Cross-Lingual Automatic Short Answer Grading

Chapter · March 2022

DOI: 10.1007/978-981-16-7527-0_9

2 authors, including:

Tim Schlippe
IU International University of Applied Sciences
63 PUBLICATIONS **998** CITATIONS

Cross-Lingual Automatic Short Answer Grading

Tim Schlippe and Jörg Sawatzki

IU International University of Applied Sciences
tim.schlippe@iu.org

Abstract. Massive open online courses and other online study opportunities are providing easier access to education for more and more people around the world. However, one big challenge is still the language barrier: Most courses are available in English, but only 16% of the world's population speaks English [1]. The language challenge is especially evident in written exams, which are usually not provided in the student's native language. To overcome these inequities, we analyze AI-driven cross-lingual automatic short answer grading. Our system is based on a Multilingual Bidirectional Encoder Representations from Transformers model [2] and is able to fairly score free-text answers in 26 languages in a fully-automatic way with the potential to be extended to 104 languages. Augmenting training data with machine translated task-specific data for fine-tuning even improves performance. Our results are a first step to allow more international students to participate fairly in education.

Keywords: cross-lingual automatic short answer grading, artificial intelligence in education, natural language processing, deep learning.

1 Introduction

Access to education is one of people's most important assets and ensuring inclusive and equitable quality education is goal 4 of United Nation's Sustainable Development Goals [3]. Distance learning in particular can create education in areas where no educational institutions are located or in times of a pandemic. There are more and more offers for distance learning worldwide and challenges like the physical absence of the teacher and the classmates or the lack of motivation of the students are countered with technical solutions like videoconferencing systems [4] and gamification of learning [5]. The research area "AI in Education" addresses the application and evaluation of Artificial Intelligence (AI) methods in the context of education and training [5]. One of the main focuses of this research is to analyze and improve teaching and learning processes. However, a major challenge is still the language barrier: Most courses are offered in English, but only 16% of the world population speaks English [1]. Figure 1 illustrates the 15 languages in the world which are spoken as first or second language. To reach the rest of the people with massive open online courses and other online study opportunities, courses would need to better support more languages. The linguistic challenge is especially evident in written exams, which are usually not provided in the student's native language.

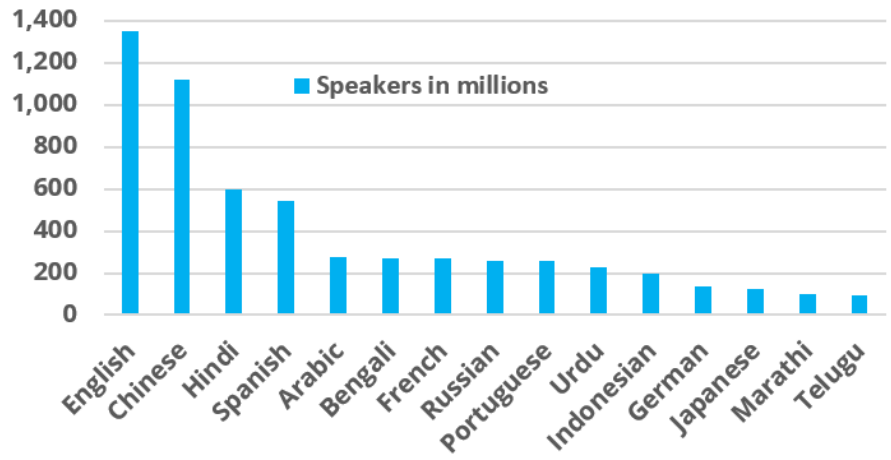

Fig. 1. The 15 most spoken *L1* and *L2* languages (based on [1]).

To overcome these inequalities, we analyze AI-driven cross-lingual automatic short answer grading (ASAG). While the focus of related work in ASAG has been on the performance of a corpus in only 1 language–whether using monolingual or multilingual pre-trained natural language processing (NLP) models–the focus of this paper is on leveraging the benefits of a multilingual NLP model for the application on multiple languages in the context of cross-lingual transfer. The Multilingual Bidirectional Encoder Representations from Transformers model (*M-BERT*) [2] is such a multilingual NLP "model pre-trained from monolingual corpora in 104 languages" which can be adapted to a certain task with task-specific labeled text data in 1 or more languages (*transfer learning*) and then perform this learned task in other languages (*cross-lingual transfer*) [7].

