Predicting Difficulty and Discrimination of Natural Language Questions

Matthew A. Byrd Shashank Srivastava
University of North Carolina at Chapel Hill
matthew_a_byrd@outlook.com, ssrivastava@cs.unc.edu

Abstract

Item Response Theory (IRT) has been extensively used to numerically characterize question difficulty and discrimination for human subjects in domains including cognitive psychology and education (Primi et al., 2014; Downing, 2003). More recently, IRT has been used to similarly characterize item difficulty and discrimination for natural language models across various datasets (Lalor et al., 2019; Vania et al., 2021; Rodriguez et al., 2021). In this work, we explore predictive models for directly estimating and explaining these traits for natural language questions in a question-answering context. We use HotpotQA for illustration. 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; Lalor et al., 2019; Vania et al., 2021; Rodriguez et al., 2021).

While it can be helpful to know which questions are difficult/discriminatory, it can be equally important to be able to determine a question's difficulty/discrimination parameters without having to use it in a testing environment (as is required to estimate IRT parameters). Some recent work, such as Ha et al. (2019), has explored using features derived from the text of a question to predict the difficulty in the context of multiple-choice medical exams. While others (Benedetto et al., 2020) have used tf-idf features to predict the difficulty of questions as measured by IRT. We differ from these works in two ways: Firstly, while Ha et al. (2019); Benedetto et al. (2020) both predict the difficulty of items for humans, we are interested in predicting the difficulty (and discrimination) of items for QA models. Secondly, we choose a question-answering dataset, HotpotQA (Yang et al., 2018), as our testbed. We utilize this dataset to generate a rich and varied feature set across each item's question, answer, and associated contexts. We can then employ these features to analyze our difficulty and discrimination predictions, giving us insights into both our underlying QA model and factors that can increase the difficulty/discrimination of a question.

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.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 119 - 130

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]

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 the 1PL by adding 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' contexts.

2.1 Estimating IRT Parameters

We estimate the IRT parameters for the questions in HotpotQA's dev set ($7,405$ questions). However, collecting human responses for each question, which is necessary to estimate IRT parameters, is infeasible. Motivated by Lalor et al. (2019), we create an artificial crowd of QA models in place of a crowd of human respondents. For this, we train 148 instances of DFGN (Qiu et al., 2019) models on HotpotQA's train set.[3] 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 $\mathcal{U}(0, 1)$. Otherwise, each model was trained with the hyperparameters described in Qiu et al. (2019). Next, we generate an *item-response matrix* indicating which questions from the HotpotQA dev set each model answered correctly (i.e., the model's answer exactly matched the correct answer). 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, as well, their discrimination parameter estimate tends towards zero (unable to distinguish between high and low performing models). Our final dataset is a subset of $4,000$ questions ($2,000$ train, $1,000$ dev, and $1,000$ test). 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

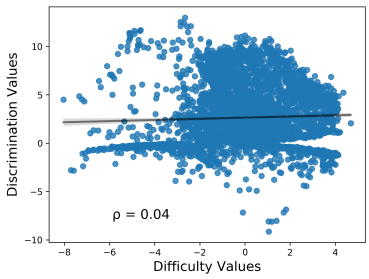

Figure 1: 2PL discrimination vs 1PL difficulty for questions.

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 of the highest per-

[2]Code, models, and data for all experiments are available at https://github.com/ByrdOfAFeather/pred_irt

[3]We choose DFGN due to its competitive performance on the HotpotQA leaderboard, the number of models we train is primarily driven by computational limits.

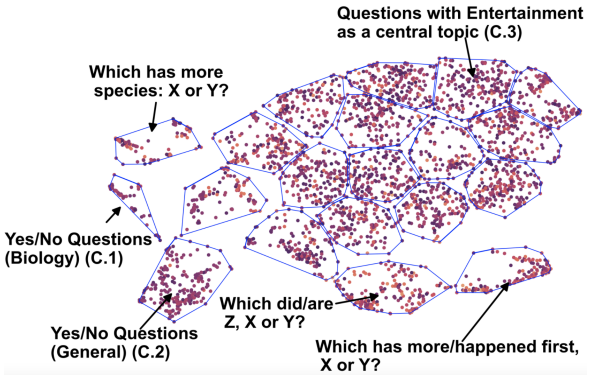

Figure 2: All 3000 questions from our train/dev set as UMAP-reduced BERT embeddings, color-coded by difficulty (darker is more difficult). We find that clusters produced by KMeans ($K = 20$) naturally cluster together questions that are similar in how they are asked or topics that are asked about. We 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.

forming models giving an answer which is either a subspan of or contains the ground-truth answer of questions that were otherwise answered correctly by lower-performing models. Overall, there is a weak positive correlation between discrimination and difficulty ($\rho = 0.04$).

