Question Difficulty Estimation from Text with Natural Language Processing Techniques

Parav Goyal · Pinaki Tyagi · Anshika Gupta

Abstract

In recent years, significant research efforts have been devoted to the task of Question Difficulty Estimation from Text (QDET) using Natural Language Processing (NLP) techniques. This research aims to address the limitations of traditional question calibration approaches. However, prior investigations have primarily focused on individual domains without conducting quantitative comparisons between different models from diverse educational domains. This work aims to bridge this gap by quantitatively analyzing various approaches proposed in previous research and comparing their performance on a publicly available real-world dataset containing Science MCQs. Our findings indicate that hybrid models tend to outperform single-feature-based models, linguistic features excel in reading comprehension questions, while frequency-based features (TF-IDF) yield better results in domain knowledge assessment.

Keywords: Question Difficulty Estimation, Natural Language Processing

1 Introduction

Estimating the difficulty of exam questions is crucial in educational settings, as it helps identify low-quality questions and those unsuitable for specific learner groups. Traditionally, pretesting and manual calibration have been used for this purpose, but they are time-consuming and expensive. Recently, research on Question Difficulty Estimation from Text (QDET) using NLP techniques has gained popularity, aiming to overcome the limitations of traditional approaches. By estimating question difficulty from the text at the time of question creation, the need for extensive pretesting and manual calibration could be reduced. However, most previous works focused on specific domains without comparing different approaches or datasets with various question types. In this study, we quantitatively evaluate previous QDET approaches and compare their performance on a publicly available dataset science MCQs (ARC). Our findings reveal that hybrid models often outperform single-feature-based ones, linguistic features excel in reading comprehension questions, while frequency-based features (TF-IDF) demonstrate better performance in domain knowledge assessment.

2 Related Works

While the field of question difficulty estimation has a long history, the adoption of NLP approaches in this task gained prominence only in recent times. A survey [1] discussed overarching trends in the domain, and another work [6] provided an overview of previous approaches with a proposed taxonomy for categorization. However, neither of these studies conducted a quantitative experimental evaluation.

Past research has primarily focused on two domains: language assessment, which evaluates learners' language proficiency, and content knowledge assessment, which

assesses knowledge about specific topics. Specifically, four papers concentrated on reading comprehension Multiple Choice Questions (MCQs), which are relevant to language assessment and consist of a reading passage, a stem, and answer options. The reading passage significantly influences question difficulty, and previous QDET models incorporated it in their estimation. In one approach [22], reading difficulty served as a proxy for question difficulty. In another study [5], five linguistic features were computed from the question and passage text, and their values were compared to a threshold to determine difficulty levels. Authors in [18] estimated difficulty using word2vec embeddings and a fully-connected neural network. Lastly, [21] explicitly considered the relationship between the reading passage and the question by employing an attention mechanism [29] to model sentence relevance in the passage.

The remaining literature focused on knowledge questions, which do not have accompanying passages and appear in both language and domain knowledge assessments. Most of these approaches involved a clear distinction between a feature engineering phase and a subsequent regression phase, with only a few utilizing end-to-end neural networks. the most commonly used machine learning regression algorithms are RandomForests [5,30,31], SVM [32,21], and linear regression [12,20,28], with someworks also experimenting with weighted softmax [36] and ridge regression [23].

A variety of features have been proposed in previous research, such as linguistic features [15,28], word2vec embeddings [16], frequency-based features [3], and readability indexes. Additionally, several papers explored hybrid approaches, combining some of these features. For instance, combinations of linguistic features and word2vec embeddings [42], linguistic and frequency-based features [36], frequency-based features, linguistic features, and readability indexes, and word embeddings, linguistic features, and frequency-based features [30,31] were examined.

Regarding end-to-end neural networks, Transformers and the attention mechanism were commonly used. The attention mechanism was applied in [6,24], BERT [11] was utilized in [33,3,27], and DistilBERT [34] in [2].

3 Evaluated Models

We experiment with i) linguistic features, ii) readability indexes, iii) TF-IDF (Term Frequency - Inverse Document Frequency),

Linguistic features: In previous studies [11, 20, 28], linguistic features have been employed in various forms to assess the number and length of words and sentences in questions, answer choices (for MCQs), and context (for reading comprehension questions). To carry out our research, we utilize seventeen linguistic features, drawing from existing literature, and incorporate them as input for a Random Forest regression model.

