Probing Difficulty and Discrimination of Natural Language Questions With Item Response Theory

Anonymous ACL submission

Abstract

Item Response Theory (IRT) has been extensively used to characterize question difficulty for human subjects in domains including cognitive psychology and education (Primi et al., 2014; Downing, 2003). In this work, we explore IRT to characterize the *difficulty* and *discrimination* of natural language questions in Question-Answering datasets. We use HotPotQA for illustration. Our analysis reveals significant variations along these traits, as well as interdependence between them. Additionally, we explore predictive models for directly estimating these traits from the text of the questions and answers. Our experiments show that it is possible to predict both difficulty and discrimination parameters for new questions, and these traits are correlated with features of questions, answers, and associated contexts. Our findings can have significant implications for the creation of new datasets and tests on the one hand and strategies such as active learning and curriculum learning on the other.

1 Introduction

The use of question answering for testing learning often relies on characterizing questions on aspects such as *difficulty* and *discrimination*[1]. For example, ordering questions by difficulty can enable curriculum learning (Bengio et al., 2009). Similarly, discrimination is used in standardized exams such as the SAT to ensure that questions are varied enough to discriminate between high-ability and low-ability respondents. Item Response Theory (IRT) (Wright and Stone, 1979; Lord, 1980) has been a widely applied framework to jointly estimate such parameters for questions (or *items*) and the abilities of *respondents*. While IRT has its inception in psychometrics and has traditionally been used with human respondents, recently, it has been explored for analyzing predictions from an 'artificial crowd' of ML models (Prudêncio et al., 2015; Plumed et al., 2016; Martínez-Plumed et al., 2019).

However, despite recent work such as Lalor et al. (2019), the use of IRT in NLP, especially Question Answering(QA), remains under-explored. For example, the questions in QA datasets often vary in terms of the skills required to answer them, but accuracy-based evaluation metrics are insensitive to which questions a model gets right and susceptible to dataset biases. IRT provides a natural framework to analyze such complexity. In this work, we use IRT to explore questions in QA datasets. We use HotPotQA (Yang et al., 2018) as our testbed, but our analysis can be applied to other datasets. We diverge from previous work (Prudêncio et al., 2015; Plumed et al., 2016; Martínez-Plumed et al., 2019; Lalor et al., 2019) that has used IRT to analyze AI datasets in two ways: firstly, we explore both difficulty and discrimination, whereas earlier work has focused on the 1PL IRT model that only estimates difficulty. Secondly, IRT analysis for large datasets suffers from the issue of computational scaling. Estimating the parameters for a single new question requires re-fitting the IRT model on a large number of question-model pairs. To address this, we explore directly predicting IRT parameters from the text of a question and its related contexts. In doing this, we experiment with multiple categories of NLP-centric and cognitive features.

Our analysis shows significant variations among questions and reveals some surprising patterns. We show that it is possible to predict both difficulty and discrimination of natural language questions, which can have multiple applications in education and pedagogy. Additionally, we see that different surface-level features are associated with high discrimination and high difficulty, which can inform

[1] By difficulty, we refer to how likely a respondent is to answer a question correctly, whereas by discrimination we refer to the value of a question in identifying a given level of ability in respondents. A question like '2+2=?' has low difficulty but potentially high discrimination, since a respondent who answers incorrectly is likely to have no arithmetic ability.

new evaluation methods and the creation of new datasets. Further, we identify attributes for predicting difficulty and discrimination that are general enough to be adapted to various QA datasets.

2 IRT Analysis of HotPotQA

IRT background: We begin by summarizing the 1PL and 2PL models from IRT, which form the basis of our later analysis. The 1PL (1 Parameter Logistic) model describes the probability of respondent i correctly answering the j'th item (question) in terms of scalar-valued parameters for question difficulty (d_j) and respondent ability (θ_i). These parameters are estimated from data $y_{ij} \in \{0, 1\}$ for a set of i, j pairs. Here, $y_{ij} = 1$ indicates a correct answer. The 1PL model is described by:

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

The 2PL model extends this through a scalar-valued parameter α_j, which represents the discrimination of the j'th item. Intuitively, this parameter denotes how sharply the probability of answering a question correctly changes as the ability of the respondent increases. The 2PL model is described by:

$$p(y_{ij} = 1 | \theta_i, d_j, \alpha_j) = \frac{1}{1 + e^{-\alpha_j(\theta_i - d_j)}}$$

Dataset description: We chose HotPotQA for our analysis since it is significantly more complex than other datasets such as SQuAD (Rajpurkar et al., 2016), due to the questions requiring multi-hop reasoning and having more complex language. In HotPotQA, each question is paired with two paragraphs considered 'gold' contexts and several other paragraphs considered 'distractor' contexts. The answer to each question is a span in one of the gold contexts, but correctly answering the question requires combining information from both 'gold' paragraphs.