Compared to separate monolingual ASAG systems, cross-lingual ASAG has the following advantages: First, only one model is required to cover many languages instead of separate models which saves storage space and is easier to maintain. Second, we do not need task-specific data in each target language for fine-tuning due to the cross-lingual transfer. To investigate cross-lingual ASAG, we compared the performances of three different approaches:

1. *M-BERT* fine-tuned on a single language.
2. *M-BERT* jointly trained on 6 different languages and
3. monolingual *BERT* models.

In the next section, we will present the latest approaches of other researchers for ASAG. Section 3 will describe the experimental setup for our study of cross-lingual ASAG with 26 languages. Our experiments and results are outlined in Section 4. We will conclude our work in Section 5 and suggest further steps.

2 Related Work

A good overview of approaches in ASAG before the deep learning era is given in [8]. Newer publications are based on *bag-of-words*, a procedure based on term frequencies [9,10].

The latest trend which has proven to outperform traditional approaches is to use neural network-based embeddings, such as *Word2vec* [11]. [12] have developed *Ans2vec*, a *feature extraction architecture* based approach. They evaluated their concept with the English data set of the University of North Texas [13]. The advantage of this data set is that it contains scored student answers, while the answers of other short answer grading corpora, e.g., the SemEval-2013Task7 data sets [14] are only categorized into 3 classes–there is no point-based grading. [15] investigate and compare state-of-the-art deep learning techniques for ASAG and outperform [12] on the data set of the University of North Texas with a *fine-tuning architecture* based on the *Bidirectional Encoder Representations from Transformers* (*BERT*) [2] model. [16], [17] and [18] also deal in their work with *BERT* fine-tuning architectures. [18] report that their multilingual *RoBERTa* model [19] shows a stronger generalization across languages on English and German.

We extend their approach to 26 languages and use the smaller *M-BERT* [20] model to conduct a larger study concerning the cross-lingual transfer. While in most ASAG systems answers are categorized into only 3 classes, we focus on point-based grading. Our goal is to give a detailed analysis over the languages and investigate if cross-lingual ASAG allows students to write answers in exams in their native language and graders to rely on the scores of the system.

3 Experimental Setup

3.1 Evaluation metrics

As in related literature, we evaluate our results with the Mean Absolute Error (*MAE*) which is calculated from the average deviations of the prediction from the target value. This metric provides an intuitive understanding in terms of the deviation of points which makes it possible to compare the systems' performance to human graders.

3.2 Data Set

The short answer grading data set of the University of North Texas [13] is used for our experiments. Table 1 summarizes the features of this data set.

Table 1. Information of the short answer grading data set of the University of North Texas.

	English
Subject	Data Structures
#questions with model answer	87
#answers (total)	2.442
#answers per question	28.1
Ø length of answer (#words)	18.4
Maximum scores (in points)	5

It contains 87 questions with corresponding model answer and on average 28.1 manually graded answers per question about the topic *Data Structures* from undergraduate studies. Each student answer received a score from 0–5 points from two independent graders. We used the average of these 2 scores as our prediction target. We randomly selected 80% of the ASAG data set (1.953 student answers) for training and the remaining 20% (489 student answers) for evaluation. Table 2 shows a typical question, the corresponding model answer and two student answers. One of them was given the full score, the other one is a rather weak answer. The structure of the questions is demonstrated in the example.

Table 2. Original sample question and answers from the English data set.

Question	What is a variable?
Model answer	A location in memory that can store a value.
Example: Answer 1	A variable is a location in memory where a value can be stored.
Grading: Answer 1	5 of 5 points
Example: Answer 2	Variable can be an integer or a string in a program.
Grading: Answer 2	2 of 5 points

In order to produce artificial student and model answers for the adaptation, evaluation and comparison of the multilingual and the monolingual *BERT* models, we translated the English ASAG text data into 25 languages using Google's Neural Machine Translation System [21]. This procedure is also done by other researchers who experiment with multilingual NLP models [22] since this machine translation system comes close to the performance of professional translators [23–25]. An overview of the BLEU scores over languages is given in [24,25]. The translation of the questions and answers in Table 1 into German and Chinese are demonstrated in Table 3 and 4.

