Towards the application of calibrated Transformers to the unsupervised estimation of question difficulty from text

Ekaterina Loginova
Ghent University, Ghent, Belgium
ekaterina.loginova@ugent.be

Luca Benedetto
Politecnico di Milano, Milan, Italy
luca.benedetto@polimi.it

Dries Benoit
Ghent University, Ghent, Belgium
dries.benoit@ugent.be

Paolo Cremonesi
Politecnico di Milano, Milan, Italy
paolo.cremonesi@polimi.it

Abstract

Being able to accurately perform Question Difficulty Estimation (QDE) can improve the accuracy of students' assessment and better their learning experience. Traditional approaches to QDE are either subjective or introduce a long delay before new questions can be used to assess students. Thus, recent work proposed machine learning-based approaches to overcome these limitations. They use questions of known difficulty to train models capable of inferring the difficulty of questions from their text. Once trained, they can be used to perform QDE of newly created questions. Existing approaches employ supervised models which are domain-dependent and require a large dataset of questions of known difficulty for training. Therefore, they cannot be used if such a dataset is not available (e.g. for new courses on an e-learning platform). In this work, we experiment with the possibility of performing QDE from text in an unsupervised manner. Specifically, we use the uncertainty of calibrated question answering models as a proxy of human-perceived difficulty. Our experiments show promising results, suggesting that model uncertainty could be successfully leveraged to perform QDE from text, reducing both costs and elapsed time.

1 Introduction

Question Difficulty Estimation (QDE), also known as "question calibration", is a crucial task in education. In Computerized Adaptive Testing (Linden et al., 2000), for instance, students are shown questions that are suitable for their skill level. When a question is miscalibrated (i.e. its difficulty has been erroneously estimated), it can be either too hard or too easy for a student, which would negatively affect their learning outcome (Wang, 2014). If the questions are too hard, students might get frustrated and lose motivation; if they are too easy, students are not adequately challenged (Papousek et al., 2016). In either case, their learning experience is worse than if the questions were of appropriate difficulty, which, especially in the context of large-scale online courses, gives rise to the methods of automated difficulty estimation.

Traditionally, QDE is performed either i) manually (Attali et al., 2014) or ii) with pretesting (Lane et al., 2015). Manual calibration involves one (or more) human experts labelling the question by selecting a numerical value representing its difficulty, a method that is subjective and not scalable. Meanwhile, pretesting involves deploying the new question in an exam as if it was a standard exam question, but without using it to assess students and without telling them that there is a question under pretesting. The other questions of the exams are then used to actually assess the students, and their answers help to calibrate the question under pretesting. This approach indeed leads to an accurate calibration, but it introduces a long delay between the time of question creation and when the new question can be used to assess students. Besides, it requires the new questions to be shown to students before being actually used for scoring them, which is undesirable.

Recent studies tried to address the limitations of the traditional approaches by performing QDE with Natural Language Processing (NLP) techniques (Ha et al., 2019; Qiu et al., 2019; Benedetto et al., 2020b, 2021; Xue et al., 2020; Huang et al., 2017). They are all based upon the same general idea: starting from a set of calibrated questions, we train a supervised machine learning model to infer the difficulty of questions from their text. Once the model is trained, it is used to calibrate newly-generated questions, overcoming (or at least reducing) the need for pretesting and manual calibration. Although these techniques were proposed to enable an immediate calibration of new questions, they

Proceedings of Recent Advances in Natural Language Processing, pages 846–855
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_097

have two major limitations due to their supervised manner: i) they require thousands of calibrated questions as a training set, and ii) they cannot perform cross-domain QDE. In other words, the training questions must assess the same topics as the new questions, which the model will later be used on, thus limiting its applicability (e.g. when introducing new courses in an e-learning platform). These limitations are intrinsic to such approaches and cannot be addressed by improving the accuracy of the models.