2.1 Estimating IRT Parameters

Due to the size of HotPotQA's dev set (7,405 questions), collecting human responses for each question, as needed by IRT, is infeasible. Motivated by Lalor et al. (2019), we create an artificial crowd of QA models. For this, we train 148 instances of DFGN (Qiu et al., 2019) models on HotPotQA's train set.We choose DFGN due to its competitive performance on the HotPotQA leaderboard, and the number of models we train is primarily driven by computational limits. To ensure diversity, we uniformly sample the number of training epochs from 1 to 15 and sample the fraction of the training data used for model training from $\mathcal{U}(0, 1)$. Each model was otherwise trained with the hyperparameters described in Qiu et al. (2019). Next, we generate a binary matrix indicating which questions from the HotPotQA dev set each model answered correctly (based on exact matches with the respective answers). We remove any questions that received no correct answers or no incorrect answers. This is done as during the estimation process, these questions tend towards (+/-) infinity in their difficulty parameters and have zero discriminatory ability. Our final dataset is a subset of 4,000 questions. Finally, we fit the 1PL and 2PL models on the foresaid item-response matrix using the variational IRT training procedure from Natesan et al. (2016).

2.2 Analysis of Estimated Parameters

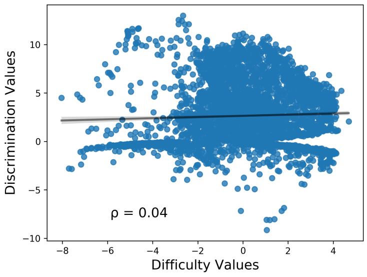

Figure 1: 2PL discrimination vs 1PL difficulty for questions.

Figure 1 shows a scatter-plot of estimated difficulty and discrimination values for individual questions. We note that some discrimination values asymptotically approach 0. This occurs when some questions receive very few or many correct answers; these questions cannot discriminate high-performing from low-performing models. We also note that some questions have negative discrimination, i.e., as a model's ability increases, its probability of answering the question correctly decreases. This is primarily a result of some higher-performing models giving an answer which is either a subspan of or contains the ground-truth answer. Overall, there is a weak positive correlation between discrimination and difficulty ($\rho = 0.04$).

To visualize any correlation between question type and difficulty levels, we clustered questions based on their BERT embeddings using KMeans ($K = 20$) clustering (2D UMAP shown in 2 Through manually examining and labeling the clus-

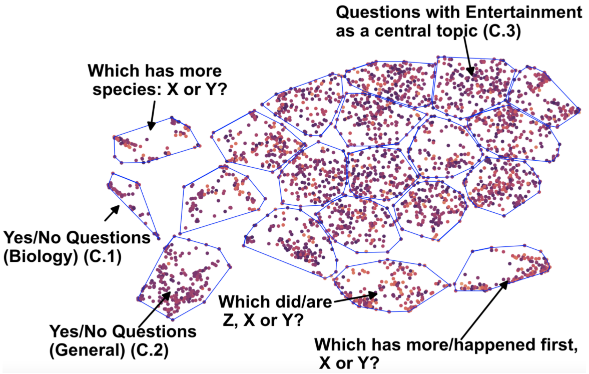

Figure 2: All 3000 questions from our train/dev set as UMAP-reduced BERT embeddings, color-coded by difficulty (darker is more difficult). We find that clusters produced by KMeans ($K = 20$) naturally cluster together questions that are similar in how they are asked/topics that are covered. We label some clusters according to these types. We specially mark C.1, C.2, and C.3. C.1 and C.2 have uniformity in the type of question being asked, as well as lower variance than other clusters. C.3 is uniform in topic but can vary in the type of question.

ters, we found that questions got clustered according to their style (e.g. yes/no questions) and topic. Some clusters, such as C.3, have a large variety in the types of questions being asked, in particular, a large variety in the type of answer to the questions. Both *Q: Khushi Ek Roag is broadcast by a company based out of where? A: Dubai* and *Q: To Catch a Predator was devoted to impersonating people below the age of consent for which in North America varies by what? A: jurisdiction* are in C.3.