To visualize any correlation between the semantic and syntactic information of questions and their respective difficulty levels, we clustered questions based on their BERT embeddings using KMeans ($K = 20$) clustering (2D UMAP reduction shown in Figure 2). Through manually examining and labeling the clusters, we found that many clusters could be described with a specific style (e.g., yes/no questions) or general topic. Some clusters, such as C.3, have a large variety in the phrasing of questions being asked and the potential answers in both syntactic and semantic features. For example, 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. The standard deviation values for C.1 and C.2 are 1.08 and 1.19 respectively, the average standard deviation value is 2.27. We further explore how these factors affect predicting the difficulty values in section 4.

3 Predicting IRT Parameters

We next discuss predictive models for discrimination and difficulty using features from the question, answer, and associated context. First, we describe our feature set, then provide an ablation study, a feature importance study, and finally qualitatively analyze the predictions of our best model.

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 C. 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

Table 1 and Table 2 show the regression performance of our models for predicting the IRT difficulty/discrimination parameters of the questions in our dev/test sets 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 ($\rho < 0.10$) predictive performance (R^2 of 0.17 and 0.13) against our baseline (Mean).

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 are typically better at capturing difficulty, while features derived from the question better predict the discrimination parameters. However, the per-

Features	Dev MSE	Dev R^2	Test MSE	Test R^2
All	**5.14**	**0.11**	**4.72**	**0.17**
All (Q)	5.43	0.07	5.10	0.10
All (A)	5.41	0.08	5.05	0.11
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
H.C. (A)	5.45	0.06	5.20	0.08
Lex. & Syn. (Q)	5.62	0.01	5.37	0.05
Lex. & Syn. (A)	5.47	0.03	5.36	0.06
Read. (Q)	5.80	0.00	5.71	0.00
Read. (A)	5.63	0.02	5.48	0.03
Sem. Ambiguity (Q)	5.76	0.01	5.55	0.02
Sem. Ambiguity (A)	5.81	0.01	5.68	0.00
P.O.S. (Q)	5.37	0.05	5.23	0.08
P.O.S. (A)	5.60	0.01	5.28	0.07
A/Q/C Overlap	5.39	0.05	4.92	0.13
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 question/answer respectively. H.C. (Q)/(A) are the human-centric features for the question/answer respectively. A/Q/C Overlap is using only the overlap counts between question, answer, and contexts.

Features	Dev MSE	Dev R^2	Test MSE	Test R^2
All	9.08	**0.13**	**9.14**	**0.13**
All (Q)	9.32	0.10	9.50	0.09
All (A)	9.59	0.08	9.98	0.04
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
H.C (A)	10.09	0.03	10.31	0.02
Lex. & Syn. (Q)	9.75	0.04	9.86	0.06
Lex. & Syn. (A)	10.13	0.01	10.21	0.03
Read. (Q)	10.08	0.01	10.17	0.03
Read. (A)	10.13	0.02	10.31	0.01
Sem. Ambiguity (Q)	10.05	0.02	10.16	0.03
Sem. Ambiguity (A)	10.21	0.00	10.47	0.00
P.O.S. (Q)	9.96	0.04	10.10	0.03
P.O.S. (A)	9.78	0.03	9.82	0.06
A/Q/C Overlap	9.56	0.06	9.63	0.08
Mean	10.21	0.00	10.53	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 in performance when using BERT (Q) compared to using All is not statistically significant. See Appendix D for significance tests.

formance 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 and the experiment setup in Appendix A; as well, significance tests can be found in Appendix D.

3.3 Feature Importance Study

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 at least .01 in tables 3 and 4.

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) showed 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. 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.

Feature	Change in MSE	Interval	Corr.
# Commas A.	0.06	$\pm$ 0.02	0.10
# Complex Words A.	0.05	$\pm$ 0.01	-0.04
# NNP A.	0.05	$\pm$ 0.02	-0.16
# SNP A/G.C.	0.02	$\pm$ 0.01	0.04
# Commas Q.	0.01	$\pm$ 0.01	-0.11

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.