Table 3. Machine-translated sample question and answers in German.

Question	Was ist eine Variable?
Model answer	Eine Stelle im Speicher, die einen Wert speichern kann.
Example: Answer 1	Eine Variable ist ein Ort im Speicher, an dem ein Wert gespeichert werden kann.
Grading: Answer 1	5 of 5 points
Example: Answer 2	Eine Variable kann in einem Programm ein Integer oder ein String sein.
Grading: Answer 2	2 of 5 points

Table 4. Machine-translated sample question and answers in Chinese.

Question	什么是变量？
Model answer	内存中可以存储值的位置。
Example: Answer 1	变量是内存中可以存储值的位置。
Grading: Answer 1	5 of 5 points
Example: Answer 2	变量可以是整数，也可以是程序中的字符串。
Grading: Answer 2	2 of 5 points

To get a first impression of how people evaluate our translations in particular, we had 33 German students evaluate the German translations. Most of them stated that the translations are linguistically correct and understandable, as shown in Figure 2.

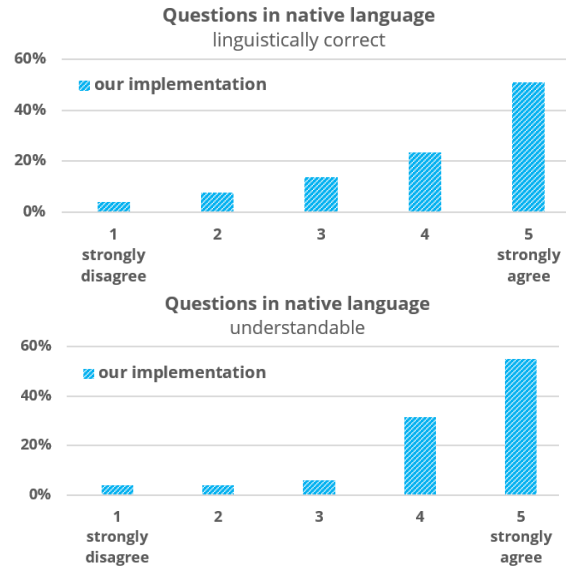

Figure 2. Feedback of 33 students on the machine-translated ASAG data set in German.

We produced ASAG data sets in the 26 languages that have the most Wikipedia articles [26]. These languages are spoken by more than 2.9 billion people (38% of the world population) and cover the language families Indo-European, Austronesian, Austroasiatic, Japonic, Afroasiatic, Sino-Tibetan, Koreanic, and Uralic [26].

3.3 Natural Language Processing Models

Our goal was to analyze the performance of cross-lingual ASAG with the help of a multilingual model in comparison to monolingual ASAG.

To investigate cross-lingual ASAG for our languages, we experimented with NLP models based on *BERT* [2] since *BERT* models are small compared to other NLP models, e.g., *RoBERTa* [19], but still provide high performances on several NLP tasks [2]. Our evaluated NLP model *M-BERT* refers to a multilingual *BERT* which could support 104 languages [20].

To compare *M-BERT* to monolingual models from different languages families, we use the following 6 models:

- *bert-base-cased* (*en*),
- *bert-base-german-dbmdz-cased* (*de*),
- *bert-base-chinese* (*zh*),
- *wietsedv/bert-base-dutch-cased* (*nl*),
- *TurkuNLP/bert-base-finnish-cased-v1* (*fi*), and
- *cl-tohoku/bert-base-japanese-char* (*ja*).

The models were downloaded and fine-tuned through the *simpletransformers* library [27], which is based on the Transformers library [28]. We trained 6 epochs with a batch size of 8 using the AdamW optimizer with an initial learning rate of 0.00004. We supplemented each fine-tuned *BERT* model with a linear regression layer that outputs a prediction of the *score* given an answer. The model expects the model answer and the student answer as input.