Other clusters, such as C.1 and C.2, (yes/no clusters), only vary in topic rather than the type of question. In particular, for these clusters, the estimated difficulty has significantly lower variance than the other clusters ($\rho = 0.02$, $\rho = 0.04$ respectively), indicating that these yes/no questions tend to be consistent in their difficulty. On the other hand, the difficulty of questions in other clusters cannot be described entirely by their type or topic. We further explore how these factors affect predicting the difficulty values in Appendix B.

3 Predicting IRT Parameters

We next discuss predictive models for discrimination and difficulty using features from the question, answer, and associated context.

3.1 Feature Design

We experiment with two categories of features: Human-centric and Machine-centric features. For Human-centric features, we considered (1) counting-based **Lexical & Syntactic features** extracted for both questions and answers like ContentWords, Type-token ratio, Avg. Word Length, Complex Words ($>$ 3 syllables)); (2) **Semantic-Ambiguity features** measuring a question's or answer's ambiguity (Ha et al., 2019); and (3) **Readability features** based on measures like Fleisch Kincaid index. More feature details can be found in Appendix D. For Machine-Centric Features we considered (1) **Contextual Embeddings** for questions and answers from BERT (Devlin et al., 2019); (2) n-gram **Overlap Counts** between the question and answer, and between question/answer and the gold/ distractor paragraphs; and (3) **POS Counts** from the Stanford Tagset (Toutanova et al., 2003) for the question and answer.

3.2 Quantitative Analysis and Ablation

Features	Dev MSE	Dev R^2	Test MSE	Test R^2
All	**5.14**	**0.11**	**4.72**	**0.17**
All (A)	5.41	0.08	5.05	0.11
All (Q)	5.43	0.07	5.10	0.10
BERT (Q)	5.41	0.07	4.99	0.12
BERT (A)	5.25	0.10	5.05	0.11
H.C. (Q)	5.62	0.01	5.38	0.05
Lex. & Syn. (Q)	5.62	0.01	5.37	0.05
Read. (Q)	5.80	0.00	5.71	0.00
Sem. Ambiguity (Q)	5.76	0.01	5.55	0.02
H.C. (A)	5.45	0.06	5.20	0.08
Lex. & Syn. (A)	5.47	0.03	5.36	0.06
Read. (A)	5.63	0.02	5.48	0.03
Sem. Ambiguity (A)	5.81	0.01	5.68	0.00
A/Q/C Overlap	5.39	0.05	4.92	0.13
P.O.S. (Q)	5.37	0.05	5.23	0.08
P.O.S. (A)	5.60	0.01	5.28	0.07
Mean	5.82	0.00	5.69	0.00

Table 1: Results for predicting the 1PL difficulty parameters. BERT (Q) and BERT (A) use the BERT embeddings for the ques./ans. respectively. H.C. (Q)/(A) are the human-centric features for the ques./ans. respectively. A/Q Con. is using only the overlap counts between question, answer, and contexts.

Table 1 and Table 2 show the regression performance of our predictive models for predicting IRT difficulty/discrimination parameters using the feature-sets described before. The reported results are averaged over a 10-fold cross-validation. We note that the best models for both difficulty and discrimination show significant ($\rho < 0.1$) predictive performance (R^2 of 0.169 and 0.126).

The best performance is achieved in both tasks by considering all features. In both cases, there is a significant difference ($\rho < 0.1$) in performance between using any single set and using all features, except the best performing BERT feature set. We also note that features derived from the answer