In this work, we explore an approach that is a total shift in the paradigm of QDE from text and which could potentially overcome both limitations. The intuition is to build an end-to-end Question Answering (QA) model that answers Multiple Choice Questions (MCQs) and use its uncertainty (which can be interpreted as representing the machine-perceived difficulty) as a proxy for human-perceived difficulty, which is the final target of the estimation. Previous research already hypothesized that there might be a relation between human-perceived difficulty and machine-perceived difficulty (Ha et al., 2019), but leveraged the machine-perceived difficulty as a feature for a supervised model, thus facing the same limitations as the approaches to QDE from text mentioned above. On the contrary, the approach we propose here performs QDE by leveraging the confidence of a trained QA model: specifically, we compute the variance of the probability distribution over the possible answer choices of the MCQ under calibration. Crucially, this approach is model agnostic and can be used on any QA model that outputs scores for the possible choices of an MCQ. Architecture-wise, in this study, we experiment with Transformers (Vaswani et al., 2017) for QA. In order to understand how well the proposed approach performs with different QA models, we use three different Transformers: BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and XLNet (Yang et al., 2019). We choose them because they offer variety and differ in sizes, and were all demonstrated to perform well in several language understanding tasks.

However, large neural classification models tend to be overconfident in their predictions (Desai and Durrett, 2020), which might hinder the possibility of using their uncertainty as a proxy for question difficulty. In order to understand whether this is really an issue for unsupervised QDE from text and to explore possible solutions, we also experiment with calibrated QA models. Calibrating a model involves aligning the posterior probabilities with the empirical likelihoods (Guo et al., 2017). For instance, if we consider all the predictions for which a model has the confidence of 75%, the true accuracy must be 75% if the model is perfectly calibrated. We remark here that calibration and accuracy are not directly related: in fact, a model might be accurate but not calibrated or, on the other hand, not very accurate but well-calibrated. Calibration can be intuitively interpreted as the "awareness" of the model of its capabilities. In practice, several techniques can be used for calibrating neural models, and they will be discussed in Section 3. For simplicity and computational reasons, we use the ensembling technique.

We experiment on the large question-answering dataset RACE (Lai et al., 2017) to assess the QDE capabilities of the proposed approach. Specifically, we evaluate it on the task of pairwise difficulty prediction: given a pair of questions from RACE, the objective is to indicate which question of the pair is more difficult. As the gold standard for the difficulty, we use i) the difficulty level available in RACE and ii) additional human labels obtained with crowd-sourcing. The experimental results are promising and suggest that the proposed approach could be used to perform QDE from text in an unsupervised manner leveraging the uncertainty of QA models. We also show that choosing the underlying QA model is not straightforward, and the best results are obtained leveraging models that are both accurate and calibrated.

Our contributions are as follows: i) we propose an unsupervised way for QDE from text that does not require answer logs or question difficulty labels, only the text of the MCQ and the possible choices, ii) we experiment with modern Transformer-based architectures to demonstrate the viability of the proposed approach.

We share our code publicly[1].

2 Related Work

The earliest research about QDE from text focused on MCQs, using bag-of-words and the similarities between question, correct choice, and distractors (incorrect choices) for the estimation (Alsubait et al., 2013; Ha and Yaneva, 2018; Kurdi et al., 2016; V and Puligundla, 2015). However, they are

[1] https://bit.ly/3b4tPLN

generally outperformed by the more recent models, which are based on machine learning techniques.

Ha et al. (2019) introduced a model to estimate from a question text its *correctness*, which is defined as the fraction of students who correctly answered the question. This model was trained using question texts and a large dataset of medical documents (i.e. books, papers). Similarly, the model proposed by Qiu et al. (2019) is trained on a dataset of medical documents and question texts to estimate the *wrongness* of newly generated questions. Benedetto et al. (2020b,a) proposed R2DE, a model that estimates the difficulty of newly generated MCQs, using as input only the text of the questions and the text of the possible choices, without any additional data. Xue et al. (2020) explored the effects of transfer learning for question calibration from text. Specifically, the authors fine-tune pre-trained ELMo embeddings (Peters et al., 2018) for the task of response time prediction and subsequently perform a second fine-tuning for the task of QDE. Lastly, Huang et al. (2017) propose a neural model for the estimation of the difficulty of reading comprehension questions.

From a high-level perspective, all these models are based on the same idea: the real question difficulty (obtained either with pretesting or manual calibration) is used as the target value for training a supervised machine learning model that performs QDE from text for newly generated questions. The downside of this approach is that it can work only as long as the new questions belong to the same domain as the training questions. Moreover, such models need a large number of calibrated questions for training, which might be too costly to obtain, especially for smaller institutions. (Wang et al., 2014) employs a pairwise difficulty comparison scheme similar to the one we will, but they still require the user responses for the algorithm to work, same as (Narayanan et al., 2017).