Feature	Change in MSE	Interval	Corr.
# CD A.	0.25	$\pm$ 0.03	0.17
# Commas Q.	0.08	$\pm$ 0.02	-0.11
Avg. Sense/Adverb A.	0.01	$\pm$ 0.02	-0.03

Table 4: Feature importances for discrimination parameters (all features considered)

4 Qualitative Analysis

We qualitatively analyze the difficulty predictions to understand the predictions of our best-performing model. Similar to Figure 2, Figure 3

shows a UMAP scatterplot[4] for questions on our test split of the estimated IRT parameters. In this case, instead of color-coding by difficulty as in Figure 2, we instead color-code by the absolute error between our predictions and the measured difficulty of each question. We again apply KMeans ($k = 10$) to our data with a smaller number of clusters due to the smaller size of the test set. We highlight CT.1, like C.1 and C.2 of Figure 2, this cluster consists primarily of yes/no questions. The difficulty in CT.1 has significantly smaller variance in the estimated difficulties than the rest of the clusters ($\rho = 0.02$). As well, the prediction error for CT.1 has significantly smaller variance ($\rho = 0.04$) and had the smallest average prediction error compared to the other clusters (0.68). This indicates that the model is able to recognize when question groupings, such as yes/no questions, have consistent difficulties (as discussed in 2.1) and has consistently lower error when predicting difficulty for these questions. However, the prediction error tends to vary more when the surface-level question types are not sufficient to characterize their difficulty.

We explore this further through a small counterfactual experiment. 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 selected an item with >2 absolute error to perform this experiment. The question we use in this study is: *Which university is this American philosopher, theologian, and Christian apologist who supports theistic science, professor at?* with an answer of *Biola University*. The predicted difficulty was -0.51. We found that simple changes to the question, such as using synonyms and removing unnecessary information, can increase the predicted difficulty up to -0.21. However, by modifying the answer (and by necessity the question) to be either a date or yes, we achieve a higher difficulty prediction (0.53 and 1.02, respectively). This further indicates the model's bias towards yes/no questions being of a higher difficulty regardless of the style or topic of question being asked. Some of our changes and their corresponding predictions are listed in Appendix E.

[4]Similar plots for the discrimination parameters are included in Appendix G

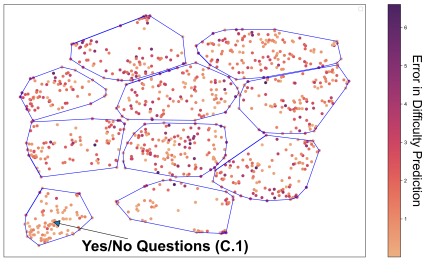

Figure 3: UMAP scatterplot of questions color coded by prediction error for difficulty. (Test set)

5 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. We note that our work is limited in two important ways: firstly, we only use the DFGN model in our artificial crowd, which may have introduced a bias in which some factors that make questions difficult/discriminatory are only applicable to this model. Secondly, we only explore the HotPotQA dataset, which may further limit our analysis to only be applicable to HotPotQA or similar datasets. Future work could incorporate multiple models and datasets to explore a more easily generalizable difficulty/discrimination prediction pipeline. We also note that our analysis here focused on QA. However, 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, automatically predicting 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.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. R2DE: a NLP approach to estimating IRT parameters of newly generated questions. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pages 412–421.

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.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4486–4503, Online. Association for Computational Linguistics.

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.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang,

Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1141–1158, Online. 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 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 the answer and question. A full list of these features can be found in Appendix F
  - **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 Appendix C.
- **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.

C Reading Difficulty Features

We list the reading difficulty features we used in our experiments and an overview of their calculations. Each calculation has its own coefficients 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)

D Significance Tests

We provide significance tests for the difficulty and discrimination predictions in tables 5 and 6. We see that the BERT features and using all features are able to beat the baseline with statistical significance ($\rho \leq .1$). 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	**0.034**
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 5: 1PL difficulty predictions. P-values for feature set performance (MSE) tested against the baseline.

E Counterfactual Results

- – Question (original): Which university is this American philosopher, theologian, and Christian apologist, who supports theistic science, professor at?'

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 6: 2PL discrimination 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.