Features	Dev MSE	Dev R^2	Test MSE	Test R^2
All	9.08	**0.13**	**9.14**	**0.13**
All (A)	9.59	0.08	9.98	0.04
All (Q)	9.32	0.10	9.50	0.09
BERT (Q)	**9.02**	0.11	9.27	0.11
BERT (A)	9.52	0.08	9.64	0.08
H.C (Q)	9.76	0.04	9.86	0.06
Lex. & Syn. (Q)	9.75	0.04	9.86	0.06
Read. (Q)	10.1	0.01	10.2	0.03
Sem. Ambiguity (Q)	10.1	0.02	10.2	0.03
H.C (A)	10.1	0.03	10.3	0.02
Lex. & Syn. (A)	10.1	0.01	10.2	0.03
Read. (A)	10.1	0.02	10.3	0.01
Sem. Ambiguity (A)	10.2	0.00	10.5	0.00
A/Q/C Overlap	9.56	0.06	9.63	0.08
P.O.S. (Q)	9.96	0.04	10.1	0.03
P.O.S. (A)	9.78	0.03	9.82	0.06
Mean	10.2	0.00	10.5	0.00

Table 2: Results for predicting the 2PL discrimination parameters. The setup is the same as in table 1. BERT (Q) has the highest performance. However, the difference between in performance when using BERT (Q) and All is not statistically significant. See Appendix for significance tests

are typically better at capturing difficulty, while features derived from the question better predict the discrimination parameters. However, the performance of All (Q) and All (A) for both the discrimination and difficulty is weaker than using All features. Since the difference is not statistically significant, it is unclear how much predictive power is added when considering both answer and question features in these predictions.

The features that focus on human difficulty are among the less effective feature sets, indicating that the human difficulty features of a question do not fully capture difficulty for QA models. We provide details of models and their training, the experiment setup, and significance tests in the Appendix.

3.3 Feature Importance Study

Feature	Change in MSE	Interval	Corr.
# Commas A.	.056	± .015	.096
# Complex Words A.	.051	± .011	-.042
# NNP A.	.048	± .016	-.157
# SNP A/G.C.	.016	± .006	.037
# Commas Q.	.013	± .006	-.114

Table 3: Feature Importances for Difficulty Parameters (All Features Considered). A. refers to a feature capturing information from the answer, Q. refers to a feature capturing information from the question. A/G.C. refers to a feature measuring overlap between the answer and gold contexts.

We estimated feature importance by permuting each feature individually and measuring the change in MSE on the dev set. We list features that caused a change in MSE of least .01 in tables 4 and 3. The feature importances for difficulty and discrimination predictions are listed in tables 3 and 4.

Feature	Change in MSE	Interval	Corr.
# CD A.	.245	± .034	.166
# Commas Q.	.078	± .017	-.113
Avg. Sense/Adverb A.	.008	± .017	-.026

Table 4: Feature Importances for Discrimination Parameters (All Features Considered)

We point out that for predicting the discrimination, the number of cardinal digits in the answer was the most important indicator of high discrimination. The positive correlation between the number of digits in the answer and the discrimination of a question is expected. Qiu et al. (2019) show that the DFGN model has a significant weakness in numeric operations. This gives questions with numeric answers a high discrimination value as DFGN models are naturally inhibited in this regard, and thus only a few models with the most training data will be capable of answering these questions. We find a similar positive Pearson score ($\rho = 0.14$) between the difficulty and the number of cardinal digits in the answer. However, we do not see this feature in the importances for the difficulty. While this weakness of the DFGN model cannot be applied to an arbitrary QA model, the methodology used to determine this weakness can be applied arbitrarily, which can give solid grounding to claims about model weaknesses.

4 Conclusion

In this paper, we explored QA datasets through the lens of Item Response Theory. We have demonstrated a way to build regression models that can describe the difficulty and discrimination of a question. It is important to point out that our models' predictive abilities depend to some degree on the type of question. More expressive models may be able to improve the performance over our interpretable linear models. We also note that our analysis here focused on QA, there are many NLP tasks in which the difficulty or discrimination of an item may be important. Our work here could naturally extend to these domains. Finally, being able to automatically predict these traits without relying on user responses can engender a host of creative educational applications. Future work can also leverage such predictive models to explore more efficient strategies for learning and evaluation.

References

Moez Ali. 2020. PyCaret: An open source, low-code machine learning library in Python. PyCaret version 2.2.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven M Downing. 2003. Item response theory: applications of modern test theory in medical education. Medical Education, 37(8):739–745.

Eileen B. Entin and George R. Klare. 1978. Some inter-relationships of readability, cloze and multiple choice scores on a reading comprehension test. Journal of Reading Behavior, 10(4):417–436.