Readability indexes, which gauge the comprehensibility of a reading passage, have also been utilized in QDET research, as demonstrated in [17]. Inspired by previous work, we experiment with several readability indexes, including Flesch Reading Ease [13], Flesch-Kincaid Grade Level [19], ARI [25], Gunning FOG Index

[14], Coleman-Liau Index [8], Linsear Write Formula [20], and Dale-Chall Readability Score [10]. These indexes are used as input features for a Random Forest regression model.

Furthermore, frequency-based features, previously employed in [26, 5], are incorporated in our study using TF-IDF [22]. The TF-IDF weights represent the importance of a word (or a set of words) in a document within a corpus. This importance increases with the word's occurrences in the document while being tempered by its frequency across the entire corpus. To encode the questions, we consider three approaches: i) QO, focusing solely on the question, ii) QC, which appends the text of the correct option to the question, and iii) QA, concatenating all options (both correct and wrong) to the question; however, QC and QA are applicable only to MCQs. These TF-IDF features are then used as input for a Random Forest regression model.

In summary, we draw on previous research to utilize linguistic features, readability indexes, and frequency-based features with TF-IDF in our study, employing them as input for Random Forest regression models.

Word2vec [31] has been the most common technique for building word embeddings in previous research [14], therefore this is the non-contextualized word embedding technique we evaluate. We experiment with the same three approaches to create the embeddings as with TF-IDF (QO, QC, and QA), and use the word2vec features as input to a Random Forest regression model.Hybrid Approaches were also used in previous research, and they are all obtained by concatenating features from two (or more) of the approaches presented above, and using them as input to a single Random Forest regression model. Specifically, we evaluate i) linguistic and readability features [2,4,27], ii)linguistic, readability, and TF-IDF [4], iii) linguistic features and word embed-dings [32], iv) linguistic features, TF-IDF, and word embeddings [30,31]

4 Experimental Datasets

ARC [7] is a dataset of science MCQs, each being assigned a level between 3 and 9, which we use as the gold standard for QDET. The original dataset contains questions with a varying number of answer choices we only keep items with three distractors and one correct option; the resulting train, and test splits contain 3,358 and 3,530 questions respectively. ARC is very unbalanced: the two most common labels (8 and 5) appear about 1400 and 700 times respectively, the least common (6) around 100. Thus, we partially balance its train split, by randomly subsampling the two most common labels to keep only 500 questions for each of them; all the details and results shown here are for the balanced dataset.

5 Result

QDET is a sentence regression task, commonly evaluated by comparing the estimated values with the gold standard references. Here, we use the metrics that are most common in the literature: Root Mean Squared Error (RMSE) and R2 score. To study how stable each model is, we performfive independent training runs and show the mean and standard deviation of the metrics on the test set. It is important to remark here that although the difficulty in ARC are discrete levels, all the QDET models are trained as regression models and output a continuous difficulty, which we convert to one of the discrete values with

simple thresholds (i.e., mapping to the closest label)the model outperform the two baselines and Transformers are significantly better than the others according to all metrics; most likely, the attention can capture the relations between the passage and the question. As for the other features, the Linguistic Perform better than Readability, and TF-IDF. Hybrid models seem to bring some advantages: most of the combinations outperform the single features, and using a larger number of different features leads to greater advantages

Model	RMSE	R2 score
Linguistic	0.633 ±0.007	0.113 ±0.007
Readability	0.612 ±0.002	0.135 ±0.002
W2V	1.707 ±0.002	0.030 ±0.002
Ling.+Read.	0.592 ±0.004	0.157 ±0.004
W2V, Ling.	1.619 ±0.004	0.128 ±0.004
W2V QC, Ling., TF-IDF	1.583 ±0.005	0.166 ±0.006
Ling.+Read.+TF-IDF	0.602 ±0.003	0.146 ±0.003

6 Conclusions

In this study, we conducted a comprehensive quantitative analysis of previous Question Difficulty Estimation from Text (QDET) approaches to discern their comparative performance on a publicly available dataset. Through training the models on dataset, we noticed that Transformers exhibit superior performance even on smaller datasets (though we ensured datasets contained at least 4,000 questions). On the other hand, linguistic features proved effective for readability comprehension MCQs but showed limited efficacy for content knowledge assessment questions, where TF-IDF performed better.

The imbalance in difficulty labels affected the estimation process, indicating the need to strike a balance between class balancing and having sufficient questions for effective model training. Furthermore, future research could delve into analyzing the models' performance across different difficulty levels, potentially employing diverse techniques to map continuous estimation to discrete levels and employ scale linking to improve estimation accuracy across the entire range of difficulties.