Motivated by this, in this work, we explore the possibility of completely shifting the paradigm for the task of QDE from text. The proposed approach leverages the uncertainty of a QA model as a proxy for question difficulty and uses this model-perceived difficulty to calibrate newly generated questions. This unsupervised approach does not require ground truth difficulty labels, but only the text of the questions and, in the case of MCQs, of the possible answer choices.

3 Calibration of Neural Models

Several techniques exist to calibrate neural network classifiers. The most popular ones are the following: i) *vanilla*: maximum softmax probability, which usually does not lead to calibrated classifiers (Hendrycks and Gimpel, 2017). ii) *Temperature scaling*: a posterior calibration technique using a validation set (Guo et al., 2017; Desai and Durrett, 2020). iii) *Bayesian deep learning*, which requires alterations to the training procedure and is computationally expensive. iv) *Ensembles*: consists in independently training M models on the entire dataset using different random initializations (Lakshminarayanan et al., 2017) or dropout (Gal and Ghahramani, 2016; Srivastava et al., 2015) and averaging their predictions.

Focusing on pre-trained Transformers, previous research (Desai and Durrett, 2020) showed that pre-trained BERT is fairly well-calibrated for in-domain tasks, but it is miscalibrated for out-of-domain tasks. To the best of our knowledge, no previous research experimented with the calibration of DistilBERT and XLNet.

In this work, we use deep ensembles with M equal to 3, since we observed that this approach led to fairly well-calibrated models for all the Transformer architectures under evaluation. Additionally, we also experimented with ensembles made of the combination of models with different architectures (e.g. BERT with XLNet, etc.).

Considering the metrics existing to evaluate model calibration, one of the most commonly used approaches is the Expected Calibration Error (ECE) (Naeini et al., 2015), which compares the confidence and the accuracy of the model. More precisely, it defines miscalibration as the difference in expectation between confidence and accuracy. Thus, ECE approximates the miscalibration by partitioning the predictions in a number M of bins and averaging the difference between the accuracy and confidence obtained in each bin.

4 Models

The approach proposed in this paper leverages the confidence of a QA model to perform QDE for MCQs in an unsupervised manner. Specifically, it leverages the output scores, one for each possible answer choice, produced by the model for each question, and the only requirement is that such output scores represent a probability distribution (i.e. they sum to 1, such as the softmax scores of

a neural network). We experiment here on MCQ with four possible choices, but the approach can be easily scaled to MCQ with different numbers of answer choices. It is important to note that this approach to unsupervised QDE from text is "QA model agnostic", in the sense that it does not need any information about the underlying model but only the scores produced by it. We leverage the softmax scores produced by the underlying QA model to measure the model-perceived question difficulty, which is considered a proxy for human-perceived difficulty. In practice, we convert the raw softmax scores of each question into a single numerical value by computing their variance. We assume that larger values of variance indicate easier questions (since it means that the model is more certain in the estimation).

We also experimented with some alternatives, but they were generally outperformed by using the score variance (although the difference was not major). Specifically, we also experimented with: i) keeping only the highest softmax score, and ii) computing the difference between the highest and the second-highest softmax score.

In order to understand how the proposed approach performs with softmax scores from different underlying QA models, we experiment with three Transformer models: i) BERT (Devlin et al., 2019), ii) DistilBERT (Sanh et al., 2019), and iii) XLNet (Yang et al., 2019). The details of each architecture are beyond the scope of this research; we refer to the original papers for their description. For our study, it is essential to know that they are all publicly available neural network models, which are pre-trained on several language understanding tasks, and that they can be fine-tuned for different downstream tasks with minimal changes to the architecture. Precisely, in this study, we add a multiple-choice classification layer on top of each original model, which is a common task in the literature. The three models have diverse sizes and architectures, even though DistilBERT is strongly related to BERT since it is obtained from it by using knowledge distillation (Hinton et al., 2015) to reduce the number of hidden layers.