R. Flesch. A new readability yardstick. Journal of applied psychology, 32(3).

R. Gunning. 1952. The Technique of Clear Writing. McGraw-Hill, New York.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 11–20, Florence, Italy. Association for Computational Linguistics.

J. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

John P Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

G. Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. Journal of Reading, 12(8):639–646.

Frederic M. Lord. 1980. Applications of Item Response Theory to Practical Testing Problems. Routledge.

Fernando Martínez-Plumed, Ricardo B.C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in ai: Analysing machine learning classifiers at the instance level. Artificial Intelligence, 271:18 – 42.

George A. Miller. 1995. WordNet: A lexical database for English. Commun. ACM, 38(11):39–41.

P Natesan, R Nandakumar, T Minka, and JD Rubright. 2016. Bayesian prior choice in irt estimation using mcmc and variational bayes. Front. Psychol. 7: 1422. doi: 10.3389/fpsyg.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Fernando Plumed, Ricardo Prudêncio, Adolfo Martínez-Usó, and Jose Hernandez-Orallo. 2016. Making sense of Item Response Theory in machine learning.

Caterina Primi, Kinga Morsanyi, Maria Anna Donati, and Francesca Chiesi. 2014. Item Response Theory analysis of the Cognitive Reflection Test: Testing the psychometric properties of the original scale and a newly developed 8-item version, pages 2799–2804.

R. Prudêncio, J. Hernández-Orallo, and A. Martınez-Usó. 2015. Analysis of instance hardness in machine learning using item response theory.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6140–6150, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

F. A. Smith and R.J. Senter. 1967. Automated readability index. Technical Report AMRL-TR-6620.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, page 173–180, USA. Association for Computational Linguistics.

Benjamin D. Wright and Mark H. Stone. 1979. Best test design. Mesa Press.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Conference on Empirical Methods in Natural Language Processing (EMNLP).

A Models & Training

For the 1PL and 2PL prediction, we considered linear models with L1 & L2 regularization, random forests, gradient boosted regressors, and bayesian ridge models. All hyperparameters were kept constant as the default in the sklearn package (Pedregosa et al., 2011) We performed 10-fold cross-validation using PyCaret (Ali, 2020). All models were trained on a consumer grade processor.

B Qualitative Analysis

We qualitatively analyze the difficulty predictions to understand the performance of our model[2]. Figure 3 shows a UMAP scatterplot for questions on a test split for our difficulty prediction analysis, color-coded by the absolute error in our predictions. In the spirit of figure 2, we apply KMeans ($k = 10$), to our data. We highlight C.1, in a similar vein to C.1 and C.2 of Figure 2, this cluster consists of yes/no question. Highlighting this similarity, the prediction error in this cluster has significantly smaller variance than the rest of the clusters ($\rho = .02$). We find that model appears to have tuned into the consistent difficulty of yes/no questions, and as such has consistently lower error in predicting difficulty for questions of this type, while the more granular details required to aptly describe the difficulty in other question types isn't always captured by the model.

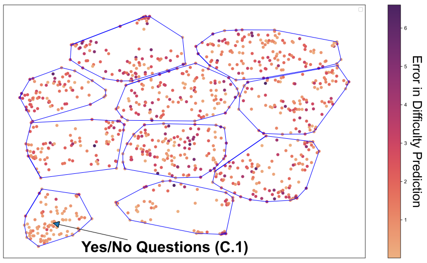

Figure 3: UMAP scatterplot of questions color coded by prediction error for difficulty. (Test set)

C Feature Definitions

- **Human-Centric Features**
  - **Lexical & Syntactic features**: These consist primarily of counting features: ContentWords, Type-token ratio, Avg. Word Length, Complex Words (> 3 syllables). These are calculated for both answer and question. A complete list can be found in the Appendix (G).
  - **Semantic-Ambiguity features**: We use WordNet (Miller, 1995) to calculate the ambiguity of sentences similar to Ha et al. (2019). These are calculated for both answer and question.
  - **Readability features**: We use previous work (Kincaid et al., 1975; Gunning, 1952; Laughlin, 1969) to model the readability of a question/answer (e.g. Fleisch Kincaid index). These are further expanded on in section D.
- **Machine-Centric Features**
  - **Contextual Embeddings**: We use the BERT-base model (Devlin et al., 2019) to obtain sentence embeddings for questions and answers.
  - **Overlap Counts**: We count overlaps between the question and answer of n-grams up to $n = 3$. We also compute overlap counts between the question/answer and the gold and distractor paragraphs.
  - **Part of Speech Counts**: We count POS tags for tags from the Stanford NLP tagset (Toutanova et al., 2003) for both the question and answer.

