Int J Artif Intell Educ (2024) 34:1294–1320
https://doi.org/10.1007/s40593-023-00370-1

ARTICLE

Crosslingual Content Scoring in Five Languages Using Machine-Translation and Multilingual Transformer Models

Andrea Horbach[1,2] · Joey Pehlke[1] · Ronja Laarmann-Quante[3] · Yuning Ding[1]

Accepted: 15 August 2023 / Published online: 3 November 2023

Abstract

This paper investigates crosslingual content scoring, a scenario where scoring models trained on learner data in one language are applied to data in a different language. We analyze data in five different languages (Chinese, English, French, German and Spanish) collected for three prompts of the established English ASAP content scoring dataset. We cross the language barrier by means of both shallow and deep learning crosslingual classification models using both machine translation and multilingual transformer models. We find that a combination of machine translation and multilingual models outperforms each method individually - our best results are reached when combining the available data in different languages, i.e. first training a model on the large English ASAP dataset before fine-tuning on smaller amounts of training data in the target language.

Keywords Automated content scoring · Automated short-answer scoring · Crosslingual methods · Domain transfer

Ronja Laarmann-Quante and Yuning Ding contributed equally to this work.

✉ Andrea Horbach
horbach@uni-hildesheim.de

Joey Pehlke
joey.pehlke@fernuni-hagen.de

Ronja Laarmann-Quante
ronja.laarmann-quante@rub.de

Yuning Ding
yuning.ding@fernuni-hagen.de

[1] CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics, FernUniversität in Hagen, Universitätsstraße 47, 58097 Hagen, Germany

[2] Hildesheim University, Institut für Informationswissenschaft und Sprachtechnologie, Universitätsplatz 1, 31141 Hildesheim, Germany

[3] Ruhr University Bochum, Faculty of Philology, Department of Linguistics, Universitätsstraße 150, 44801 Bochum, Germany

Introduction

In this paper, we investigate crosslingual methods for automated content scoring, a scenario where scoring models trained on data in a source language are applied to data in a different target language. In content scoring in general, students answer to a prompt in written form, typically with one or a few sentences. These answers are scored on the basis of their content rather than their form. It is only important that the student understood the right concepts, while form aspects, such as orthography and grammar, are not considered for the final score. In a real-life educational setting, such a procedure can discriminate against non-native speaking students who might conceptually understand the topic in question, but are unable to express their understanding in the language of instruction. One solution to this problem could be that students are allowed to answer a question in a language they are proficient in. As only the content matters, the form, including the language, is unimportant. This setting, however, would require that a teacher scoring an item is also proficient in the language used by the student, which would still restrict the available language options. We therefore investigate in this paper the feasibility of crosslingual automated scoring, where training data in one language is used to predict scores on test items in a different language.

Besides fostering educational equality, crosslingual scoring can also help to overcome data sparsity, as available data in one language can be used to score the same item in a new language without the need to collect and humanly annotate large amounts of new training data in the target language.

Our work builds on earlier work by Horbach et al. (2018) who collected ASAP$_{de}$, a German version of the established English ASAP dataset (henceforth called ASAP$_{orig}$) and conducted crosslingual scoring experiments with the help of automatic machine translation finding a substantial performance loss compared to monolingual scoring.

We extend this work in several respects: First, we extend the crosslingual setup to include a broader variety of languages. We collected new versions of the ASAP dataset in Spanish and French through crowd-sourcing studies and include a recently collected Chinese version (Ding et al., 2020). While German and English are closely related Germanic languages for which machine translation usually works very well, including those further languages allows us to examine the influence of language proximity on crosslingual scoring performance. Second, previous work (Horbach et al., 2018) has speculated that part of the performance drop reported in a crosslingual setting could be attributed rather to the data collection setup than to the language difference, as ASAP$_{orig}$ was collected from high school students while ASAP$_{de}$ was acquired through crowdsourcing. To further address this question, we also collected a crowdsourced ASAP version in English, ASAP$_{en}$. For each of these setups (ASAP$_{es}$, ASAP$_{fr}$, ASAP$_{en}$), we collected about 300 learner answers for each of the three individual prompts already used in ASAP$_{de}$. These datasets were manually double-annotated by trained annotators. Figure 1 shows as an example the English version of the prompt material from prompt 1. Table 1 presents example answers from the different language-specific datasets in response to that prompt. Note that the shown answers from different datasets are conceptually very similar and thus all received the same score. We take this as a hint

Prompt 1 - Acid Rain

A group of students wrote the following procedure for their investigation.
Procedure:

1. Determine the mass of four different samples.
2. Pour vinegar in each of four separate, but identical, containers.
3. Place a sample of one material into one container and label. Repeat with remaining samples, placing a single sample into a single container.
4. After 24 hours, remove the samples from the containers and rinse each sample with distilled water.
5. Allow the samples to sit and dry for 30 minutes.
6. Determine the mass of each sample.

The students' data are recorded in the table below.

Sample	Starting Mass (g)	Ending Mass (g)	Difference in Mass (g)
Marble	9.8	9.4	–0.4
Limestone	10.4	9.1	–1.3
Wood	11.2	11.2	0.0
Plastic	7.2	7.1	–0.1

After reading the group's procedure, describe what additional information you would need in order to replicate the experiment. Make sure to include at least three pieces of information.

Fig. 1 Prompt 1 from the original English ASAP dataset

that the annotators' scoring, who followed the scoring instructions published with ASAP$_{orig}$, is indeed consistent across datasets.

Third, we methodologically extend existing crosslingual work to include recent developments in neural content scoring as well as crosslingual text-classification. We follow Horbach et al. (2018) in using machine translation to either automatically translate training data into the language of the test data or vice versa as a baseline condition. As machine translation has the risk of introducing noise into the data, especially in the case of learner data with misspellings and other language problems, we investigate ways to circumvent the machine translation step through the use of a multilingual BERT classification model (Devlin et al., 2019) capable of dealing

Table 1 Example answers in different languages for prompt 1

Dataset	Example Answer
ASAP$_{orig}$	Some additional information that I would need is the amount of vinegar they poured.
ASAP$_{en}$	you need to know what amount of vinegar to put on each sample.
ASAP$_{de}$	Eventuell sollte man wissen, wie viel Essig in die Behälter gegeben werden soll.
ASAP$_{es}$	Cantidad de vinagre usada en cada experimento.
ASAP$_{fr}$	la quantité de vinaigre on ajouté dans les echantillon.
ASAP$_{zh}$	加入醋的量

All answers have been scored with 1 point (out of a range between 0 and 3)

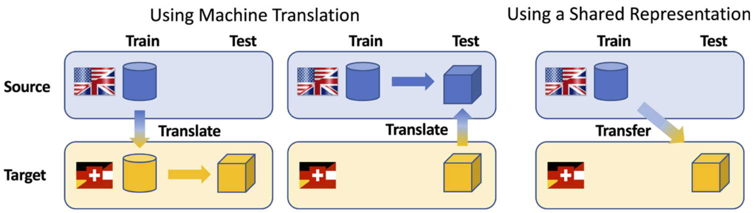


Fig. 2 Visualization of the crosslingual scoring process at the example of English as source and German as target language

with multilingual input. We use such a model as another experimental condition and explore ways how the two approaches, machine translation and the use of multilingual transformer models, can be combined. (See Fig. 2 for a visualization of the different variants of the scoring process.) In a domain adaptation approach, we also investigate how smaller amounts of in-domain data from the same language as the test data can be used to fine-tune a model trained on larger amounts of out-of-domain data, i.e. in our case the large amounts of English answers from the original dataset (but always using training data from the same prompt as the test data).