Figure 3, Figure 4, and Figure 5 demonstrate the training and testing procedures of our monolingual (*mono*) and multilingual ASAG systems: Our monolingual ASAG systems are exclusively fine-tuned with data from the target language, e.g., Chinese (*zh*) as shown in Figure 3.

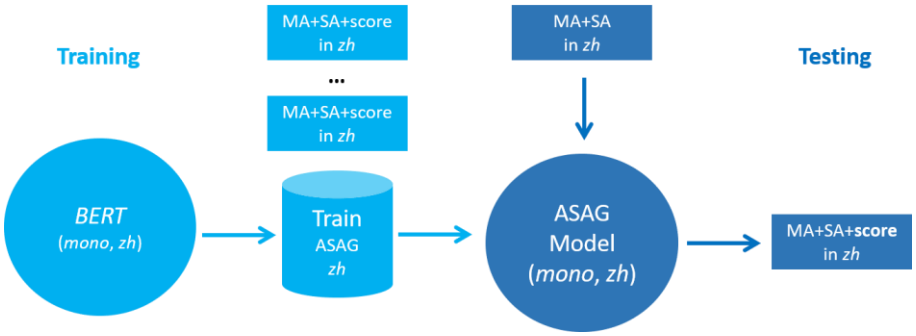

Figure 3. Monolingual ASAG system (here trained with Chinese).

As illustrated in Figure 4, the multilingual systems only need to be fine-tuned with ASAG data of 1 language, e.g., with the original English ASAG data (*Train ASAG en*). Then the multilingual *ASAG model* is able to receive a model answer together with a student answer in 1 of the other 103 languages and return a score in terms of points–without the need of fine-tuning with ASAG data in the other languages (*cross-lingual transfer*).

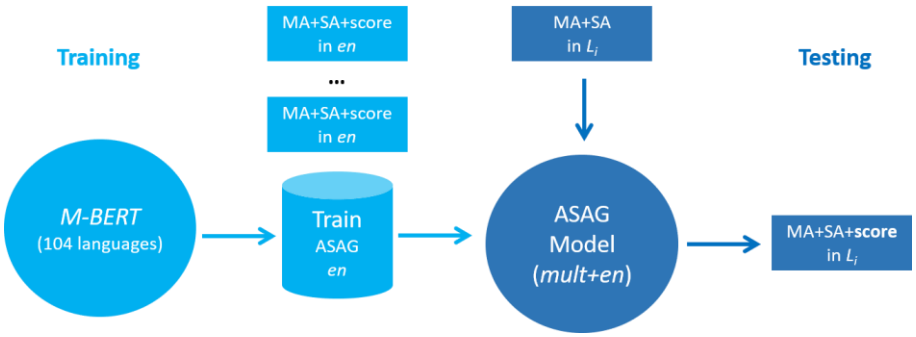

Figure 4. Multilingual ASAG system with *cross-lingual transfer* (here trained with English).

As shown in Figure 5, we additionally investigated if adding translations of the ASAG data in more languages improves fine-tuning and performance, respectively.

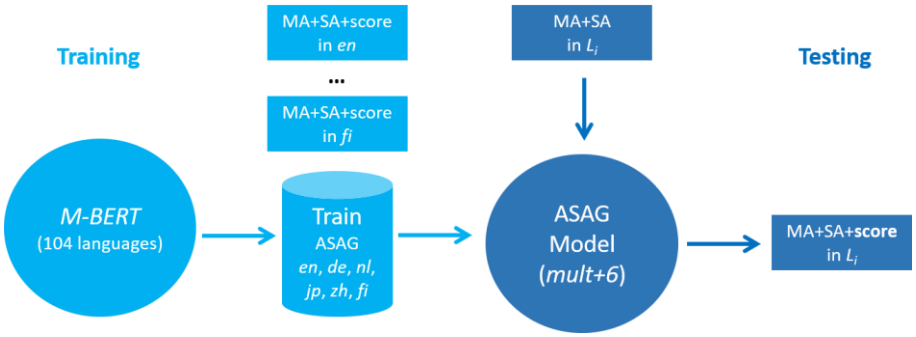

Figure 5. Multilingual ASAG system with *cross-lingual transfer* (trained with English, German, Dutch, Japanese, Chinese, and Finnish).