We implement all the models with the HuggingFace transformers library (Wolf et al., 2019), using in all cases the pre-trained *base-cased* version. We fine-tune them using Google Colab GPUs. The parameter configuration for the QA models is taken from the literature; the specific values are shown in Table 1. The table also shows, for each model, the accuracy we obtain in the QA task on RACE, which is in line with previously reported results.

Parameter	BERT	DistilBERT	XLNet
input text len.	256	512	256
learning rate	2e-5	5e-5	2e-5
Adam epsilon	1e-6	1e-8	1e-6
weight decay	0.05	0.05	0.01
n. epochs	2	3	2

Table 1: Training configuration of the QA models.

As suggested by (Lakshminarayanan et al., 2017), we ensemble models to reduce miscalibration. In practice, we proceed as follows. First, we i) train five instances for each architecture (i.e. BERT, DistilBERT, XLNet), and each instance is trained on the entire training dataset (randomly shuffled), with a different random initialization. Then, we ii) pick the three best performing instances of each architecture, considering the test accuracy on the QA task. Lastly, we iii) build the ensembles by averaging (separately for each test question) the softmax scores produced by the three instances of each architecture so that each of the four answer options is assigned a single score from 0 to 1. These scores indicate the probability (according to the model) of each option being correct. In addition to building an ensemble for each architecture, we also build "hybrid" ensembles in the same way but averaging the predictions of instances of different architectures.

5 Data

In this study, we use the reading comprehension RACE dataset[2] and two datasets derived from it, which contain pairs of questions and a label indicating which question of the pair is more difficult. The entire RACE dataset is used to train the QA models, while the two other datasets are used for the experiments on QDE from text.

5.1 RACE

The original RACE dataset contains 25,000 passages in English from middle and high school reading comprehension exams, with four MCQ associated with each text (100,000 questions in total). All the questions are MCQ with four possible choices (see an example in Figure 1). The questions are

[2] www.cs.cmu.edu/~glai1/data/race/

Bungee jumping is an activity about jumping from a tall structure while connected to a large elastic cord . The tall structure is usually a fixed object, such as a building, bridge or crane; but it is also possible to jump from a movable object, such as a hot-air balloon or helicopter...

Question 1

Which of the following is NOT suitable for bungee jumping?

A) The fixed-wing aircraft
B) The helicopter
C) The hot-air balloon
D) The mobile crane

Answer: A

Figure 1: An example question from RACE.

designed to require more extensive reasoning skills than other QA datasets such as SQuaD (Rajpurkar et al., 2016) and differ in reasoning types required to answer them. For example, they cover passage summarization and attitude analysis, which means that the answer cannot always be extracted directly from the passage. As a result, neural methods have a significant performance gap compared to humans: 66.7% (with XLNet-base) and 94.5% accuracy, respectively. The questions can be separated into two types based on their syntax: interrogative or cloze. Interrogative questions are the ones that end with a question mark, while cloze questions contain a gap that has to be filled in.

5.2 PairRACE_HM

For constructing this dataset, we use the *level* label available in RACE, which indicates the level of examination (high or middle) of each question. Specifically, we use *level* as an indication of question difficulty and prepare 2,062,096 pairs of question, such that each pair contains one *middle* question and one *high* question (related to different passages). This dataset is then used to evaluate the proposed approach in the task of pairwise difficulty estimation: basically, given a pair of question, we check whether the proposed approach labels the *high* question as being the more difficult one.

5.3 PairRACE_CS

The *level* label does not contain a numerical estimation of question difficulty, and there is no way of knowing how much harder the *high* questions are. Therefore, we also build a dataset to evaluate how well the proposed approach performs at a more focused level difficulty estimation, which differentiates the difficulty of the question within the *middle* level and within the *high* level. Such fine-grained information about the difficulty is not available in RACE; thus we manually annotate a subset of the question pairs by crowd-sourcing on the Amazon Mechanical Turk platform.

