Exploration of Annotation Strategies for Automatic Short Answer Grading

Aner Egaña, Itziar Aldabe[0000−0003−2017−2910], and Oier Lopez de Lacalle[0000−0003−4969−2055]

HiTZ Center - University of the Basque Country UPV/EHU, San Sebastian, Spain
{aegana028,itziar.aldabe,oier.lopezdelacalle}@ehu.eus

Abstract. Automatic Short Answer Grading aims to automatically grade short answers authored by students. Recent work has shown that this task can be effectively reformulated as a Natural Language Inference problem. State-of-the-art is defined by the use of large pretrained language models fine-tuned in the domain dataset. But how to quantify the effectiveness of the models in small data regimes still remains an open issue. In this work we present a set of experiments to analyse the impact of different annotation strategies when not enough training examples for fine-tuning the model are available. We find that when annotating few examples, it is preferable to have more question variability than more answers per question. With this annotation strategy, our model outperforms state-of-the-art systems utilizing only 10% of the full-training set. Finally, experiments show that the use of out-of-domain annotated question-answer examples can be harmful when fine-tuning the models.

Keywords: Automatic Short Answer Grading · Natural Language Processing · Natural Language Inference · Transfer Learning.

1 Introduction

Automatic content scoring is an important application in the area of automatic educational assessment. In this context, the evaluation of short answers authored by students is referred to as Automatic Short Answer Grading (ASAG) and the available datasets usually consist of questions, reference answers and student answers. Current state-of-the-art in Natural Language Processing (NLP) has shown that task reformulation (e.g., transforming specific tasks as Natural Language Inference or Question Answering) is an effective way to transfer knowledge across tasks and improve results [16, 15]. Similarly, recent work in ASAG has demonstrated that reformulating the ASAG as an entailment problem is an effective method to obtain strong results [3].

Methods that fine-tune large pretrained language models (LM) with large amounts of labelled data have established the state-of-the-art [10]. Nevertheless, due to differing languages, topic of questions, grading scale and the cost of human annotation, there is typically only a small number of labelled examples in real-world applications —and these models perform poorly. As an alternative, methods that require few (few-shot) or no (zero-shot) examples have emerged. Still, the way we should select training examples is an open question in ASAG. In this paper we focus on using entailment models to explore zero- and few-shot learning in student short answer grading. We define different scenarios where we assume there are no sufficient training examples for fine-tuning the model, and pose the following research questions in order to devise better strategies for data annotation:

RQ1 Having a task-agnostic generic entailment model, what would be the best way to annotate data and how much data would be needed to obtain state-of-the-art results?

RQ2 Can we effectively transfer task knowledge to new domains? And similarly, having a NLI-based fine-tuned model in one domain, how much data do we need to be annotated in a new one?

We attempt to answer the questions stated above empirically conducting experiments in the Semeval-2013 SRA dataset [5] and make the following contributions: 1) We show that annotation strategy can have a significant impact on results. This is because the annotation that increases the variability on the question side, at the cost of decreasing the amount of annotated answers per question, is preferable to having the same number of annotated examples with fewer questions and more answers. 2) Reformulating ASAG as an entailment problem and fine-tuning an entailment model allows us to obtain state-of-the-art results. 3) Related to this, we demonstrate that zero-shot entailment models can perform close to state-of-the-art results.4) We illustrate that the impact of the domain can be larger than the knowledge that can be acquired from the task. That is, using a generic entailment model is more effective than fine-tuning it with out-of-domain examples.

2 Related Work

Current approaches for ASAG can be categorized into three types [8]: 1) Hand-engineered feature-based machine learning (ML) approaches, which still get competitive results, 2) supervised deep learning approaches that fit parameters directly from training, and 3) large pretrained language models fine-tuned on the target task.

Hand-crafted features rely on the extraction of features from the questions and reference and student answers in order to find lexical, syntactic and semantic similarities between the student answer with the reference answer [12, 9].