  - Answer: "Biola University"
  - Pred. Diff: -0.51

- - Question : Which school is this philosopher and theologian who supports science, professor at?
  - Answer: "Biola University"
  - Pred. Diff: -0.21

- - 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: 0.53

- - 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

F 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 WN-Root
- 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

G Discrimination UMAP plots

In the following section, we provide the UMAP reduction plots for the discrimination parameters (darker being more discriminatory), as well as the prediction error UMAP plot for our best model (darker meaning higher error).

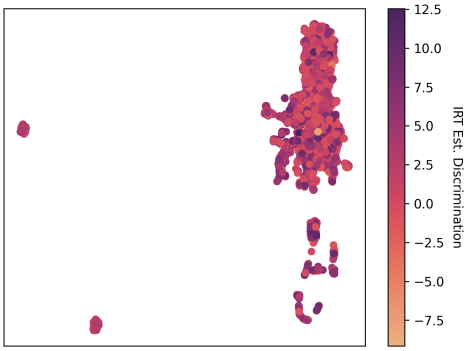

Figure 4: Answer BERT UMAP Reduction VS Discrimination values, train/dev set

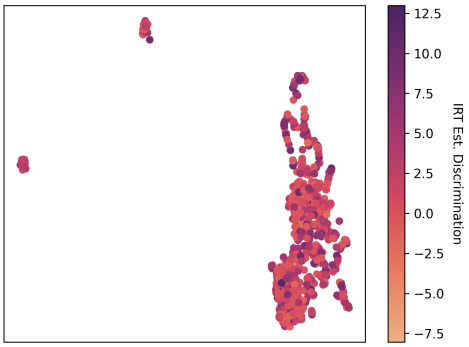

Figure 5: Answer BERT UMAP Reduction VS Discrimination values, test set

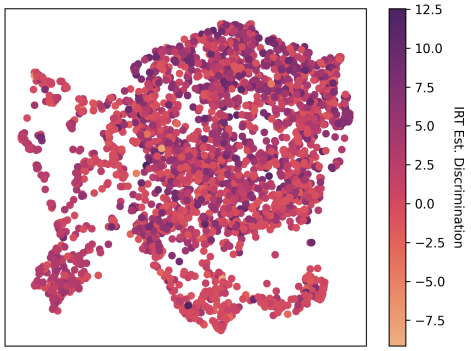

Figure 6: Question BERT UMAP Reduction VS Discrimination values, train/dev set

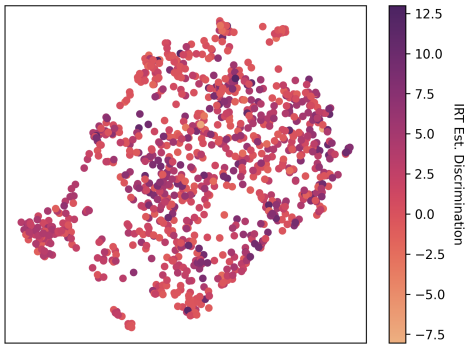

Figure 7: Question BERT UMAP Reduction VS Discrimination values, test set

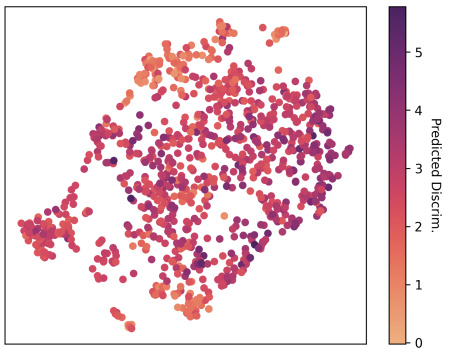

Figure 8: Question BERT UMAP Reduction VS Predicted Discrimination values, test set

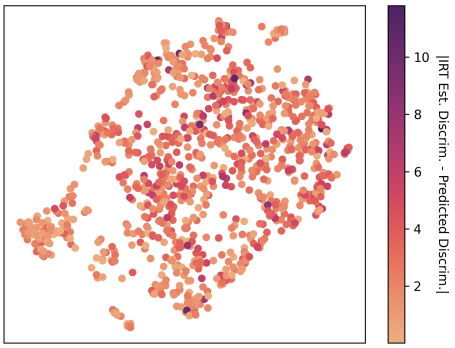

Figure 9: Question BERT UMAP Reduction VS Discrimination prediction error, test set