First, 80 pairs of questions from the test set were randomly chosen (both questions in each pair have the same level and correspond to the same passage, with 77 unique passages used in total). An annotator was then presented with a passage and a pair of questions, along with their answer options, and asked to identify the more difficult question. Each question was first labelled by two of the authors (with Cohen's kappa of 0.30) and then passed to turkers. The only condition imposed on turkers was to be native English speakers. Each question was answered by one crowd-worker; thus, in combination with our labels, we obtained three labels per question pair. To encourage a thoughtful approach, we added an obligatory field in which we asked the turkers to provide a brief motivation for their choice. An example reasoning is the following: "Question 1 requires you to think beyond the passage content, to extrapolate and predict the next step, while 2 just asks to give a title to the content." There were multiple recurring indicators of how humans estimate difficulty. As expected, the questions which require searching for a named entity or finding a simple fact statement are considered simpler, while summarising information or giving a title is more challenging. The text's location contributes as well – it is easier to answer the question if the cue can be found in the first or the last sentences of the passage. Ultimately, the Fleiss' kappa[3] agreement was 0.21, which is "fair", according to guidelines by (Landis and Koch, 1977). Still, we recognise the possible issues with a relatively low agreement, and, in the experimental evaluation, we separately consider the performance on the question pairs with full agreement.

6 Experiments

The goal of our experiments is to perform unsupervised QDE from text, using only the softmax scores produced by QA models. More precisely, we do not estimate directly question difficulty but evaluate the proposed approach on the task of pairwise difficulty prediction: given a pair of questions, the task consists in identifying the one that is more difficult. The manual labels are obtained, also within the pairwise comparison framework. This way,

[3] is used when the annotators are drawn from a random distribution

Model	QA Accuracy eval	QA Accuracy test	ECE test
BERT (0)	0.64	0.62	0.15
BERT (3)	0.64	0.62	0.14
BERT (42)	0.64	0.62	0.14
DistilBERT (1)	0.48	0.46	0.03
DistilBERT (3)	0.44	0.42	0.02
DistilBERT (42)	0.50	0.48	0.14
XLNet (2)	0.66	0.65	0.15
XLNet (3)	0.66	0.66	0.15
XLNet (4)	0.67	0.65	0.15
BERT (E)	0.66	0.63	0.10
DistilBERT (E)	0.49	0.47	0.07
XLNet (E)	0.68	0.66	0.11
BERT-DB (E)	0.64	0.62	0.05
BERT-XLNet (E)	0.48	0.47	0.09
DB-XLNet (E)	0.33	0.32	0.21
BERT-DB-XLNet (E)	0.56	0.55	0.06

Table 2: Evaluation of QA accuracy and calibration of the underlying models used for QDE from text, both single instances and ensembles. In the hybrid ensembles, "DB" means "DistilBERT".

we investigate i) whether the machine uncertainty leads to a notion of difficulty that aligns well with the human one – which is represented by the *level* and the crowdsourced labels – and ii) whether it can be useful in practical applications when logs of answers or calibrated questions are not available for training.

6.1 Evaluating QA accuracy and calibration

Before actually evaluating the proposed approach on the task of pairwise difficulty prediction, we evaluate the QA accuracy and the calibration of the underlying models to explore the relations between these and the accuracy in the task of pairwise difficulty prediction. Results are shown in Table 2.

For QA, we use accuracy as a metric (the higher, the better), while for calibration, we use Expected Calibration Error[4] (ECE, the lower, the better). It should be noted that ECE is used to diagnose whether the model's confidence can be a reliable proxy for difficulty or not.

For each underlying model (i.e. BERT, DistilBERT, XLNet), we present the results for three single instances and the calibrated ensemble. The single instances are identified by a number, which is the random seed used during training. While the specific value of the random seed is not meaningful in itself for this analysis, we use it to distinguish between different instances of the same architecture. Ensembles are indicated by "E".

As for the QA accuracy, we can see that, considering the same architecture (e.g. BERT) the accuracy of the ensembles is generally better than the single models, both on the *eval* dataset and the *test* dataset. However, this is not true for the "hybrid" ensembles, which perform worse than the model they are obtained from.

Similar results can also be seen for model calibration: ensembles generally have lower ECE, meaning that they are better calibrated. Indeed, BERT (E) has an ECE of 0.10, while the average of the single models is over 0.14, and XLNet (E) has an ECE of 0.11, the average of the single instances being 0.15. This trend is not as visible for DistilBERT, as the ensemble model has an ECE only slightly higher than the average of the single instances (0.07 compared to 0.06).