Deep-learning models contributed to significantly improving results in ASAG. They provide the opportunity to learn from different related tasks (e.g. transfer learning) and representations (e.g. feature types). For instance, [14] trained a bidirectional LSTM on the SNLI dataset [2] and adapted the feature extraction in combination with hand-crafted features for ASAG in the Semeval-2013 dataset [5].

Regarding transfer learning, today's state-of-the-art in ASAG is defined by the use of large pretrained language models fine-tuned in the specific ASAG dataset [3]. The main difference between these approaches comes with the selection of the pretrained language model and the strategies to fine-tune it. [10] explore the potential of using T5 and XLNET, among others, as pretrained models for ASAG, and [17] proposed new ways to enhance the performance of BERT by further pretraining it as a language model on domain specific data such as textbooks and use labeled automatic short answer grading data.

Textual Entailment as a pivoting task The task of Textual Entailment, better known as Natural Language Inference (NLI), was first introduced by [4]. Given a textual premise and hypothesis, the task consists in classifying whether the premise entails or contradicts (or is neutral to) the hypothesis. The current state-of-the-art uses large pretrained LMs fine-tuned in NLI datasets [18].

Textual Entailment has been shown to be useful as a pivot task for few/zero-shot learning. For instance, entailment models are highly effective for text classification [16], and Information Extraction tasks [15], among others. The core idea relies on recasting the task at hand as an entailment task in which the original input is transformed into a classification problem of entailment, contradiction or neutral. Pretraining LMs in existing large entailment datasets and recasting the task into an entailment problem has demonstrated that it is possible to reduce annotation effort and still obtain state-of-the-art results [15].

Regarding ASAG, [3] explore the effect of transfer learning by fine-tuning a variety of pretrained LM models on the Semeval-2013 dataset. In a similar fashion to us, they also explored the impact of transfer learning with a model fine-tuned on the MNLI dataset [19]. They showed that models trained on this dataset are capable of transferring knowledge to the task of short answer grading, but did not quantify the effectiveness of the model in small data regimes. Other research lines are also appearing. To mention a few, [13] explore how to evaluate an automated grader in small-scale testing scenarios to help teachers and students in the use of such systems, [1] propose a similarity-based model, and [6] examine human-in-the-loop frameworks to guarantee grading quality.

3 Entailment Based Answer Grading

3.1 Problem Formulation

The Automatic Short Answer Grading task can be defined as follows. Given a triplet of *question*, *reference answers*, and *student answers* as input in our system, the system must assess the student answer by classifying it with a label that denotes the degree of correctness. We conduct our experiments in SciEntsBank and Beetle datasets, which were made available in the Semeval-2013 task 7 [5] and are one of the most used datasets. Note that Semeval-2013 datasets include three sets of labels that correspond to 2-, 3- and 5-way task problems, respectively. In this paper, we focus on the 3-way task, in which each answer is labeled as either correct, contradictory, or incorrect. Figure 1 depicts an example of a correct answer to a given question and the corresponding reference answer.

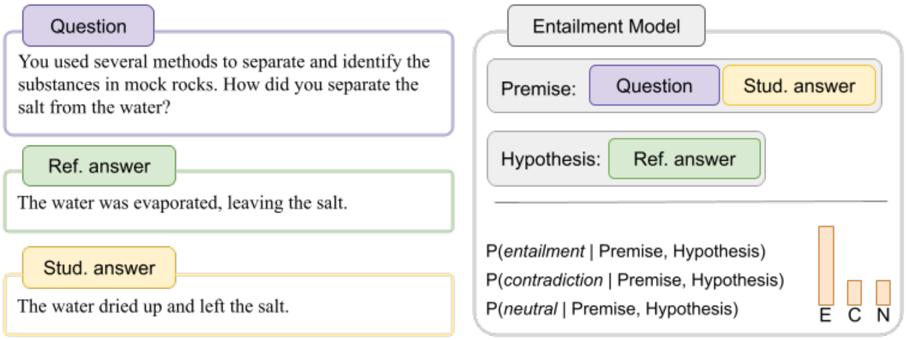
Question
You used several methods to separate and identify the substances in mock rocks. How did you separate the salt from the water?
Ref. answer
The water was evaporated, leaving the salt.
Stud. answer
The water dried up and left the salt.
Entailment Model
Premise:
Question
Stud. answer
Hypothesis:
Ref. answer
P(*entailment* | Premise, Hypothesis)
P(*contradiction* | Premise, Hypothesis)
P(*neutral* | Premise, Hypothesis)
E
C
N