Our paper thus makes the following contributions:

- We compare and discuss five additional language variants of the ASAP dataset, three of which we have newly collected. [1]
- We present crosslingual experiments methodologically covering approaches based on (a) machine translation, (b) multilingual transformer models and (c) a combination thereof, showing that the latter works best in most setups.
- We show that results can be further improved in a two-step training process, where a model is first trained on large amounts of data in a source language and then fine-tuned on the smaller available amounts of target language training data.

The remainder of the paper is organized as follows. We discuss related work in "Related Work". In "Datasets", we discuss and analyze the multilingual short-answer datasets used in our study. "Experimental Studies" presents two experimental studies. The first one uses monolingual scoring for comparison. The second one uses crosslingual scoring with machine translation or multilingual transformer models or both to cross the language boundary. We further analyze the datasets and classification results in "Further Analyses" before concluding in "Conclusion and Future Work" and discussing ethical considerations in "Ethical Considerations".

Related Work

In the following, we first describe options how NLP tasks in general can be placed in a crosslingual setting. We then review the automated scoring literature with respect to multi- and crosslingual setups.

[1] The data is publicly available under https://github.com/andreahorbach/CrosslingualScoring/.

Crosslingual Methods

Most NLP applications have been developed in a monolingual setting, which makes sense given that any machine learning task works best if the training data is as similar as possible to the actual test data at application time. However, given that the acquisition of new training data is a time-consuming task, crosslingual methods have been investigated as one way of performing domain transfer as the language of the data is one domain aspect in which two data sets can differ. In our paper, we mean by crosslingual NLP tasks any NLP task with training data in one source language and test data in a different target language.

There are several ways how the language boundary can be crossed, the most straightforward one being by means of machine translation. In information retrieval, for example, queries have been automatically translated from English to Spanish and vice-versa in a crosslingual information retrieval setup (Ballesteros and Croft, 1996). Another approach for crosslingual text-classification consists in translating a learnt n-gram model such as Shi et al. (2010). They accounted for the fact that individual words can have multiple translations by learning feature translation probabilities. Avoiding machine translation, Prettenhofer and Stein (2010) proposed a structural correspondence learning method for text classification, where large amounts of unlabeled data and some bilingual word correspondence lists are used to learn structural correspondences between two languages, i.e. basically to learn a task-specific machine translation model.

In the context of recent neural network advances, crosslingual word-embeddings have become popular, for example in work by Klementiev et al. (2012), who learn representations of words in different languages in the same vector space, such that *dog*, *chien*, *Hund*, *perro* and 狗 are represented by similar feature vectors. Finally, large pre-trained language models such as BERT (Devlin et al., 2019), which we employ in our work, have also been trained on documents in a multitude of languages making them multilingual. Such approaches thus seemingly remove the need to translate training or test data.

Crosslingual Automated Scoring

The existence of available data in more than one language is a necessary pre-requisite for multi- and crosslingual scoring. However, the main body of work in the area of automated scoring concerns monolingual setups in English. This holds for both automated content scoring, where only conceptual correctness is scored (Horbach and Zesch, 2019), and essay scoring, where both content and linguistic form are evaluated (Klebanov and Madnani, 2021). We therefore first provide pointers to approaches on languages other than English, before discussing the few instances of truly multi- or crosslingual scoring approaches that have been developed so far.

Our work is an instance of automated content scoring, with an ample body of work for monolingual English scenarios and a fair number of established English-language datasets (e.g., Bailey and Meurers (2008); Basu et al. (2013); Dzikovska et al. (2013); Mohler and Mihalcea (2009). In contrast, approaches and datasets for languages other than English are much scarcer, especially when it comes to approaches with datasets that are freely available for research[2]. They include languages such as Arabic (Abdul Salam et al., 2022; Ouahrani and Bennouar, year), German (Meurers et al., 2011; Pado and Kiefer, 2015; Sawatzki et al., 2021), Hebrew (Ariely et al., 2023), Indonesian (Herwanto et al., 2018; Wijaya, 2021), Japanese (Funayama et al., 2023), Portuguese (Galhardi et al., 2018; Gomes et al., 2021), Punjabi (Walia et al., 2019), Swedish (Weegar and Idestam-Almquist, 2023) or Turkish (Çınar et al., 2020)[3].

The English ASAP-SAS dataset[4] used as one of the datasets in this paper has received substantial attention in the content scoring community (e.g. Heilman and Madnani (2015); Higgins et al. (2014); Kumar et al. (2019)). Crosslingual content scoring approaches instead are rare, with Horbach et al. (2018) as one of the first instances. Recently, Schlippe and Sawatzki (2022) proposed a method for crosslingual scoring, unfortunately on datasets which are not publicly available. Similar to our approach, they also make use of multilingual transformer models. The work by Camus and Filighera (2020), while not presenting genuine crosslingual experiments, investigates the influence of translating whole content scoring datasets automatically into a foreign language. Note that for content scoring in general, most approaches fall either into the *instance-based* or the *similarity-based* algorithmic design (see Horbach and Zesch (2019)). The majority of current approaches belong to the first category where a classifier learns from a set of labeled learner answers properties of a correct or incorrect answer. An alternative is the similarity-based approach where learner answers are compared to one or several reference answers to determine their label. Both approaches have been found to reach similar performances Bexte et al. (2022, 2023). In this paper, we stick to the instance-based approach, as we extend prior, also instance-based, cross-lingual approaches. A similarity-based cross-lingual scenario opens up a vast additional parameter space - Do we fine-tune a monolingual or cross-lingual similarity metric? Should we translate the answers used to train the metric, the reference answers or the learner answers to be scored? A thorough investigation of these additional questions goes beyond the scope of this paper and will be tackled in future work.

The situation regarding data availability is similar for the essay scoring task. An abundance of work has been published for the English ASAP-AES dataset[5], a de-facto standard in automated essay scoring. Work with publicly available datasets in other languages mostly comes from a foreign-language learning perspective and provides data for CEFR classification, such as the Swedish SWELL corpus (Volodina et al., 2016), the German Falko corpus (Lüdeling et al., 2008) and the Portuguese COPLE2 corpus (Mendes et al., 2016). While most of these datasets are monolingual, the Merlin corpus (Boyd et al., 2014) is an exception in that it covers three languages,

[2] We acknowledge that the public release of educational data is always problematic and often requires substantial additional work in terms of anonymization, even more so in longer essays, which might contain personal information, rather than in shorter content scoring answers. See Megyesi et al. (2021) as an exemplary in-depth description of a thorough anonymization effort.

[3] See also https://catalpa-cl.github.io/EduScoringDatasets/ for an overview of educational scoring datasets.

[4] https://www.kaggle.com/c/asap-sas/

[5] https://www.kaggle.com/c/asap-aes

Czech, German and Italian, and has thus allowed for crosslingual CEFR classification experiments (Vajjala and Rama, 2018). They use a set of essay scoring features that generalize across languages, such as features based on part-of-speech tags and Universal Dependency relations. Such features are suitable for the domain of essay scoring and especially proficiency classification, where linguistic form is important, but less so for content scoring.

An important research branch also concerns large-scale assessment. This scoring branch is not discussed in detail here because these data sets can typically not be made public. Yet PISA (Peña-López et al., 2012) and similar international standardized assessments of educational attainment would be ideal deatasets for crosslingual scoring studies as data often comes in a wide variety of languages targeting exactly the same prompts per language. While there is research targeting the (semi-)automatic assessment of such data, e.g. by means of clustering approaches (Andersen et al., 2023; Zehner et al., 2016) for the German PISA subset, there has - to the best of our knowledge - not been a systematic investigation of crosslingual scoring for this data.

Datasets

In this section, we describe the data sets used in our study. We first discuss requirements for crosslingual data sets. Then we describe the data collection and annotation process. In order to assess how similar the new datasets are to each other and to the original English ASAP data, we analyze and compare them regarding answer length, label distribution and linguistic variance in the data.