6.2 Pairwise difficulty prediction

Given as input a question pair, the unsupervised pairwise difficulty prediction task consists of predicting which question of the pair is more difficult. We evaluate the proposed approach using different underlying models (both single instances and calibrated ensembles), and compare it with three baselines based on ELMo (Peters et al., 2018). It should be noted that internally there is no difference indicated between comprehension and knowledge questions and each model is thus evaluated on the same subset.

i) ELMo_C (comprehension): we calculate the cosine distance between the ELMo embeddings of the question and of the passage; the question with the larger distance is labelled as more difficult.

ii) ELMo_K (knowledge): we calculate the average distance between the correct answer option and the distractors; the more difficult question has the lowest distance. This is a standard approach in the literature (Alsubait et al., 2013; Kurdi et al., 2016).

iii) ELMo_QA: a QA model built upon ELMo. Given a passage and an MCQ, it selects the answer by picking the choice which has the lowest cosine distance to the passage. It produces one score for each possible choice, and we use our approach directly on these scores (after normalization).

As introduced in Section 4, the proposed ap-

[4] calculated with pypi.org/project/netcal/netcal

	PairRACE_HM			PairRACE_CS			
	All (n=2M)	CA		All (n=80)	TA (n=37)	TA & CA	
Model	Acc.	n	Acc.	Acc.	Acc.	n	Acc.
Random	0.50	-	-	0.50	0.50	-	-
ELMo_{C}	0.57	-	-	0.65	0.76	-	-
ELMo_{K}	0.57	-	-	0.50	0.57	-	-
ELMo_{QA}	0.55	0.2M	0.57	0.45	0.41	3	-
BERT (0)	0.60	0.8M	0.62	0.55	0.54	20	0.50
BERT (3)	0.60	0.8M	0.61	0.60	0.65	19	0.63
BERT (42)	0.62	0.8M	0.62	0.57	0.59	15	0.53
DistilBERT (1)	0.60	0.5M	0.59	0.51	0.46	6	-
DistilBERT (3)	0.56	0.4M	0.58	0.52	0.51	5	-
DistilBERT (42)	0.60	0.5M	0.61	0.49	0.46	9	-
XLNet (2)	0.57	0.9M	0.60	0.62	0.65	20	0.60
XLNet (3)	0.58	0.9M	0.60	0.60	0.65	21	0.71
XLNet (4)	0.57	0.9M	0.59	0.62	0.68	19	0.68
BERT (E)	0.60	0.9M	0.61	0.59	0.65	19	0.63
DistilBERT (E)	0.58	0.5M	0.58	0.49	0.51	5	-
XLNet (E)	0.57	0.9M	0.60	0.61	0.65	19	0.68
BERT-DistilBERT (E)	0.59	0.9M	0.60	0.49	0.54	19	0.63
BERT-XLNet (E)	0.57	0.5M	0.55	0.57	0.41	8	-
DistilBERT-XLNet (E)	0.56	0.2M	0.55	0.56	0.46	2	-
BERT-DistilBERT-XLNet (E)	0.58	0.7M	0.56	0.56	0.43	8	-

Table 3: Evaluation of pairwise difficulty prediction on PairRACE_HM and PairRACE_CS. For PairRACE_HM we separately present the accuracy i) on the whole dataset (2M pairs of questions), and ii) on the questions which were Correctly Answered (CA) by each model, showing the number of question pairs (n). For PairRACE_CS we separately present the accuracy i) on the whole dataset (80 pairs of questions), ii) on the question pairs with Total Agreement (TA) between the human annotators (37 pairs of questions), and iii) on the question pairs with Total Agreement which were Correctly Answered by each model (TA & CA).

proach consists of converting the raw softmax scores of the QA models into a unique value that can be used for pairwise difficulty prediction. This is done by computing the variance of the raw softmax scores and assuming that the question with the lower variance is the more difficult one.

Table 3 presents the results obtained on PairRACE_HM and PairRACE_CS, using accuracy as evaluation metric. Each row shows the results for a different model, and we can identify three groups: the baselines, the single instances, and the ensembles.

6.2.1 High vs middle (PairRACE_HM)

The columns on the left present the results obtained using the *level* label as ground truth difficulty. We separately present the accuracy i) on all the question pairs and ii) on the pairs in which both questions were Correctly Answered (CA) by the QA models; n is the number of question pairs.