Fig. 1. Schema of the NLI-based ASAG model where the input of *question*, *reference answer*, and *student answer* are reformulated as an entailment model. Concatenation of the question and student answer form the *premise* of the NLI model, whereas the *hypothesis* is generated with the reference answer. Prediction of the entailment model is then mapped to the ASAG 3-way label.

Evaluation scenarios The SemEval-2013 challenge gives three different test scenarios in order to evaluate model generalization capabilities across problems and domains:

- **Unseen answers** (UA): A set containing held-out student answers from questions which are available for training the system and contain some other student answers.
- **Unseen questions** (UQ): A set containing held-out questions in order to assess the system in non-seen questions, but still laying in the same domain as the one used for training.
- **Unseen domains** (UD): Available only for SciEntsBank, a domain independent test set of responses to topics not seen in the training data.

3.2 Model Description

According to the standard definition of Textual Entailment, given two text fragments called Premise (P) and Hypothesis (H), P entails H if, typically, a human reading P would infer that H is most likely true [4]. In a typical answer assessment scenario, we expect that a correct student answer would entail the reference answer, while an incorrect answer would not. However, students often skip details that are mentioned in the question or may be inferred from it, while reference answers often repeat or make explicit information that appears in or is implied from the question [5]. Hence, a more precise formulation of the task in this context considers the entailing text P as consisting of both the original question and the student answer, while H is the reference answer.

Figure 1 shows the schema of our entailment-based ASAG model, where the input of *question*, *reference answer*, and *student answer* are reformulated as a textual entailment problem. Concatenation of the question and student answer form the premise (P) of the NLI model, whereas the hypothesis (H) is created using the reference answer.

In our experiments we focus on the 3-way classification task so the predictions of the entailment model are mapped to the 3-way set of labels in the Semeval-2013 dataset. That is, the predictions of *entailment*, *contradiction*, and *neutral* of the NLI model are mapped to `correct`, `contradictory` or `incorrect`, respectively.

3.3 Fine-tuning the ASAG Model

We take advantage of NLI's ability to represent other NLP downstream tasks, ASAG in this case. Taking a large LM fine-tuned on MNLI as a base (RoBERTa-MNLI) [11], fine-tuning the ASAG model is carried out by reformulating the triplets (question, reference answer, student answer) provided in both SciEntsBank and Beetle datasets as traditional inference pairs (premise, hypothesis) as shown in Figure 1.

4 Annotation Strategies

The paper's main contribution is to explore the effectiveness of different annotation strategies when there is a need to have new annotated examples. RQ1 not only deals with data quantity, but also selects new samples to effectively save time and effort. Similarly, RQ2 takes into account the importance of selecting unseen data wisely in order to take advantage of the annotation to the fullest extent possible.

In order to answer these research questions, we explore two strategies of data annotation using the SemEval-2013 dataset. As the dataset has multiple student answers for a given question, the sampling of labeled data can be done answer- or question-wise. Specifically, we define two ways for sampling the training set of our experiments:

- *One question per student (1Q1S)* This scenario annotates a unique question and student answer pair. That is, if we had 10 students, we would create 10 different questions and would have 10 different answers. The goal of this strategy is to increase the variability of the questions, losing the capacity to generalize over the answers. Note that having very few examples for a given question might necessarily be a better strategy. Note as well that in some cases it is not possible to sample the defined dataset as there are not enough questions in the dataset. In those cases, we tried to generate an approximated dataset as in the previous strategy.
- *One question for all students (1Q4A)* This scenario annotates multiple answers for a single question with the goal of having larger variability on the answers side. That is, if we had 10 students, we would create and ask a single question to all the students in order to get many answers for the question. Note that in most cases there are not enough answers for a single question, so we attempted to sample a dataset that approximated it as much as possible.