Requirements for Crosslingual Data

When working on crosslingual data, language should be the only or, at least, the main factor where data sets differ from each other. Horbach et al. (2018) propose a set of requirements for crosslingual data sets that includes independence of the prompts of the respective culture, language or curriculum. Especially when extending an already existing dataset, the guidelines used to score this original dataset should be available to and applicable by new annotators. Following this rationale, we extend the multilingual ASAP data sets already used in their study by integrating a Chinese data set by Ding et al. (2020) and three new data set variants: English, French and Spanish. As mentioned above, although the original ASAP data is in English, we collected new English answers via crowd-sourcing in order to keep the population providing the answers similar across all languages in the multilingual dataset and have a way to compare ecologically valid educational data with crowd-sourced data in the same language.

Data Collection

Prompts 1, 2 and 10 in ASAP are science-related tasks, which do not have a strong cultural background, and are therefore considered as appropriate to be transferred to other languages and learner populations. In addition, earlier work found that annotators were unable to apply the annotation guidelines successfully for the other prompts (Horbach et al., 2018).

Therefore, we manually translated the original prompt material and collected answers in all languages for these three prompts.

The French data was collected by distributing the prompts via different platforms and groups (e.g. Facebook). The majority of answers were obtained from Amazon Mechanical Turk (Nguimkeng, 2021). In total, 1,070 answers were collected, of which 672 were usable (manually eliminating nonsense answers or those given in another language). The participants came from many different countries according to their self-report, predominantly from France and the USA, followed by Brazil, England, India and Italy. The new English data, as well as the Spanish data was collected via Amazon Mechanical Turk only, with the majority of participants coming from the US and (for Spanish) from Venezuela. To determine payment for crowd-workers, we tested how long it takes on average to read the instructions and answer the questions and calculated per-item payments in a way that crowd-workers would earn the same as a student assistant at university. Those participants in the French data collection not recruited via AMT were unpaid volunteers.

For the Chinese data, Ding et al. (2020) collected 314 answers per prompt from high school students in grades 9-12, which is a comparable population in age and educational background to the population in the original ASAP-SAS data set. The answers are manually transcribed from handwriting into digital form. The German dataset by Horbach et al. (2018) consists of 301 crowdsourced answers per prompt.

Table 2 shows key statistics for the different data sets. Apart from the original ASAP data set, all other data sets come with comparable amounts of answers per prompt.

Dataset Annotation

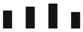
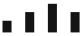
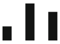
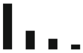
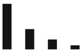
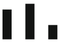
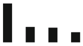
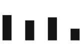
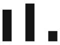
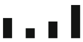
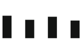
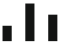
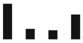
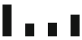
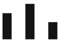
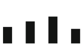
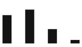
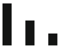
For the German, French, Spanish and Chinese datasets, two native speakers of the respective language scored the answers based on the original scoring guidelines. The newly collected English dataset was scored by two native speakers of German with advanced English language skills. As defined in the original scoring guidelines,

Table 2 Key statistics for the datasets used in our study

Dataset	Language	Prompt 1		Prompt 2		Prompt 10	
		# answ.	IAA	# answ.	IAA	# answ.	IAA
ASAP$_{orig}$	English	2,229	.94	1,704	.92	2,186	.88
ASAP$_{en}$	English	330	.83	328	.65	330	.69
ASAP$_{de}$	German	301	.85	301	.67	301	.58
ASAP$_{es}$	Spanish	325	.84	297	.66	393	.61
ASAP$_{fr}$	French	274	.65	187	.91	211	.70
ASAP$_{zh}$	Chinese	314	.72	314	.70	314	.69

Inter-annotator agreement (IAA) is measured in QWK

Table 3 Label distribution for the original ASAP dataset and the different language versions

Dataset	Prompt		
	1	2	10
ASAP$_{orig}$			
ASAP$_{en}$			
ASAP$_{de}$			
ASAP$_{es}$			
ASAP$_{fr}$			
ASAP$_{zh}$			

prompts 1 and 2 were rated on a scale from 0 to 3 points and prompt 10 on a scale from 0 to 2 points. All annotators were first trained on a set of answers from the original English ASAP-SAS dataset before they scored the new data. The inter-annotator agreement measured in quadratically weighted kappa is shown in Table 2.[6] The final gold standard was created in that the annotators discussed and agreed on a final label. Note that for the original English data, IAA only refers to the training set consisting of about 75% of the answers per prompt. IAA results for the other datasets are lower than for the original ASAP data, although Higgins et al. (2014) have already wondered whether these extremely high reported agreement values are really the result of completely independent ratings, while Shermis (2014) described kappa values between .62 and .85 as "a typical range for human rater performance in statewide high-stakes testing programs" (p. 58).

Dataset Analysis

We further analyze properties of our data sets assuming that any differences between datasets can be due either to the difference in language or to the difference in learner population. By including also a crowdsourced version of the English data we aim at decoupling those two influence factors. We investigate label distribution, answer length and linguistic diversity operationalized by type-token-ratio.

Label Distribution

We evaluate the relative frequency for each label per prompt and dataset, comparing the original ASAP dataset and the different additional versions for different languages (see Table 3). We see that all crowd-sourced datasets have a higher amount of low-scoring answers than the original ASAP data and attribute that to the lower intrinsic motivation in crowd-workers compared to the high school students involved in creating the original

[6] For Spanish, invalid answers were not filtered out beforehand but as part of the annotation process. In this case, inter-annotator agreement is reported only for answers which both annotators saw as valid answers.

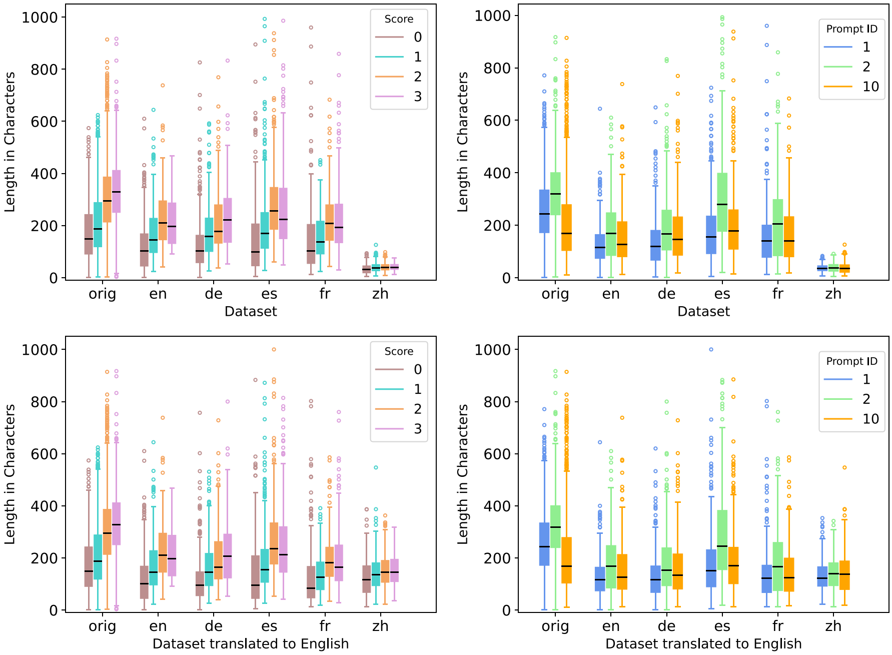


Fig. 3 Distribution of answer length in characters for all answers with a certain score (left part) and for all answers of a certain prompt (right part). We show results on the original datasets (upper part) and their English translations (lower part)

ASAP data and to the fact that the tasks were probably aligned with the science classes of the students, even though we tried to select prompts that were not particularly curriculum-dependent. These high amounts of low-scoring answers also lead to more imbalanced datasets, which bear the risk of being harder to score automatically. It seems that the prompts have different difficulty for individual populations, e.g. prompt 10 seems to be particularly difficult for Chinese respondents, while prompt 1 has high amounts of low-scoring answers among the crowdsourced English and German population.