D Reading Difficulty Features

We list the reading difficulty features we used in our experiments as well as an overview of their calculations. Each calculation has it's own coefficents that can be found in their respective citations.

- Flesch Reading Ease - linear combination of words/sentence and syllables/word (Flesch)

- Flesch Kincaid Grade Level - linear combination of word/sentence and syllables/word (Kincaid et al., 1975)

- Automated Readability Index (ARI) - linear combination of characters/word and words/sentence (Smith and Senter, 1967)

- Gunning Fog index - linear combination of words/sentence and complex words/words. Complex words are words with 3 syllabus (Gunning, 1952)

- Coleman-Liau - linear combination of letters/100 words and sentences/100 words.(Entin and Klare, 1978)

- SMOG index - calculates the grade level by considering the number of complex words/sentence (Laughlin, 1969)

[2] Similar analysis for the discrimination parameters is included in the Appendix (H).

E Significance Tests

We provide significance tests for difficulty and discrimination predictions in tables 6 and 5. We see that the BERT features and using all features are able to beat the baseline with statistical significance ($\rho <= .1$). Note that we compare using MSE rather than R^2 as the baseline always has an R^2 score of 0. We also provide in table 7 the significance tests for using all features against BERT features. We find that the best performing BERT feature set does not have a statistically significant improvement in performance when compared to the all feature set. In this case we use R^2 as the performance metric.

Features	p
All	**0.007**
BERT (Q)	**0.013**
BERT (A)	**0.098**
H.C. (Q)	0.165
H.C. (A)	0.726
A/Q Con.	0.831
P.O.S. (Q)	0.656
P.O.S. (A)	0.174

Table 5: 2PL discrimination predictions. P-values for feature set performance (MSE) tested against the baseline.

Features	p
All	**0.034**
BERT (Q)	0.211
BERT (A)	**0.078**
H.C. (Q)	0.551
H.C. (A)	0.261
A/Q Con.	0.674
P.O.S. (Q)	0.501
P.O.S. (A)	0.523

Table 6: 1PL difficulty predictions. P-values for feature set performance (MSE) tested against the baseline.

Features	p
BERT (Q) (Diff.)	**0.042**
BERT (Q) (Discrim.)	0.769
BERT (A) (Diff.)	0.278
BERT (A) (Discrim.)	**0.089**

Table 7: 1PL and 2PL Difficulty and Discrimination predictions. P-values for BERT performance (R^2) tested against all features performance.

F Counterfactual Analysis

We explore our model through a counterfactual analysis. Motivated from our qualitative analysis, we look at a question in the form of clusters (2) and (3) from the figure provided in the text. In particular, we are interested in taking an item with high prediction error and slightly tweaking it to understand how the model's predictions can change with changes in the question and answer. We randomly selected an item from clusters 1 and 2 with >2 absolute error to perform this experiment. The question we use in this study is in figure 4. We found that most simple changes to the question (using synonyms, removing unnecessary information, etc.) are unable to significantly change the difficulty predictions. We modify the answer in two ways to achieve a higher difficulty prediction. First, we modify the answer to be a date. Second we modify the answer to be “yes." We see larger gains in difficulty when using yes as an answer. However, both increase the difficulty far more than changes to the question did.

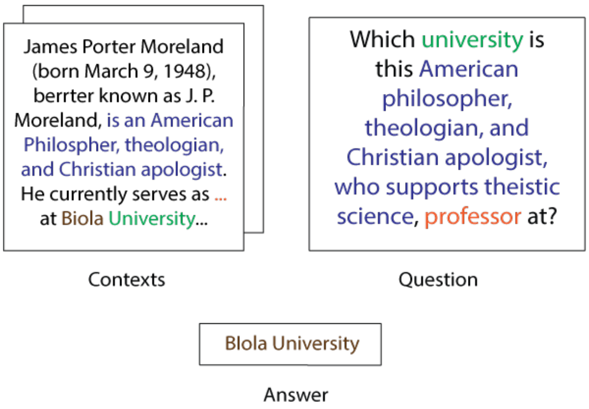
James Porter Moreland (born March 9, 1948), bertter known as J. P. Moreland, is an American Philospher, theologian, and Christian apologist. He currently serves as ... at Biola University...
Which university is this American philosopher, theologian, and Christian apologist, who supports theistic science, professor at?
Contexts
Question
Blola University
Answer

Figure 4: Our question used for our counter-factual analysis. This question had an estimated difficulty of 4.10 and a predicted difficulty of -.507.