Table 1 shows the number of questions and student answers that each few-shot setting contains. As can be seen, the ideal 1Q1S and 1Q4A annotations are not always collected since there are limited questions and student answers per question. Even so, each annotation strategy aims to add more variability to either the questions or the student answers at hand.

Table 1. Number of questions (#Q) and student answers (#A) for each few-shot scenario according to the specific annotation strategy (Ann.) as well as the number of training examples (Total) for each few-shot setting and dataset. FT stands for full training.

Dataset	Ann.	1%			2%			5%			10%			FT
		#Q	#A	Total	#Q	#A	Total	#Q	#A	Total	#Q	#A	Total	Total
SciEntsBank	1Q1S	40	1	40	80	1	80	100	2	200	100	4	400	3966
	1Q4A	4	10		8	10		20	10		40	10		
Beetle	1Q1S	28	1	28	28	2	56	35	4	140	35	8	280	2833
	1Q4A	1	28		1	56		2	70		5	56		

5 Experimental Setting

We use the data provided in the SemEval-2013 shared task for our experiments. As explained above, the dataset consists of two distinct subsets: SciEntsBank and Beetle. The former is based on a corpus of student answers to assessment questions collected in 15 science domains, whereas the latter is based on transcripts of students interacting with the Beetle II dialogue system in the basic electricity and electronics domain. Although both subsets show similar structure, Beetle contains more than one reference answer for each question[1] while a single reference answer is given in SciEntsBank. SciEntsBank includes 150 assessment questions with 150 reference answers and 6242 student answers in total. By comparison, Beetle is a smaller subset, which includes 56 questions, 283 reference answers, and 5199 student answers in total.

Since there is no validation set in the SemEval-2013 dataset, we created one by separating some examples from the original training set. We obtained a specific validation set for each test scenario. For the UA scenario, the selection of validation examples was done answer-wise and we held out a set of student answers for questions existing in the training part. For UQ, the selection was carried out question-wise and we selected the same number of questions that were extracted for the test set. We sample 15 and 9 questions from the SciEntsBank and Beetle training datasets, respectively. As all the training data belongs to the same domain, it was not possible to create a validation set that met the

[1] We use one reference answer in our experiments.

Table 2. Number of validation examples for unseen answers and unseen question scenarios.

	SciEntBank		Beetle	
	UA	UQ	UA	UQ
#Questions	120	15	38	9
#Reference answers	120	15	170	35
#Student answers	472	531	351	757

conditions for the UD test scenario. In all the cases, we select the validation examples so that class distribution is kept as similar as possible to the training dataset. Table 2 displays the sizes of the validation sets in terms of number of questions, reference answers, and student answers for each dataset and test scenario.

In order to measure the effectiveness of the annotation strategies in different few-shot scenarios, we generate the same training sizes for 1Q1S and 1Q4A. We created samples of 1%, 2%, 5%, 10% of the remaining training set and we reduced the validation set according to the same ratio. Table 1 shows the number of training examples for each few-shot scenario in the *Total* columns. Although FT denotes full training, it actually contains 1003 and 1108 fewer examples than the original training set as a consequence of utilizing a certain amount of training examples as a validation set.

As in [3], we used the RoBERTa large [11] fine-tuned on the MNLI dataset as the base model for our zero- and few-shot experiments. The model is publicly available at Huggingface[2]. We performed the following hyperparameter exploration for each few-shot scenario: We ran our model for 25 epochs with a batch size of 4 and selected the best learning rate between 1e-5, 5e-5, and 4e-6, as well as the best gradient accumulation between 8 and 32. For the model selection we took the checkpoint with the lowest loss (cross-entropy) value in those 25 epochs.

6 Few-shot Experiments

Table 3 shows the results of the effect of the annotation strategies in the few-shot scenario and seeks to answer the question the question concerning what would be the best strategy to annotate new data with a pretrained NLI model. It displays the macro F-score for the few-shot experiments in which we fine-tune an entailment model (NLI-roberta) using 1%, 2%, 5%, and 10% of training data and evaluated in unseen answers (UA), unseen questions (UQ), and unseen domains (UD).