7 References

1. AlKhuzaey, S., Grasso, F., Payne, T.R., Tamma, V.: A systematic review of data-driven approaches to item difficulty prediction. In: International Conference on Artificial Intelligence in Education. pp. 29–41. Springer (2021)

2. Benedetto, L., Aradelli, G., Cremonesi, P., Cappelli, A., Giussani, A., Turrin, R.: On the application of transformers for estimating the difficulty of multiple-choice questions from text. In: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 147–157 (2021)
3. Benedetto, L., Cappelli, A., Turrin, R., Cremonesi, P.: R2de: a nlp approach to estimating irt parameters of newly generated questions. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 412–421 (2020)
4. Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., Turrin, R.: A survey on recent approaches to question difficulty estimation from text. ACM Computing Surveys (CSUR) (2022)
5. Bi, S., Cheng, X., Li, Y.F., Qu, L., Shen, S., Qi, G., Pan, L., Jiang, Y.: Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 4645–4654 (2021)
6. Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., Ma, H., Hu, G.: Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 2397–2400 (2019)
7. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457 (2018)
8. Coleman, E.B.: On understanding prose: some determiners of its complexity. NSF Final Report GB-2604. Washington, DC: National Science Foundation (1965)
9. Culligan, B.: A comparison of three test formats to assess word difficulty. Language Testing 32(4), 503–520 (2015)
10. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin pp. 37–54 (1948)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (2019)
12. El Masri, Y.H., Ferrara, S., Foltz, P.W., Baird, J.A.: Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. The Curriculum Journal 28(1), 59–82 (2017)
13. Flesch, R.: A new readability yardstick. Journal of applied psychology 32(3), 221 (1948)
14. Gunning, R., et al.: Technique of clear writing (1952)
15. Hou, J., Maximilian, K., Quecedo, J.M.H., Stoyanova, N., Yangarber, R.: Modeling language learning using specialized elo rating. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 494–506 (2019)
16. Hsu, F.Y., Lee, H.M., Chang, T.H., Sung, Y.T.: Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. Information Processing & Management 54(6), 969–984 (2018)
17. Huang, Y.T., Chen, M.C., Sun, Y.S.: Development and evaluation of a personalized computer-aided question generation for english learners to improve proficiency and correct mistakes. arXiv preprint arXiv:1808.09732 (2018)
18. Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., Hu, G.: Question difficulty prediction for reading problems in standard tests. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

19. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
20. Klare, G.R.: Assessing readability. Reading research quarterly pp. 62–102 (1974)
21. Lin, L.H., Chang, T.H., Hsu, F.Y.: Automated prediction of item difficulty in reading comprehension using long short-term memory. In: 2019 International Conference on Asian Language Processing (IALP). pp. 132–135. IEEE (2019)
22. Manning, C.D.: Introduction to information retrieval. Syngress Publishing, (2008)
23. Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R.D., Brefeld, U.: Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. International Journal of Artificial Intelligence in Education 29(3), 342–367 (2019)
24. Qiu, Z., Wu, X., Fan, W.: Question difficulty prediction for multiple choice problems in medical exams. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 139–148 (2019)
25. Senter, R., Smith, E.A.: Automated readability index. Tech. rep., Cincinnati Univ OH (1967)
26. Settles, B., T. LaFlair, G., Hagiwara, M.: Machine learning–driven language assessment. Transactions of the Association for computational Linguistics 8, 247–263 (2020)
27. Tong, H., Zhou, Y., Wang, Z.: Exercise hierarchical feature enhanced knowledge tracing. In: International Conference on Artificial Intelligence in Education. pp. 324–328. Springer (2020)
28. Trace, J., Brown, J.D., Janssen, G., Kozhevnikova, L.: Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. Language Testing 34(2), 151–174 (2017)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
30. Yaneva, V., Baldwin, P., Mee, J., et al.: 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. pp. 11–20 (2019)
31. Yaneva, V., Baldwin, P., Mee, J., et al.: Predicting item survival for multiple choice questions in a high-stakes medical exam. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 6812–6818 (2020)
32. Yang, H., Suyong, E.: Feature analysis on english word difficulty by gaussian mixture model. In: 2018 International Conference on Information and Communication Technology Convergence (ICTC). pp. 191–194. IEEE (2018)
33. Zhou, Y., Tao, C.: Multi-task bert for problem difficulty prediction. In: 2020 International Conference on Communications, Information System and Computer Engineering (CISCE). pp. 213–216. IEEE (2020)