Answer Length

We measure answer length in characters and report for every dataset the distribution of lengths for all answers per score across all prompts (Fig. 3, left part) and for all answers per prompt across all scores (Fig. 3, right part).[7] We observe that across all datasets, better answers, i.e. answers with a higher score, tend to be longer than answers with a lower score. We also see differences between prompts, with prompt 2 eliciting the longest answers. Effects from data collection contexts are visible in

[7] In Fig. 3, extreme outlier answers with a length of > 1,000 characters have been removed prior to plotting in order to allow for a more concise visualization. This affected 23 answers in the original datasets and 22 answers in the translated datasets. The longest answer was 4,493 characters long.

that English crowd-worker answers are shorter than answers from the original ASAP dataset. To factor out the presumed language influence in the length comparison (most pronounced for the Chinese dataset), we also compare a version of the dataset with all data automatically translated to English (Fig. 3, lower part) using DeepL (as we did later in the crosslingual scoring experiments using machine translation). One can see that even when translated to English, the answers in the Chinese dataset tend to be shorter than those in the other datasets. By inspecting the data, we found that the Chinese answers tend to appear in the form of bullet points. Instead of writing e.g. *After reading this procedure, I have noticed that I would need to know the Ph of the vinegar* in ASAP$_{orig}$, answers like *(1)PH of the vinegar sample* are often found in the Chinese dataset. Across all datasets, we observe that crowdsourced answers tend to be shorter than the original data.

Linguistic Diversity

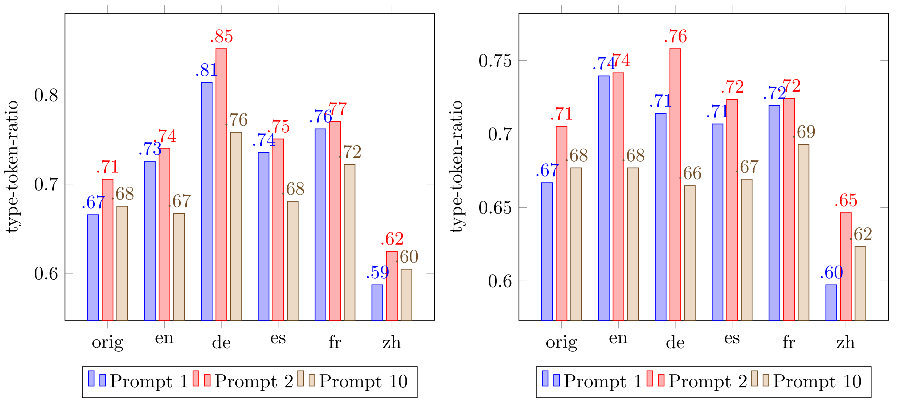
Variance in learner answers has been identified as one parameter influencing the scoreability of datasets (Horbach and Zesch, 2019). One example of such variance is linguistic diversity. Intuitively, a very repetitive data set is much easier to score than one with a broad range of different words and word combinations.

We measure linguistic diversity by type-token-ratio (TTR), i.e. the number of unique words of a text divided by the overall number of words. We use TTR of unlemmatized word-forms treating each dataset as an individual text. Type-token ratio is known to be influenced by text length, so that a comparison between TTR values is only possible if texts of the same length are compared. To overcome this problem, we use a method similar to the moving average TTR (Covington and McFall, 2010) and randomly draw subsamples of a fixed size from the datasets and report averages over the individual values.

For Chinese, TTR could either be calculated on the token level or on the character level (Cui et al., 2022). Here, we rely on a character-based representation because, in our content scoring scenario, the accuracy of tokenization is lower, since most of the answers are not complete sentences. From the results in Fig. 4 we see that most datasets fall into the same TTR range after translation to English with the exception of Chinese displaying a lower TTR ratio than the other datasets indicating that there seems to be slightly less lexical variance in the Chinese data.

Experimental Studies

In the following, we first describe our experimental setup, followed by three experimental studies. In Study 1, we establish baselines for each dataset by cross-validating on every dataset separately, either in its original version or translated into the other languages. In Study 2, we score crosslingually by translating training or test data into the respective other language or by using an inherently multilingual transformer model (either with or without additional translating datasets). In Study 3, we explore the



(a) on original data

(b) on English translations

Fig. 4 Type-token-ratio for the six datasets computed for answers from a certain prompt (a) on the original version of each dataset and (b) on the version translated to English

effects of pretraining a model on a larger amount of English data and then fine-tuning on a smaller amount of data in the target language.

Experimental Setup

Classifiers and Features

For the shallow learning baseline we use a logistic regression classifier provided through sklearn (Pedregosa et al., 2011)[8]. For all European languages, we use word uni- to trigrams and character bi- to 5-grams as features. In the case of Chinese, tokenization of words can be a non-trivial problem as the Chinese language does not separate words by whitespaces. Therefore, we follow Ding et al. (2020) in using only character uni- to 7-grams.

In the case of deep learning, we used a standard transformer architecture with a pre-trained multilingual BERT model (*bert-base-multilingual-uncased*) with a sequence classification head, as shown in Fig. 5, and, for comparison, also a monolingual English BERT model (*bert-base-unased*) as monolingual models are known to exhibit a stronger performance (Artetxe et al., 2023). We fine-tuned for six epochs using an AdamW optimizer and CrossEntropy loss.

Note that our main interest lies in comparing different crosslingual scoring models, not finding an optimal hyperparameter setup. We therefore hold meta-parameters fixed and also chose a fixed number of epochs instead of using validation data to select the best epoch.

[8] We also experimented with other shallow learning algorithms and found that logistic regression showed on average the best results.

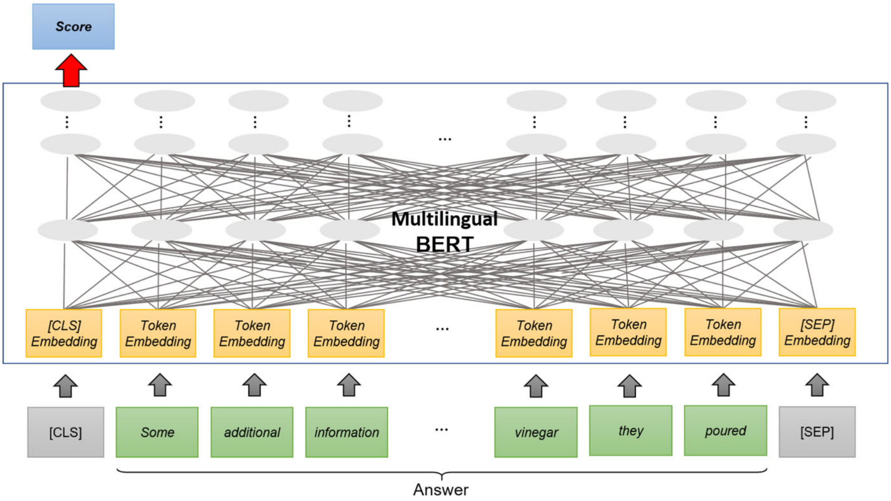


Fig. 5 Our neural architecture, using a multilingual BERT model with a sequence classification head

Datasplit and Evaluation Metric

For the ASAP$_{orig}$ dataset, we use the datasplit provided in the original data. Due to the small size of the other datasets, we use 10-fold cross-validation for these datasets. We always train separate models for each of the three prompts. Again, we do not use a separate validation dataset for deep learning.