The results indicate that, overall, increasing the number of annotated questions at the cost of reducing the number of different answers (1Q1S) seems to be the best strategy compared to increasing the variability of answers (at the

[2] https://huggingface.co/roberta-large-mnli

Table 3. Results for the few-shot experiments. 1Q1S annotation correspond to training data where we annotate one question per student and 1Q4A correspond to the one question for all students annotation procedure.

Domain	Scenario	Annotation	0%	1%	2%	5%	10%	FT
SciEntsBank	UA	1Q1S	56.2	**59.5**	**63.2**	63.9	**67.0**	71.0
		1Q4A		58.3	60.0	**64**.1	59.6	
	UQ	1Q1S	65.8	**67.0**	**66.7**	64.4	64.2	68.6
		1Q4A		62.7	65.6	**65.9**	**66.8**	
	UD	1Q1S	59.0	57.9	58.7	**58.8**	**61.2**	67.6
		1Q4A		**58.2**	**59.2**	56.0	58.2	
Beetle	UA	1Q1S	51.0	50.0	**52.3**	**52.7**	**56.6**	73.8
		1Q4A		**50.1**	50.8	52.5	51.5	
	UQ	1Q1S	36.1	**37.0**	36.8	**38.0**	**43.1**	61.8
		1Q4A		34.8	**37.8**	36.5	37.1	
Overall F-score		1Q1S	-	**55.7**	56.5	**56.6**	**59.2**	-
		1Q4A	-	55.2	56.5	54.6	55.8	-

cost of reducing the variability of seen questions) when annotating new question-answer pairs. This trend is confirmed in the bottom rows of the table, where we report the macro-average of each few-shot setting. In addition, results suggest that 1Q1S annotation strategy yields better generalization properties as we increase the number of examples. For instance, when we annotate 400 examples in SciEntsBank and 280 examples in Beetle (10% few-shot setting), 1Q1S outperforms 1Q4A by almost 4 points and F-score increases steadily compared to the rest of the few-shot settings.

7 Cross-domain Experiments

Table 4 shows the results on cross domain evaluation. Results of the top rows try to answer the questions posed in **RQ2**. First, we analyse if it is better to fine-tune in a related task but in a different domain or to simply apply a zero-shot model on the new dataset. The column with the 0% headline stands for this setting in which we have an entailment-based ASAG model fine-tuned in an out-of-domain dataset (e.g. Beetle) and evaluate it in the target domain dataset (i.e ScientsBank). We compare the fine-tuned (task-aware) model to the zero-shot entailment base model (task-agnostic, in parenthesis) in order to measure the effect of using out-of-domain task-related examples in learning.

Contrary to our expectation, task-aware fine-tuned models obtain significantly lower results compared to the task-agnostic model that is only pretrained in the MNLI dataset and not fine-tuned in the specific task. Results suggest that the impact of the domain is larger than the knowledge that can be acquired from

Table 4. Results of cross-domain few-shot evaluation. BT stands for Beetle and SB for ScientsBank. In the top rows % indicates the amount of in-domain data included in the training set, whereas the bottom rows refer to the amount of out-of-domain data.

Train → test	Scenario	0%	5%	10%
BT+%SB → SB	UA	55.8 (↓56.2)	58.9 (↓63.9)	**63.3** (↓67.0)
	UQ	59.7 (↓65.8)	62.5 (↓65.9)	**62.8** (↓66.8)
	UD	53.9 (↓59.0)	56.0 (↓58.8)	**59.3** (↓61.2)
SB+%BT → BT	UA	50.4 (↓51.0)	51.0 (↓52.7)	**54.0**(↓56.6)
	UQ	33.8(↓36.1)	34.6(↓38.0)	**37.9**(↓43.1)
SB+%BT → SB	UA	**71.0**	70.6	68.5
	UQ	68.6	69.8	**74.3**
	UD	**67.6**	66.7	64.0
BT+%SB → BT	UA	**73.8**	72.7	71.0
	UQ	**61.8**	59.9	55.7

the task. The drop is larger in the unseen questions scenario (UQ) in both Bettle and Scientsbank datasets. This can be explained assuming that unseen question scenarios require a higher capacity of generalization to perform better. In that sense, results suggest that generalization can not achieve using related tasks for transfer learning. In order to effectively transfer task related nuances the domain must be related as well.