Considering all the pairs, we can see that the proposed approach consistently performs at least as well as the baselines, both when using the single models and when using the ensembles, although the improvement is not major. The only exceptions are DistilBERT (3) and DistilBERT-XLNet (E).

Interestingly, if we compare Table 3 and Table 2, we can see that there is no clear correspondence between the accuracy in the QA task and the accuracy in pairwise difficulty prediction or between the ECE and the accuracy in pairwise difficulty prediction. For instance, DistilBERT (3) has the lowest (i.e. best) ECE, but it is outperformed by all the ensembles (except the hybrid ones) in the pairwise difficulty prediction task. This suggests that the calibration of the QA model is not the only factor to take into consideration when using its uncertainty for QDE and that its QA accuracy also has an important role.

Comparing the accuracy of the ensembles and the single models, we observe that there is no appar-

ent improvement in using calibrated ensembles. It is especially noticeable when considering only the questions that the models correctly answered. However, in a real-world unsupervised scenario, the true difficulty labels are not available for choosing the best performing model with cross-validation, and neither the accuracy in the QA task nor the ECE is sufficient to pick the best performing model. Therefore, we argue that the usage of calibrated ensembles is a better solution as it allows to avoid the oscillations of single instances (e.g. from 0.56 to 0.60 of the single DistilBERT models against the 0.58 of the ensemble). However, this is true only for ensembles of models with the same architecture. Hybrid ensembles did not lead to better performance; thus, we argue that they should not be used for the task of unsupervised QDE from text.

6.2.2 Crowdsourced labels (PairRACE_CS)

Moving to the right side of the table, we consider the crowdsourced difficulty as ground truth, and we present the results separately for i) the whole dataset, ii) the pairs of questions with Total Agreement between human annotators (TA), and iii) the pairs of questions with Total Agreement and Correctly Answered by each model (TA & CA).

The most crucial difference is that the best performing model, in this case, is ELMO_C. It means that leveraging the similarity between the provided document and the questions might be a good alternative to using the uncertainty of the QA models for comprehension MCQ. This is reasonable since the goal of comprehension questions is to find the answer to the question in the accompanying passage. However, this is not in agreement with the results obtained on PairRACE_HM. Also, we have to consider that PairRACE_CS is made of only 80 question pairs (37 when considering only the ones with total agreement) while PairRACE_HM contains 2M pairs; therefore the performance of ELMO_C is worth of further exploration. Moreover, BERT (E) and XLNet (E) clearly outperform ELMO_K and ELMO_QA, suggesting that indeed the uncertainty of accurate and calibrated QA models can be beneficial for QDE of knowledge questions (which are not provided with an accompanying passage that contains the answer).

Differently from the previous experiment, the performance of DistilBERT (E) here is clearly worse than the other ensembles (it is even worse than random), thus suggesting that – being a smaller model – they are not capable of modelling the questions as well as BERT (E) and XLNet (E).

Except for these two differences, the rest of the findings is fairly similar to PairRACE_HM, and the ensembles generally outperform the single instances of the same architecture. Crucially, the accuracy of all ensemble models (except the hybrid ones) is higher for pairs with the total agreement, thus supporting the claim that the uncertainty of QA models could really be used for unsupervised QDE from text. It is also interesting to remark that this is not always the case for the single instances of Transformers, which sometimes have worse accuracy on the pairs with total agreement. This, once again, suggests that calibrated ensembles are more suitable for unsupervised QDE from text.

Considering the pairs with the total agreement and containing questions correctly answered by the QA models (shown in column TA & CA[5]), we can see that the correctness of the QA model does not seem to have a significant impact on the accuracy. However, there are only a few question pairs of this type; therefore we cannot perform any relevant observations.

7 Conclusions

The results of this research support the idea that it is possible to estimate the human-perceived difficulty of exam questions via the uncertainty scores produced by Question Answering (QA) neural networks. The advantage of the approach we propose is that it is possible to predict the relative difficulty of questions across different domains without needing any calibrated questions or logs of students' answers. For training the QA model, it is sufficient to have access to i) the corpus of questions and (possibly) ii) the learning materials.

