 648–657. http://aclweb.org/anthology/D16-1062

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions
In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pages 107–117. http://aclweb.org/anthology/D16-1011

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp
In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pages 681–691. https://doi.org/10.18653/v1/N16-1082

Fernando Martinez-Plumed, Ricardo B. C. Prudencio, Adolfo Martinez Us, and Jos Hernandez-Orallo. 2016. Making sense of item response theory in machine learning
In ECAI, IOS Press, volume 285 of Frontiers in Artificial Intelligence and Applications, pages 1140–1148. https://doi.org/10.3233/978-1-61499-672-9-1140
Warren S McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5(4):115–133.

Tsendsuren Munkhdalai and Hong Yu. 2016. Neural semantic encoders. CoRR abs/1607.04315. http://arxiv.org/abs/1607.04315.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. Transactions of the Association of Computational Linguistics 2:311–326. http://aclweb.org/anthology/Q14-1025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Mathieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12:2825–2830.

Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In International Conference on Learning Representations.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized experience replay. In International Conference on Learning Representations. Puerto Rico.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. Nature 529:484–503. https://doi.org/10.1038/nature16961.

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. 2015. Incentivizing exploration in reinforcement learning with deep predictive models. CoRR abs/1507.00814. http://arxiv.org/abs/1507.00814.

Esther Thelen and Linda B Smith. 1996. A dynamic systems approach to the development of cognition and action. MIT press.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. CoRR abs/1611.03530. http://arxiv.org/abs/1611.03530.