Similarly, the results of columns 5% and 10% in the top rows of Table 4 also try to answer the question posed in **RQ2**. In this scenario we assume that we already have an entailment-based ASAG model fine-tuned in an out-of-domain dataset (e.g SciEntBank) and we get some annotated examples of our target domain (i.e Beetle). We evaluate the performance of adding target domain examples into the out-of-domain model.

Results demonstrate that adding few in-domain examples improves the outcome compared to the model trained only in the out-of-domain scenario. However, they are significantly worse compared to in-domain few-shot models (figures in parenthesis). The results are in accordance with those obtained in Table 3 and suggest that the domain differences can affect negatively even if we are modeling the same task (which is something unexpected according to some recent work [15]). That is, we can conclude that, having an entailment model, it is better to start from scratch rather than learning an out-of-domain ASAG model and retraining with a few in-domain examples.

When we defined a new setting where we do have an in-domain ASAG model (NLI model fine-tuned with target domain examples) and added some out-of-domain examples, we observed the model behaved similarly as in the previous settings. Results are shown in the bottom rows of Table 4. In general, we can conclude that mixing in-domain examples with out-of-domain examples is not helpful (only the unseen questions scenario in SciEntsBank obtains any improvement).

Table 5. Comparison to SOTA F-Macro results. Underlined figures denote that current results outperform previous state-of-the-art models. * for results not directly comparable with ours. Bold for best among comparable results. In ours FT and 10% experiments, the validation examples are included in the training set.

Model	SciEntsBank			Beetle	
	UA	UQ	UD	UA	UQ
CoMeT [12]	64.0	38.0	40.4	71.5	46.6
ETS [9]	64.7	45.9	43.9	71.0	58.5
(Galhardi et al., 2018) [7]	70.2	49.3	53.7	67.7	58.8
(Saha et al., 2018) [14]	66.6	49.1	47.9	-	-
(Sung et al., 2019) [17]	72.0*	57.5*	57.9*	-	-
(Camus and Filighera, 2020) [3]	78.3*	65.7*	70.9*	-	-
Ours 10% (1Q1S)	67.1	67.3	62.5	58.9	48.2
Ours FT	**76.5**	**72.3**	**69.1**	**76.7**	**70.0**

8 Comparison to the State-of-the-Art

Table 5 details the comparison of our model with state-of-the-art systems in SciEntsBank and Beetle datasets and the corresponding evaluation scenarios: Unseen answers (UA), unseen questions (UQ), and unseen domain (UD). The table is organized into three groups: 1) top rows include the best systems that took part in the Semeval-2013 shared-task, which correspond to hand-engineered feature-based systems; 2) middle rows include systems that rely on fine-tuned language models; 3) bottom rows include our model, fine-tuned using 10% of the data annotated with the 1Q1S strategy and fine-tuned utilizing the whole set of the original training examples. It is worth noting that the best-performing systems in SciEntsBank [17, 3] are not directly comparable with the rest of the models as it is not clear how model selection was carried out.

Regarding our few-shot model (10%-1Q1S), results show that annotating only 10% of examples (cf. Table 1) for training following the 1Q1S strategy is effective to outperform state-of-the-art systems in SciEntsBank dataset but not in the case of Beetle. It is also remarkable that Beetle appears more demanding, as recent state-of-the-art models [7]) are not able to surpass systems that participate in the SemEval-2013 shared task.

When we fine-tune our model employing all the data available in the training set, the model yields state-of-the-art results in both datasets and shows impressive generalization capabilities in those scenarios that presumably are more challenging. For example, our few-shot model improves in 18.0 F-score points in SciEntsBank compared to the best comparable model in the unseen questions (UQ) scenario (49.3 vs 67.3) and we increase the margin up to 23.0 points when we utilize the whole training set for fine-tuning the model (ours FT). In Beetle improvements rise to 11.2 F-score points with the full-training model.