As a practical guideline, both the QA accuracy and the calibration seem to impact the accuracy of QDE. Therefore we believe that it is better to use calibrated models which are powerful enough to reach decent performance in the QA task (e.g. the BERT and XLNet ensembles used here).

Future work will focus on exploring whether improving the calibration of the QA models (e.g. increasing the number of models in the ensemble or using Bayesian neural networks) would lead to improved results in unsupervised QDE from text and will analyze the effects of combining raw softmax scores with different techniques. Weighing the

[5] We do not show the accuracy unless there are at least 10 question pairs correctly answered

scores by the accuracy of each intervening model is a way to explore whether hybrid ensembles can still improve the performance, but we leave the implementation of this approach for another study.

Another natural progression of this work is to leverage question-specific information, such as the reasoning type required to answer it, or the question format (e.g. cloze items vs interrogative items). Using the text of the possible choices might improve the accuracy, which makes it an interesting modification to explore in future studies. Further experimental investigations are also needed to use the proposed approach for creating a difficulty ranking of questions.

References

Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2013. A similarity-based theory of controlling MCQ difficulty. In *2013 second international conference on e-learning and e-technologies in education (ICEEE)*, pages 283–288. IEEE.

Yigal Attali, Luis Saldivia, Carol Jackson, Fred Schuppan, and Wilbur Wanamaker. 2014. Estimating item difficulty with comparative judgments. *ETS Research Report Series*, 2014(2):1–8.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. 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*, pages 147–157.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020a. Introducing a Framework to Assess Newly Created Questions with Natural Language Processing. 12163:43–54.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020b. 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.

Shrey Desai and Greg Durrett. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. pages 4171–4186.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1050–1059.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. 70:1321–1330.

Le An Ha and Victoria Yaneva. 2018. Automatic Distractor Suggestion for Multiple-Choice Tests Using Concept Embeddings and Information Retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*, pages 389–398. Association for Computational Linguistics.

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, BEA@ACL 2019, Florence, Italy, August 2, 2019*, pages 11–20. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for READING problems in standard tests. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1352–1359. AAAI Press.

Ghader Kurdi, Bijan Parsia, and Uli Sattler. 2016. An Experimental Evaluation of Automatically Generated Multiple Choice Questions from Ontologies. In Mauro Dragoni, María Poveda-Villalón, and Ernesto Jiménez-Ruiz, editors, *OWL: - Experiences and Directions - Reasoner Evaluation - 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, Italy, November 20, 2016, Revised Selected Papers*, volume 10161 of *Lecture Notes in Computer Science*, pages 24–39. Springer.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. pages 785–794.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Suzanne Lane, Mark R Raymond, and Thomas M Haladyna. 2015. *Handbook of test development*. Routledge.

Wim J Linden, Wim J van der Linden, and Cees AW Glas. 2000. *Computerized adaptive testing: Theory and practice*. Springer.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press.

Sankaran Narayanan, Vamsi Sai Kommuri, N. Sethu Subramanian, Kamal Bijlani, and Nandu C. Nair. 2017. Unsupervised Learning of Question Difficulty Levels Using Assessment Responses. In *Computational Science and Its Applications - ICCSA 2017 - 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part I*, volume 10404 of *Lecture Notes in Computer Science*, pages 543–552. Springer.

Jan Papousek, Vít Stanislav, and Radek Pelánek. 2016. Impact of question difficulty on engagement and learning. In *Intelligent Tutoring Systems - 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings*, volume 9684 of *Lecture Notes in Computer Science*, pages 267–272. Springer.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2019. Question difficulty prediction for multiple choice problems in medical exams. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 139–148. ACM.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2377–2385.

Vinu E. V and Sreenivasa Kumar Puligundla. 2015. A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *J. Web Semant.*, 34:40–54.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Quan Wang, Jing Liu, Bin Wang, and Li Guo. 2014. A Regularized Competition Model for Question Difficulty Estimation in Community Question Answering Services. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1115–1126. ACL.

Tzu-Hua Wang. 2014. Developing an assessment-centered e-learning system for improving student learning effectiveness. *Comput. Educ.*, 73:189–203.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2020, Online, July 10, 2020*, pages 193–197. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.