We evaluate our experiments using quadratically weighted kappa (QWK) (Cohen, 1960).

Experimental Study 1: Within-Dataset Baselines

In our first experimental study, we establish baseline scoring performance per dataset, i.e. we train and test on each dataset individually as described in the previous section.

Table 4 Performance in QWK for monolingual baseline experiments

Dataset	LR				M-BERT			
	1	2	10	avg.	1	2	10	avg.
ASAP$_{orig}$	.700	.628	.674	.667	.791	.808	.720	**.773**
ASAP$_{orig\_300}$	.671	.424	.677	.591	.653	.618	.702	**.658**
ASAP$_{en}$	.692	.503	.629	**.608**	.677	.400	.617	.565
ASAP$_{de}$	.804	.570	.592	.655	.783	.668	.554	**.668**
ASAP$_{es}$	.786	.614	.631	.677	.805	.573	.661	**.680**
ASAP$_{fr}$	.718	.740	.667	.708	.734	.760	.685	**.726**
ASAP$_{zh}$	.616	.353	.529	.499	.666	.398	.587	**.550**
avg.	.712	.547	.628	.629	.730	.604	.647	**.660**

Scoring each Dataset in its Original Language

Table 4 shows the performance per language and prompt in both the shallow learning and deep learning condition. The original ASAP dataset (ASAP$_{orig}$) reaches the highest scoring performance, which is not surprising given that this dataset contains a much higher amount of training data. Therefore, we also down-sampled the dataset to 300 instances per prompt roughly matching the training sizes of the other data sets (ASAP$_{orig\_300}$). In this scenario, results are more on par with other datasets, suggesting that also the other datasets might benefit from more training data than available in the monolingual setup - a scenario we will test later in Study 3 when combining different data sets as training data.

We further observe for most datasets that prompt 2 is the hardest among the three prompts to score with the exception of ASAP$_{orig}$ and ASAP$_{fr}$ while prompt 1 is on average and across both shallow and neural approaches the easiest. This is reflected in substantially lower inter-annotator agreement for all but these two datasets for prompts 2 and 10 compared to prompt 1, hinting at a generally higher difficulty by most of the annotators to score these two prompts accurately. We also see that in most cases the BERT model outperforms logistic regression (the highest performance per dataset averaged across the three prompts is marked in bold in the table). Performance on Chinese data is slightly worse than for the other datasets despite high inter-annotator agreements. The values that we find in our experiment, however, are comparable to those reported by Ding et al. (2020) for Chinese and Horbach et al. (2018) for English and German.

Comparison between Monolingual and Multilingual BERT Models

We wanted to keep the pre-trained transformer model fixed across experiments and therefore chose to use the same multilingual BERT model for all of our experiments. However, monolingual models are known to exhibit a higher performance. Thus, in order to assess the potential performance drop from using an inherently multilingual pre-trained model in cases where a monolingual model could be used, we compare the two alternatives at the example of the English data.

Table 5 repeats the monolingual baseline results from the previous experiment, which used a multilingual transformer model for the three English datasets (on the right) and for comparison the same experiments using a monolingual English trans-

Table 5 Performance in QWK for monolingual baseline experiments comparing a monolingual transformer model (left) to a multilingual transformer model (right) for English

Dataset	BERT				M-BERT			
	1	2	10	avg.	1	2	10	avg.
ASAP$_{orig}$	.850	.792	.737	**.793**	.791	.808	.720	.773
ASAP$_{orig\_300}$	.662	.560	.715	.646	.653	.618	.702	**.658**
ASAP$_{en}$	.698	.383	.642	**.574**	.677	.400	.617	.565
avg.	.737	.578	.698	**.671**	.730	.604	.647	.660

Table 6 Performance in QWK for monolingual baseline experiments using the logistic regression model (left) and the neural model (right)

Dataset	LR					M-BERT				
	en	de	es	fr	zh	en	de	es	fr	zh
ASAP_{orig}	.667	-.029	.014	-.024	-.007	.773	.018	.013	.012	-.001
ASAP_{orig_300}	.591	-.011	-.038	-.003	-.029	.658	-.084	-.047	-.012	-.084
ASAP_{en}	.608	-.009	-.008	-.018	-.087	.589	-.061	-.011	-.009	-.050
ASAP_{de}	-.010	.655	-.008	.013	-.034	-.015	.677	-.012	.004	-.087
ASAP_{es}	-.013	-.027	.677	-.002	-.029	.004	-.012	.718	-.036	-.062
ASAP_{fr}	.029	-.004	-.005	.708	-.013	.010	.020	.029	.726	.009
ASAP_{zh}	.027	.012	.038	.036	.499	-.013	-.040	.057	.022	.505
avg	.634	.620	.628	.630	.601	.662	.641	.668	.661	.624

Training and test data (rows) have been translated into a different language (columns). Results are averaged across three prompts. Cells marked in blue do not use machine translation, i.e., they are the baseline results from Table 4 and report absolute values. All other cells report the delta from that baseline (i.e. from the baseline value in the same row) with green values indicating an improvement over the baseline and red values a performance drop

former model. Results are mixed, but we see a tendency that the monolingual BERT model indeed has a slight advantage over the multilingual model (.671 for the monolingual versus .660 for the multilingual model). As the improvements are rather small and in order to use the same model in all experiments, we use the M-BERT model from here on.

Scoring Translated Datasets

As a second monolingual baseline, we use DeepL[9] to translate both the test and the training data into the same other language, using each of the other languages represented in our datasets as the target language. I.e. we still score in a monolingual setting but both training and test data have been machine-translated.

Results are shown in Table 6, with values highlighted in blue on the main diagonal representing the untranslated baseline condition. In most cases, results produced on translated datasets are in the same ballpark as the original data, sometimes even slightly higher especially in the case of Chinese as the original language. This points at a normalizing effect that machine translation can have, for example on spelling errors. It can also mean that a multilingual model does not perform equally well for all language pairs (Pires et al., 2019).

We manually checked a number of misspelled sentences and indeed found that a spelling variant of a word, such as *vinagar* instead of *vinegar* or *expirement* instead of *experiment* were often translated correctly. Only occasionally such a misspelled word was not translated at all as if it were a named entity. We also encountered examples of non-words spelling errors being translated to a different word in the target language

[9] www.deepl.com

than the correct form, such as English *vinager* being translated to German *Wein* (*wine*). These probing support the idea of a normalization by translation, thus the translation would be able to provide more informative embeddings or ngrams than the original version.

The case is somewhat different for Chinese as potential misspellings in Chinese were already removed by force through the digitization process (i.e. when typing the handwritten answers, a non-existing character had to be entered as an existing one), so that non-word errors in Chinese are rare and real-word errors often led to problematic translations. For example, some Chinese answers contain the compound word 涮洗, which combines two real words 涮 (*rinse*) and 洗 (*wash*) in an unusual order. The intended compound word could be 洗涮 (*wash up*). This miss-spelled compound was translated into *shabu-shabu*, because the character 涮 is somehow related to 涮涮鍋 (*shabu-shabu*, a Japanese hotpot) by machine translation.

When looking at the average over all datasets translated into the same language (the last line of Table 6), we see that values are quite similar per machine learning method, indicating that both linear regression and M-BERT are able to handle data in different languages and that differences we saw in Table 4 can be mainly attributed to the dataset, not the language.

Crosslingual Scoring within ASAP$_{orig}$

Our ultimate goal is to score cross-lingually data from one dataset with a model trained on a different dataset in a different language. In the previous step, we have established monolingual baseline performance per dataset and have also established that there is hardly any loss from applying machine translation to a data set. Next, we investigate the influence of applying a multilingual transformer model on training and test data in different languages but from the same dataset in order to separate dataset from language effects. Therefore, as an additional third baseline, we translate either the training (**MT-train**) or the test data section (**MT-test**) from the ASAP$_{orig}$ set into each of the other four languages and then train a model on one language and apply it to another language (where one of those languages is always English).