9 Conclusion

In this study we reformulate Automatic Short Answer Grading as an entailment problem and explore the extent to which annotation strategies are effective in few-shot scenarios. Experiments show that increasing the variety of questions in the annotation is more effective than annotating more answers of the same question. Our method makes effective use of available labeled examples and, utilizing only 400 annotated examples, is able to perform on par with state-of-the-art approaches in SciEntBank. Moreover, when we use full-training, our model outperforms the rest of the models in the two datasets. Our analysis indicates that employing cross-domain annotated examples is not beneficial and it is more effective to use a task-agnostic general purpose entailment model. Actually, zero-shot obtains strong results, which indicates that the reformulation of ASAG into an entailment problem can be done naturally.

In the future we hope to explore methods to improve the limitation of grading answers into a more fine-grained level of entailment (subject to some arbitrary evaluation rubrics). In that sense, using generative large language models to learn reasoning on answers assessments seems to be a promising research avenue. On the other hand, adopting active learning to find refined ways of selecting questions would be a complementary approach to be explored. It would also help in measuring the variance of the sample selection and in obtaining robust findings.

Acknowledgements Work supported by the Basque Government (excellence research group IT1570-22) and by the DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

References

1. Bexte, M., Horbach, A., Zesch, T.: Similarity-based content scoring - how to make S-BERT keep up with BERT. In: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022). pp. 118–123. Seattle, Washington (2022)
2. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 632–642. Lisbon, Portugal (2015)
3. Camus, L., Filighera, A.: Investigating transformers for automatic short answer grading. Artificial Intelligence in Education **12164**, 43 – 48 (2020)
4. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (2005)
5. Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., Dang, H.T.: SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In: Proceedings of the

Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 263–274. Atlanta, Georgia, USA (2013)

6. Funayama, H., Sato, T., Matsubayashi, Y., Mizumoto, T., Suzuki, J., Inui, K.: Balancing cost and quality: An exploration of human-in-the-loop frameworks for automated short answer scoring. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 465–476. Springer International Publishing, Cham (2022)
7. Galhardi, L.B., de Mattos Senefonte, H.C., de Souza, R.C.T., Brancher, J.D.: Exploring distinct features for automatic short answer grading. In: Anais do XV Encontro Nacional de Inteligência Artificial e Computacional. pp. 1–12. SBC (2018)
8. Haller, S., Aldea, A., Seifert, C., Strisciuglio, N.: Survey on automated short answer grading with deep learning: from word embeddings to transformers. arXiv preprint arXiv:2204.03503 (2022)
9. Heilman, M., Madnani, N.: ETS: Domain adaptation and stacking for short answer scoring. In: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 275–279. Atlanta, Georgia, USA (2013)
10. Khayi, N., Rus, V., Tamang, L.: Towards improving open student answer assessment using pretrained transformers. The International FLAIRS Conference Proceedings **34** (2021)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019), https://arxiv.org/abs/1907.11692
12. Ott, N., Ziai, R., Hahn, M., Meurers, D.: CoMeT: Integrating different levels of linguistic modeling for meaning assessment. In: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 608–616. Atlanta, Georgia, USA (2013)
13. Padó, U.: Assessing the practical benefit of automated short-answer graders. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. pp. 555–559. Springer International Publishing, Cham (2022)
14. Saha, S., Dhamecha, T.I., Marvaniya, S., Sindhgatta, R., Sengupta, B.: Sentence level or token level features for automatic short answer grading?: Use both. In: Artificial Intelligence in Education (2018)
15. Sainz, O., Gonzalez-Dios, I., Lopez de Lacalle, O., Min, B., Eneko, A.: Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In: Findings of the Association for Computational Linguistics: NAACL-HLT 2022. Seattle, Washington (2022)
16. Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 255–269. Association for Computational Linguistics (2021)
17. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: International Conference on Artificial Intelligence in Education. pp. 469–481. Springer (2019)
18. Wang, S., Fang, H., Khabsa, M., Mao, H., Ma, H.: Entailment as few-shot learner (2021)
19. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 1112–1122 (2018)