4 Experiments and Results

In our experiments we investigated the following research questions:

- How is the performance with monolingual models over languages? (Fig. 2)
- How is the performance with multilingual models over languages?
  - fine-tuned with task-specific data in target languages (Fig. 3)
  - fine-tuned with task-specific data in other languages (Fig. 4)
- How is the performance with monolingual/multilingual models compared to human graders?

Table 5 shows the deviation in context of the scoring scale from 0 to 5 points, represented by the Mean Absolute Error (*MAE*). The columns represent *M-BERT* fine-tuned on a single language (*multi+xx*), *M-BERT* trained on 6 languages (*multi+6*) and our monolingual *BERT* models (*mono*). The rows represent the evaluation of the models in 26 languages. When we look at the results, we need to consider the grader variability: The scores given by the 2 graders of the ASAG data set of the University of North Texas differ on average by 0.75 points, which is a relative difference of 15% [13]. Our results in Table 5 indicate that fine-tuning the multilingual model *M-BERT* with task-specific data in 6 languages (*multi+6*) is more beneficial than fine-tuning *M-BERT* with task-specific data in English (*multi+en*) or with task-specific data in 1 language (*multi+xx*)—even if the language *xx* is the target language (*multi+*L_{target}). If *xx* is not the target language, *multi+xx* performs worse than *multi+*L_{target} but even fine-tuning with task-specific data from 1 other language results in *MAE*s lower than 0.86 points. This shows the strong effect of the cross-lingual transfer and is an impressive result considering that no data from the target language at all was used for training and that human graders differ by 0.75 points.

Table 5. ASAG performance (*MAE*): Multilingual and monolingual models.

	multi+ en	**multi+ de**	**multi+ nl**	**multi+ jp**	**multi+ zh**	**multi+ fi**	**multi+ 6**	**mono**
en	0.45	0.61	0.64	0.68	0.63	0.63	**0.43**	0.43
ceb	0.70	0.73	0.72	0.68	0.72	0.71	**0.63**	-
sv	0.63	0.67	0.68	0.73	0.72	0.68	**0.48**	-
de	0.64	0.51	0.67	0.70	0.70	0.65	0.46	**0.45**
fr	0.61	0.66	0.64	0.67	0.70	0.67	**0.54**	-
nl	0.62	0.64	0.52	0.70	0.73	0.67	**0.45**	0.47
ru	0.68	0.73	0.83	0.74	0.75	0.78	**0.52**	-
it	0.62	0.65	0.72	0.71	0.73	0.70	**0.52**	-
es	0.61	0.68	0.76	0.68	0.72	0.65	**0.49**	-
pl	0.62	0.71	0.77	0.69	0.72	0.68	**0.51**	-
vi	0.71	0.72	0.84	0.77	0.73	0.71	**0.52**	-
jp	0.66	0.70	0.73	0.49	0.63	0.71	**0.44**	0.53
zh	0.63	0.71	0.77	0.69	0.50	0.79	**0.41**	0.44
ar	0.72	0.78	0.85	0.78	0.76	0.76	**0.59**	-
uk	0.65	0.70	0.82	0.73	0.73	0.75	**0.54**	-
pt	0.59	0.67	0.75	0.69	0.73	0.69	**0.50**	-
fa	0.64	0.66	0.71	0.67	0.70	0.69	**0.56**	-
ca	0.64	0.70	0.74	0.70	0.76	0.67	**0.53**	-
sr	0.69	0.81	0.83	0.76	0.79	0.86	**0.56**	-
id	0.66	0.68	0.69	0.70	0.79	0.63	**0.49**	-
no	0.63	0.69	0.65	0.75	0.71	0.69	**0.45**	-
ko	0.70	0.70	0.76	0.66	0.66	0.67	**0.58**	-
fi	0.69	0.79	0.77	0.77	0.73	0.52	0.47	**0.45**
hu	0.69	0.76	0.81	0.72	0.76	0.69	**0.54**	-
cs	0.62	0.77	0.82	0.72	0.78	0.71	**0.51**	-
sh	0.66	0.77	0.79	0.74	0.78	0.79	**0.53**	-

Note: Human grader variability is **0.75** points.