From the results in Table 7 we see that there is a performance loss of up to 12 percentage points for German, Spanish and French compared to the English baseline where both training and test data are the original untranslated ASAP dataset. For Chinese, performance goes down by over 50%. We expect that these values are an upper bound for the performance we can expect in true crosslingual setups without machine-

Table 7 Performance in QWK for experiments on ASAP$_{orig}$ where either the training or the test data has been translated to a different target language

Language	MT-train	MT-test
en	.77	.77
de	.72	.65
es	.69	.65
fr	.66	.66
zh	.33	.35

translation (at least when one language is English) where the effect of different datasets add to that of different languages.

Experimental Study 2: Crosslingual Scoring using Machine Translation and Pretrained Multilingual Models

The monolingual baselines established in Study 1 can serve as an upper bound for the performance we hope to reach using crosslingual methods. We compare it against two ways of how to cross the language barrier. One is by means of machine translation to translate either train or test data into the respective other language. As a second method, similar to the previous study, we use the M-BERT model in a zero-shot fashion so that neither test nor training data is translated but rather both will be represented in a shared embedding space.

Table 8 presents these crosslingual results grouped by the test data in different languages. In every block, we train on the different available training data (one training data set per line) and apply the model to the same test data making results directly comparable. We compare the different methods per column: We present results for logistic regression with translated training data (**LR MT-train**) or test data (**LR MT-test**) and the same for the neural model (**M-BERT MT-train** and **M-BERT MT-test**). Finally, **M-BERT zero-shot** relies solely on multilingual representations without any machine translation.

When comparing the three methods in Table 8, logistic regression with MT, M-BERT with MT and M-BERT in a zero-shot setting, we see that the M-BERT setting almost always outperforms logistic regression and that machine translation is beneficial (comparing M-BERT MT with zero-shot). The best crosslingual method per block (highlighted in bold print) is often M-BERT with MT with no clear tendency whether translating training or test data is more beneficial. For all languages, the best setup among those with 300 training instances, i.e. excluding ASAP$_{orig}$, is lower than the monolingual baseline with the smallest distance between monolingual baseline and best crosslingual setup for French (.726 for monolingual vs .698 for crosslingual when trained on Spanish data) and the worst performance gap for Chinese (.550 vs .457). Also the zero-shot setup sometimes yields surprisingly good results, for example for the transfer from German or Spanish to French. On the basis of the monolingual comparison between a monolingual and a multilingual transformer model from Experimental Study 1, we assume that replacing the multilingual BERT model with monolingual versions for the respective variants would benefit both the monolingual baseline as well as the MT$_{train}$ and MT$_{test}$ variants as both could be scored with a monolingual model.

Overall, we observe that the transfer often works well for highly related languages, such as Franch and Spanish, while for Chinese, which is in our setup phylogenetically farthest from the other languages, the transfer does not nearly work as well.

Table 8 Performance in QWK for crosslingual experiments either using Logistic Regression (LR) or a multilingual BERT model (M-BERT) either with machine translation (MT) or relying only on the multilingual transformer (zero-shot)

Train	**Test**	**LR**		**M-BERT**		
		MT$_{train}$	**MT**$_{test}$	**MT**$_{train}$	**MT**$_{test}$	**zero-shot**
$\mathrm{ASAP}_{\mathbf{orig}}$	$\mathrm{ASAP}_{\mathbf{en}}$	.472		.599		
$\mathrm{ASAP}_{\mathbf{orig\_300}}$	$\mathrm{ASAP}_{\mathbf{en}}$	**.604**		.525		
$\mathrm{ASAP}_{\mathbf{en}}$	$\mathrm{ASAP}_{\mathbf{en}}$	.608		.608		
$\mathrm{ASAP}_{\mathbf{de}}$	$\mathrm{ASAP}_{\mathbf{en}}$	.512	.531	.523	.540	.494
$\mathrm{ASAP}_{\mathbf{es}}$	$\mathrm{ASAP}_{\mathbf{en}}$	.539	.514	.544	.497	.509
$\mathrm{ASAP}_{\mathbf{fr}}$	$\mathrm{ASAP}_{\mathbf{en}}$	.479	.497	.416	.374	.326
$\mathrm{ASAP}_{\mathbf{zh}}$	$\mathrm{ASAP}_{\mathbf{en}}$	.476	.431	.454	.414	.209
$\mathrm{ASAP}_{\mathbf{orig}}$	$\mathrm{ASAP}_{\mathbf{de}}$	.430	.457	**.623**	.554	.469
$\mathrm{ASAP}_{\mathbf{orig\_300}}$	$\mathrm{ASAP}_{\mathbf{de}}$	.438	.371	.568	.589	.436
$\mathrm{ASAP}_{\mathbf{en}}$	$\mathrm{ASAP}_{\mathbf{de}}$	.422	.401	.408	.576	.490
$\mathrm{ASAP}_{\mathbf{de}}$	$\mathrm{ASAP}_{\mathbf{de}}$	.655		.668		
$\mathrm{ASAP}_{\mathbf{es}}$	$\mathrm{ASAP}_{\mathbf{de}}$	.489	.459	.530	.515	.516
$\mathrm{ASAP}_{\mathbf{fr}}$	$\mathrm{ASAP}_{\mathbf{de}}$	.506	.527	.486	.480	.471
$\mathrm{ASAP}_{\mathbf{zh}}$	$\mathrm{ASAP}_{\mathbf{de}}$	.374	.389	.445	.460	.211
$\mathrm{ASAP}_{\mathbf{orig}}$	$\mathrm{ASAP}_{\mathbf{es}}$	.501	.497	**.689**	.634	.559
$\mathrm{ASAP}_{\mathbf{orig\_300}}$	$\mathrm{ASAP}_{\mathbf{es}}$	.417	.407	.638	.610	.474
$\mathrm{ASAP}_{\mathbf{en}}$	$\mathrm{ASAP}_{\mathbf{es}}$	.422	.410	.650	.631	.535
$\mathrm{ASAP}_{\mathbf{de}}$	$\mathrm{ASAP}_{\mathbf{es}}$	.486	.542	.639	.633	.593
$\mathrm{ASAP}_{\mathbf{es}}$	$\mathrm{ASAP}_{\mathbf{es}}$	.677		.680		
$\mathrm{ASAP}_{\mathbf{fr}}$	$\mathrm{ASAP}_{\mathbf{es}}$	.626	.588	.591	.590	.549
$\mathrm{ASAP}_{\mathbf{zh}}$	$\mathrm{ASAP}_{\mathbf{es}}$	.367	.343	.565	.494	.331
$\mathrm{ASAP}_{\mathbf{orig}}$	$\mathrm{ASAP}_{\mathbf{fr}}$	.503	.508	.595	.593	.513
$\mathrm{ASAP}_{\mathbf{orig\_300}}$	$\mathrm{ASAP}_{\mathbf{fr}}$	.378	.418	.596	.562	.374
$\mathrm{ASAP}_{\mathbf{en}}$	$\mathrm{ASAP}_{\mathbf{fr}}$	.389	.399	.545	.540	.466
$\mathrm{ASAP}_{\mathbf{de}}$	$\mathrm{ASAP}_{\mathbf{fr}}$	.504	.533	.620	.678	.635
$\mathrm{ASAP}_{\mathbf{es}}$	$\mathrm{ASAP}_{\mathbf{fr}}$	.564	.614	.689	**.698**	.665
$\mathrm{ASAP}_{\mathbf{fr}}$	$\mathrm{ASAP}_{\mathbf{fr}}$	.708		.726		
$\mathrm{ASAP}_{\mathbf{zh}}$	$\mathrm{ASAP}_{\mathbf{fr}}$	.402	.342	.519	.501	.288
$\mathrm{ASAP}_{\mathbf{orig}}$	$\mathrm{ASAP}_{\mathbf{zh}}$	.210	.226	.378	.432	.207
$\mathrm{ASAP}_{\mathbf{orig\_300}}$	$\mathrm{ASAP}_{\mathbf{zh}}$	.294	.413	.270	.257	.100
$\mathrm{ASAP}_{\mathbf{en}}$	$\mathrm{ASAP}_{\mathbf{zh}}$	.166	.240	**.457**	.321	.135
$\mathrm{ASAP}_{\mathbf{de}}$	$\mathrm{ASAP}_{\mathbf{zh}}$	.230	.360	.374	.368	.318
$\mathrm{ASAP}_{\mathbf{es}}$	$\mathrm{ASAP}_{\mathbf{zh}}$	.306	.329	.280	.330	.304
$\mathrm{ASAP}_{\mathbf{fr}}$	$\mathrm{ASAP}_{\mathbf{zh}}$	.362	.343	.286	.319	.271
$\mathrm{ASAP}_{\mathbf{zh}}$	$\mathrm{ASAP}_{\mathbf{zh}}$	.499		.550		