We note some of the hardest questions in the test set were those that had yes/no answers or numeric answers. If the model was only considering the question type via the BERT embedding, we would expect the model to predict difficulties between 1 and 4. The prediction for the yes/no modification was 1.02, while the prediction for the date modification was 0.53. This indicates that the model is capable of taking into account individual characteristics of a question besides its surface level type and topic, however, it can be severely limited in this task depending on the question type. Some of our changes and their corresponding predictions are listed below:

- - Question (original): Which university is this American philosopher, theologian, and Christian apologist, who supports theistic science, professor at?’
  - Answer: "Biola University"
  - Pred. Diff: -.507

- - Question : Which school is this philoso-

  - pher and theologian who supports science, professor at?
  - Answer: "Biola University"
  - Pred. Diff: -.210

- - Question : What was the birth date of a professor at Biola University who is an American philosopher, theologian, and Christian apologist, who supports theistic science?
  - Answer: March 9, 1948
  - Pred. Diff: 0.534

- - Question : Does Biola University have a professor who is an American philosopher, theologian, and Christian apologist, who supports theistic science?
  - Answer: yes
  - Pred. Diff: 1.02

G Lexical Features

We list our full list of lexical features, these features are a subset of the lexical features used in Ha et al. (2019).

- Word Count
- Content Word Count
- Content Word Incidence
- Content Word Count No Stopwords
- Noun Count
- Noun Incidence
- Verb Count
- Verb Incidence
- Adjective Count
- Adjective Incidence
- Adverb Count
- Adverb Incidence
- Number Count
- Number Incidence
- Type Count
- Type Token Ratio
- Comma Count
- Comma Incidence
- Average Word Length In Syllables
- Complex Word Count
- Complex Word Incidence,
- Average Sentence Length
- Negation Count
- Negation Incidence
- Negation In Stem
- NP Count
- NP Incidence
- Average NP Length
- NP Count With Embedding
- NP Incidence With Embedding
- Average All NP Length,
- PP Count
- PP Incidence
- PPs Per Sentence Ratio
- VP Count
- VP Incidence
- Passive Active Ratio
- Proportion Active VPs
- Proportion Passive VPs
- Agentless Passive Count
- Relative Clauses Count
- Relative Clauses Incidence
- Proportion Relative Clauses
- Polysemic Word Count
- Polysemic Word Incidence
- Average Sense No Content Words
- Average Sense No Nouns
- Average Sense No Verbs

- Average Sense No Non Auxiliary Verbs
- Average Sense No Adjectives
- Average Sense No Adverbs
- Average Noun Distance To WNRoot
- Average Verb Distance To WNRoot,
- Average Noun And Verb Distance To WNRoot
- Answer Words In Word Net Ratio
- Average Word Frequency Abs
- Average Word Frequency Rel
- Average Word Frequency Rank
- Average Content Frequency Abs
- Average Content Frequency Rel
- Average Content Frequency Rank
- Not In First 2000 Count
- Not In First 2000 Incidence
- Not In First 3000 Count
- Not In First 3000 Incidence
- Not In First 4000 Count
- Not In First 4000 Incidence
- Not In First 5000 Count
- Not In First 5000 Incidence
- Imagability
- Imagability Found Only
- Imagability Ratio
- Familiarity
- Familiarity Found Only
- Familiarity Ratio
- Concreteness
- Concreteness Found Only
- Concreteness Ratio
- Age Of Acquisition
- Age Of Acquisition Found Only
- Age Of Acquisition Ratio
- Meaningfulness Colorado Found Only
- Meaningfulness Pavio Found Only
- No Imagability Rating
- No Familiarity Rating
- No Concreteness Rating
- No Age of Acquisition Rating
- Connectives Count
- Connectives Incidence
- Additive Connectives Count
- Additive Connectives Incidence
- Temporal Connectives Count
- Temporal Connectives Incidence
- Causal Connectives Count
- Causal Connectives Incidence
- Referential Pronoun Count,
- Referential Pronoun Incidence

H Discrimination UMAP plots

We provide the same UMAP plots as we did for difficulty in the paper. We see that clusters of similar questions have similar discrimination values as was the case for difficulty.