The ASAG performance of *multi+6* shows only deviations between 0.41 points (Chinese (*zh*)) and 0.63 points (Cebuano (*ceb*)) which is 8% to 13% relative. Furthermore, Table 5 shows that *multi+*L_{target} is more beneficial than *multi+en*: *multi+*L_{target} achieves a cross-lingual performance with small deviations between 0.45 points (English (*en*)) and 0.52 points (Finnish (*fi*)) which is only 9% to 10% relative. However, if target language data is not available, fine-tuning with English data (*multi+en*) is sufficient since it comes only with marginal deviations between 0.45 points (English (*en*)) and 0.72 points (Arabic (*ar*)) which is only 9% to 14% relative.

The monolingual models (*mono*) slightly outperform *M-BERT* fine-tuned and evaluated on the same language (*multi+*L_{target}) with deviations between 0.43 points (English (*en*)) and 0.53 points (Japanese (*jp*)). However, *multi+6* outperforms 4 of the 6 monolingual models demonstrating good overall performance and cross-lingual transfer.

Human graders deviate more (with 0.75 points, 15%) than the ASAG models which were cross-lingually adapted with English (worst *MAE*: 0.72), fine-tuned with the target language (worst *MAE*: 0.52), fine-tuned with our 6 languages (worst *MAE*: 0.63), and our monolingual models (worst *MAE*: 0.53).

Table 6. ASAG performance (*MAE*): *multi+en* vs. *multi+6*.

	multi+ en	**multi+ 6**	**rel. improvement**
en	0.45	0.43	4.4%
de	0.64	0.46	28.1%
nl	0.62	0.45	27.4%
jp	0.66	0.44	33.3%
zh	0.63	0.41	34.9%
fi	0.69	0.47	31.9%

Table 7. ASAG performance (*MAE*): *multi+*L_{target} vs. *multi+6*.

	multi+ L_{target}	**multi+ 6**	**rel. improvement**
en	0.45	0.43	4.4%
de	0.51	0.46	9.8%
nl	0.52	0.45	13.5%
jp	0.49	0.44	10.2%
zh	0.50	0.41	18.0%
fi	0.52	0.47	9.6%

The significant improvements of *multi+6* over *multi+en* and *multi+*L_{target} for English (*en*), German (*de*), Dutch (*nl*), Japanese (*jp*), Chinese (*zh*), and Finnish (*fi*) are listed in Table 6 and 7. With the wide range of English online study opportunities, in many cases ASAG data from English courses would be used for fine-tuning. However, in Table 6 we see that we can achieve up to 35% improvement by adding more languages. Even if we already have ASAG data in the target language, adding the 5 languages provides improvements of up to 18%, as demonstrated in Table 7.

5 Conclusion and Future Work

Our analysis on 26 languages demonstrated the potential of cross-lingual ASAG to allow students to write answers in exams in their native language and graders to rely on the scores of the system. With *MAEs* which are only between 0.41 and 0.72 points out of 5 points, our best models *multi+6*, *multi+xx* and *mono* have even less discrepancy than the 2 graders, which is 0.75 points in our corpus. Augmenting training data with machine translated task-specific data for fine-tuning improves performance of multilingual models. We are aware that our results have to be considered experimentally. Depending on the domain and the language combination, we see challenges in achieving optimal quality in machine translation. Nevertheless, we are very confident and plan to investigate this augmentation with different combinations and numbers of languages. We hope that performance is in a similar range for further languages and intend to analyze this in the future. If this is true, with multilingual models, we do not need training data in the target language at all to reach human level. To enhance online and distance learning, our next step includes to analyze the integration and application for online exams on the one hand but on the other hand for interactive training programs to prepare students optimally for exams. Figure 6 demonstrates our visualization of a multilingual interactive conversational artificial intelligence tutoring system for exam preparation [30], where students can prepare for exams in their native language, e.g., Dutch, in a gamification approach and automatically receive points for their free text answers.

Figure 6. Conversation with greeting, language selection, exam question, student answer, scoring, model answer and motivation.