Cells in blue are monolingual baseline results for comparison. The best crosslingual results per target language are printed in bold

Experimental Study 3: Pretraining on Larger Amounts of Data in another Language

So far, we examined domain transfer scenarios where only data in another language but no target language data is available for training and compare it to a within-domain approach with (only) data in the target language. But one could imagine a real-life scenario where larger amounts of data might be available in a majority language (such as the original English ASAP data set) and only limited amounts of data in a new application language. To simulate such a use case, we further pre-train the pre-trained M-BERT model on the original ASAP dataset in English and then fine-tune on a specific language using cross-validation (similar to the monolingual experiments in Experimental Study 1) in order to see whether we benefit from English pretraining. I.e. each model in Study 3 has been trained on about 2000 English answers and about 270 target language answers.

We compare several conditions. In the **pretrain zero-shot** condition, we pretrain on the original (untranslated) English ASAP dataset before fine-tuning on the (untranslated) target language dataset. In **pretrain MT-train** we translate the original ASAP data into the respective target language before pretraining. In **pretrain MT-test**, we translate the in-domain data (for both fine-tuning and testing, i.e. the whole cross-validation procedure) from the target language into English.

For comparison, we report three baseline values: **monolingual** CV without pretraining, i.e. results from Study 1, as well as model performance when only training on ASAP$_{orig}$ without further fine-tuning on the in-domain data in two versions. One where the original ASAP data has been translated into the respective language and one where the test data has been translated to English, i.e. results from Study 2. (We do not show the zero-shot condition from Study 2 as another baseline as it was clearly outperformed by the two variants using machine translation.) As our method in this study is essentially always a combination of two baseline setups, we expect results to be above the maximum of these baselines.

Table 9 shows results for different test data (per line) and different experimental conditions (per column), the three baselines on the left followed by three options to combine both test data sets in the right part of the table.

We see that in all cases results outperform all three baselines. The improvement over the monolingual baseline indicates that we indeed benefit from more training data than the approximately 300 training instances in each dataset, even if it is not in the target language. The improvement over the ASAP$_{orig}$ MT baselines shows the importance of in-domain training data, given that a modest increase in training data size (adding less than 300 instances to an existing dataset of 2000 answers) results in a relatively large performance gain.

When comparing the three experimental conditions, there is no clear winner and the zero-shot condition performs surprisingly well. We take that as an indicator that fine-tuning on the target language can compensate for a pre-training in a non-matching language. (Note that the three identical results in the first line for ASAP$_{en}$ result from source and target language both being English, thus translating any data set has no effect.) Overall, results in this study are for Spanish and French in the same

Table 9 Performance in QWK for an M-BERT model pretrained on the original ASAP data and finetuned on each of the other datasets (right part) either training or test data is translated or neither (zero-shot)

Test Data	baselines			pretrained		
	monolingual	MT_{train}	MT_{test}	zero-shot	MT_{train}	MT_{test}
ASAP_{en}	.608	.599	.599	**.700**	**.700**	**.700**
ASAP_{de}	.668	.623	.554	**.735**	.730	.727
ASAP_{es}	.680	.689	.634	.755	**.776**	.743
ASAP_{fr}	.726	.595	.593	**.783**	.777	.780
ASAP_{zh}	.550	.378	.432	.626	**.651**	.629

This is compared to three baselines (left part): monolingual cross-validation within the same dataset and training on ASAP_{orig} with either translated training or test data

region as monolingual scoring performance on the full original English ASAP dataset (with much less additional annotation effort in the new target language), and at least substantially improved over the baselines in case of crowd-sourced English, German and Chinese.

Further Analyses

We saw in our experiments that crosslingual scoring with training and test data from different datasets yields a substantial performance loss compared to monolingual scoring where only one dataset is used, even if the data has been translated into a different language. This gap only becomes smaller, when in addition to the data from a different dataset also genuine target-language data is used. We therefore conclude that part of the drop in performance is due to the nature of the different datasets, i.e. that different user populations think or at least write differently about the same prompt. In the following, we thus provide additional analyses intended to shed light on those differences. We do this in two ways: by comparing the semantic similarity between answers across data sets and by examining high-frequency lexical material per data set.

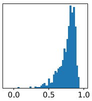
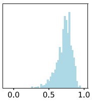
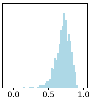
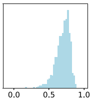
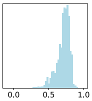
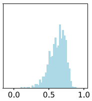
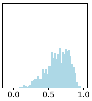
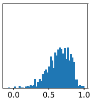
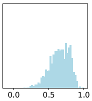
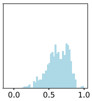
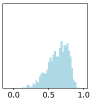
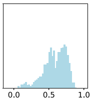
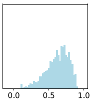
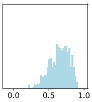
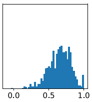
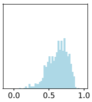
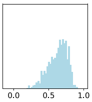
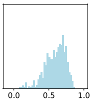
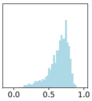
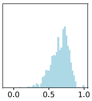
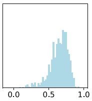
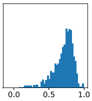
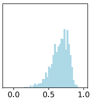
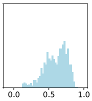
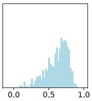
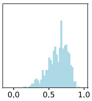
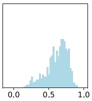
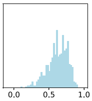
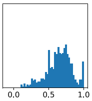
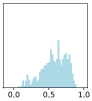
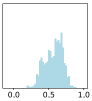
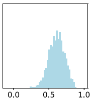
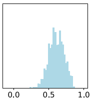
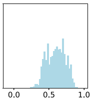
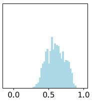
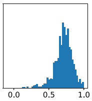
Similarity between Datasets

As a first analysis step, we measure the similarity between datasets. We operationalize this by measuring pairwise similarity between answers for the same prompt.

We are first and foremost interested in semantic similarity, not similarity on the surface. We measure semantic similarity by encoding every answer with a multilingual SBERT model (*distiluse-base-multilingual-cased-v1*) and calculate its cosine similarity with every other answer from either the same or a different dataset (only within the same prompt). For each answer in dataset A, we record the maximum similarity to any answer in dataset B, with A and B being either different or the same dataset. The idea is that if different user populations answer a prompt in conceptually different ways, we would see a low maximum answer similarity between these datasets. Likewise, we expect the answer similarities to be higher within a dataset. The reason why we look at maximum similarity only is that we want to see if a scoring model would have had the chance to see a similar answer in the training data or not.