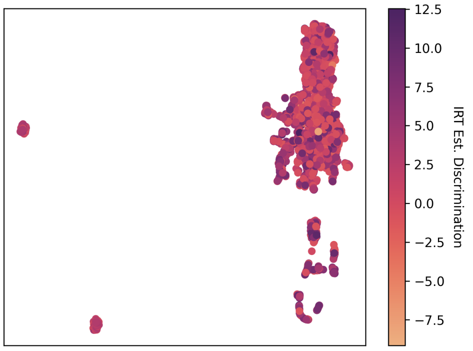

Figure 5: Answer BERT UMAP Reduction VS Discrimination values, train/dev set

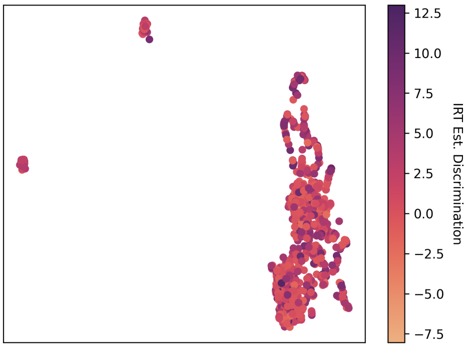

Figure 6: Answer BERT UMAP Reduction VS Discrimination values, test set

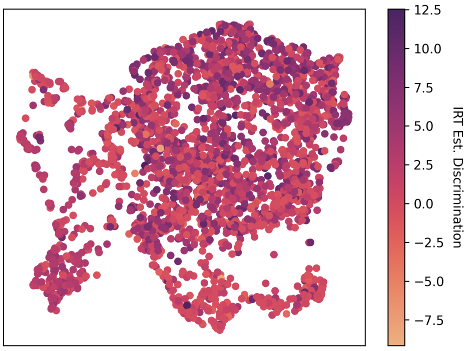

Figure 7: Question BERT UMAP Reduction VS Discrimination values, train/dev set

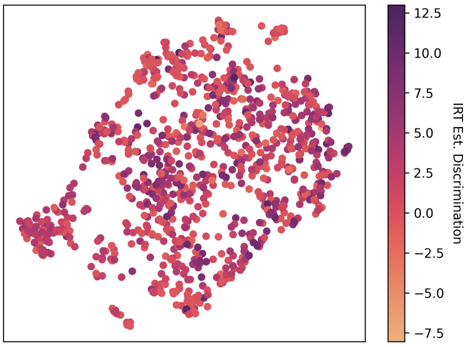

Figure 8: Question BERT UMAP Reduction VS Discrimination values, test set

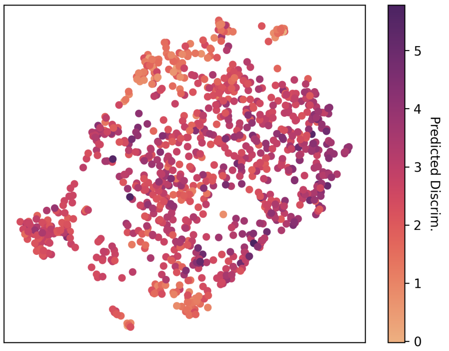

Figure 9: Question BERT UMAP Reduction VS Predicted Discrimination values, test set

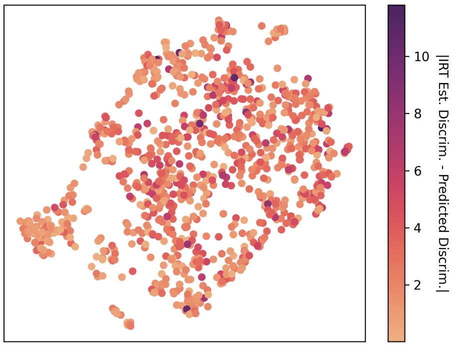

Figure 10: Question BERT UMAP Reduction VS Discrimination prediction error, test set