References

1. Statista: The Most Spoken Languages Worldwide in 2019 (2020), https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide
2. 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. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019)
3. United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021), https://www.un.org/sustainabledevelopment/sustainabledevelopment-goals
4. Correia, A.P., Liu, C., Xu, F.: Evaluating Videoconferencing Systems for the Quality of the Educational Experience. Distance Education 41(4), 429–452 (2020)
5. Koravuna, S., Surepally, U.K.: Educational Gamification and Artificial Intelligence for Promoting Digital Literacy. Association for Computing Machinery, New York,NY, USA (2020)
6. Libbrecht, P., Declerck, T., Schlippe, T., Mandl, T., Schiffner, D.: NLP for Student and Teacher: Concept for an AI based Information Literacy Tutoring System. In: The 29th ACM International Conference on Information and Knowledge Management (CIKM 2020). Galway, Ireland (19-23 October 2020)
7. Pires, T., Schlinger, E., Garrette, D.: How Multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4996–5001. Association for Computational Linguistics, Florence, Italy (2019)
8. Burrows, S., Gurevych, I., Stein, B.: The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education 25, 60–117 (2014)
9. Süzen, N., Gorban, A., Levesley, J., Mirkes, E.: Automatic Short Answer Grading and Feedback using Text Mining Methods. Procedia Computer Science 169, 726– 743 (2020)
10. Zehner, F.: Automatic Processing of Text Responses in Large-Scale Assessments. Ph.D. thesis, TU München (2016)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013)
12. Gomaa, W.H., Fahmy, A.A.: Ans2vec: A Scoring System for Short Answers. In: Hassanien, A.E., Azar, A.T., Gaber, T., Bhatnagar, R., F. Tolba, M. (eds.) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019). pp. 586–595. Springer International Publishing, Cham (2019)
13. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 752–762. Association for Computational Linguistics, Portland, Oregon, USA (2011)
14. 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. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics. Atlanta, Georgia, USA (2013)

15. Sawatzki, J., Schlippe, T., Benner-Wickner, M.: Deep Learning Techniques for Automatic Short Answer Grading: Predicting Scores for English and German Answers. In: The 2nd International Conference on Artificial Intelligence in Education Technology (AIET 2021), Wuhan, China (2021).
16. Krishnamurthy, S., Gayakwad, E., Kailasanathan, N.: Deep Learning for Short Answer Scoring. International Journal of Recent Technology and Engineering 7, 1712–1715 (2019)
17. Sung, C., Dhamecha, T., Mukhi, N.: Improving Short Answer Grading Using Transformer-Based Pre-training. Artificial Intelligence in Education pp. 469–481 (2019)
18. Camus, L., Filighera, A.: Investigating Transformers for Automatic Short Answer Grading. Artificial Intelligence in Education 12164, 43–48 (2020)
19. 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 (2019)
20. Devlin, J.: BERT-Base, Multilingual Cased (2019), https://github.com/googleresearch/bert/blob/master/multilingual.md
21. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. CoRR abs/1609.08144 (2016)
22. Budur, E., Özçelik, R., Gungor, T., Potts, C.: Data and Representation for Turkish Natural Language Inference. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8253–8267. Association for Computational Linguistics, Online (2020)
23. Stapleton, P., Leung Ka Kin, B.: Assessing the Accuracy and Teachers' Impressions of Google Translate: A study of Primary L2 Writers in Hong Kong. English for Specific Purposes 56, 18–34 (2019)
24. Aiken, M.: An Analysis of Google Translate Accuracy. Studies in Linguistics and Literature 3, 253 (2012)
25. Aiken, M.: An Updated Evaluation of Google Translate Accuracy. Studies in Linguistics and Literature 3, 253 (2019)
26. Wikimedia: List of Wikipedias (2021), https://meta.wikimedia.org/wiki/List_of_Wikipedias#All_Wikipedias_ordered_by_number_of_articles
27. Rajapakse, T.C.: Simple Transformers. https://github.com/ThilinaRajapakse/simpletransformers (2019)
28. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (2020)
29. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
30. Schlippe, T., Sawatzki, J.: AI-based Multilingual Interactive Exam Preparation. The Learning Ideas Conference 2021 (14th annual conference). ALICE - Special Conference Track on Adaptive Learning via Interactive, Collaborative and Emotional Approaches. New York, New York, USA (2021)