Table 10 presents the results. Every histogram shows the distribution of the maximum similarity of every answer in the dataset specified in the row label with the answers in the dataset specified in the column label (measured within the same prompt). The diagonal from the upper left to the lower right corner contains the results for the similarities within one dataset. We see that ASAP$_{orig300}$ is rather homogeneous in that most answers have a matching answer with a very high similarity, i.e. the histogram has a clear peak towards the right. Also when compared to answers in the other datasets (see first row), ASAP$_{orig300}$ mostly has matching answers with high similarities, with the Chinese dataset behaving a bit differently. For the Chinese dataset itself (see last row), we see that the answers are rather similar to each other within this dataset but not so much when the answers are compared to any of the other datasets. This indicates that either the Chinese data is indeed rather dissimilar to the other datasets, which is in line with the observation that Chinese was hardest to score in a crosslingual setting, or that the multilingual SBERT model is not able to capture semantic similarity across languages for Chinese. Most of the other histograms for same-dataset vs. different-

Table 10 Distribution of maximum similarity between an answer from the dataset specified in the row label and any answer from the dataset specified in the column label

	ASAP$_{orig300}$	ASAP$_{en}$	ASAP$_{de}$	ASAP$_{es}$	ASAP$_{fr}$	ASAP$_{zh}$
ASAP$_{orig300}$						
ASAP$_{en}$						
ASAP$_{de}$						
ASAP$_{es}$						
ASAP$_{fr}$						
ASAP$_{zh}$						

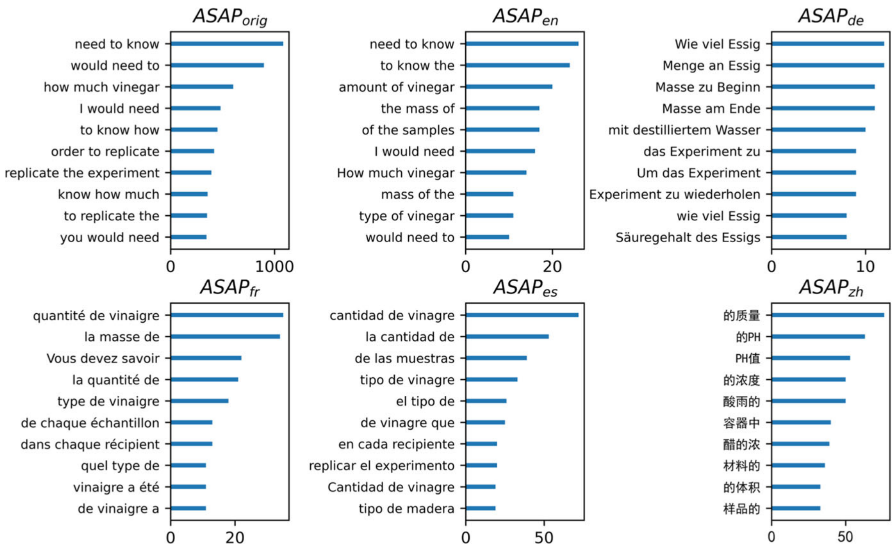


Fig. 6 Top 10 word-level trigrams in different datasets (character-level trigrams for Chinese). (The x-axes are not on the same scale because of the different size of data sets.)

dataset comparisons look rather similar to each other, indicating that the conceptual differences between datasets are not more pronounced than the conceptual differences one finds within a dataset.

Most Frequent Terms and N-Grams

To investigate the idea further that different learner or crowd worker populations simply talk about different concepts when answering the same questions we inspect the most frequent n-grams per dataset and prompt as an easy way to gain insight into which concepts are frequently mentioned.

We focus on content words and thus exclude n-grams that consist only of function words, such as *‘and for the’*. As for some of the previous analyses, we do this in two variants: We first analyze each datasets in its original language. Next, we re-run the analyses on all datasets translated into English. In doing so, we want to make the analyses more accessible to readers unfamiliar with those languages, but also see potentially detrimental effects of machine translation when one term might be consistently translated to English as a term not occurring in the English data set.

Figures 6 and 7 show exemplarily the most frequent 3-grams for Prompt 1 in all languages. We see that terms like *vinegar*, *sample* or *experiment* can be found frequently in all datasets. Many answers in German contain the factor of time, since terms *at the beginning* and *at the end* are only listed as top 10 in this dataset. For answers in German and Chinese, unlike other languages, common terms for the formulation of answers into a whole sentence, repeating phrases from the question, such as *need to know* or *I would need* are rare. This confirms our first observation that answers in German and Chinese are often in bullet points. It is interesting to see that the terms *PH value* and *acid rain* can only be found in the top 10 list in Chinese whereas answers

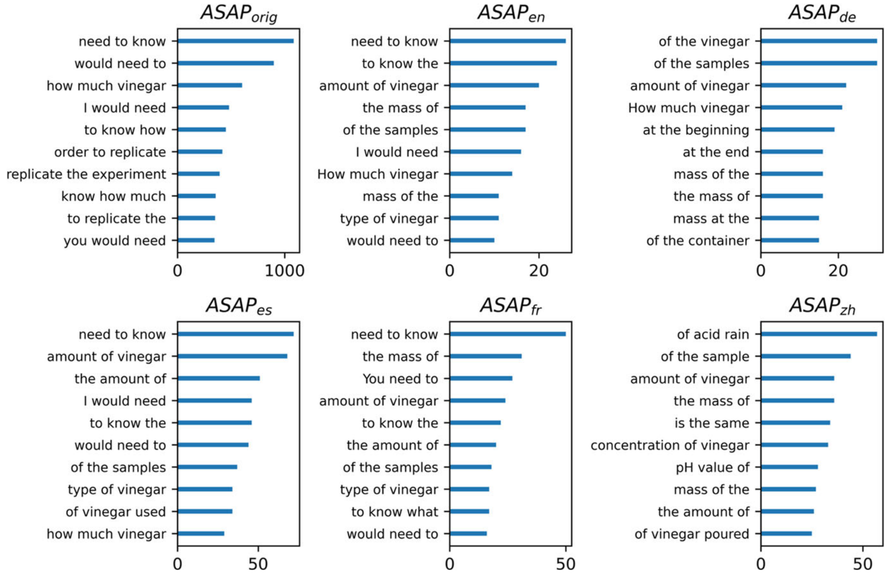


Fig. 7 Top 10 word-level 3-grams in different datasets translated to English

in other languages rather talk about the amount of vinegar and it remains an open question whether such variation reflects cultural differences or different approaches to a reading comprehension task as the term *acid rain* appeared only in the title of the prompt but not in the task itself. This supports our assumption that Chinese data is indeed conceptually different from other datasets.

Conclusion and Future Work

In this paper, we have presented new datasets from three languages addressing the same content scoring prompts as a starting point to explore the field of crosslingual content scoring. We used both machine translation and multilingual transformer models as well as a combination of the two to assess the feasibility of using training data in one language to score test data in another. We found in Study 1 that in concordance with earlier findings machine translation by itself does not introduce noise that makes automated scoring harder. Study 2 revealed that truly crosslingual experiments with training data in a source language and test data in a different target language unsurprisingly work not as well as a monolingual baseline and that the extent of the performance decrease depends on the language pair considered. Closely-related languages (such as the Spanish-French pair) are often easier than pairs that are further apart (such as Chinese paired with any other language in our study). When comparing different methods, a semantic representation works better than one relying on surface information such as n-grams. The best performance could be achieved when using a multilingual transformer model but still translating either the train or test data into the respective other languages. In Study 3, we explored a scenario where a transformer model was iteratively trained on both larger amounts of English data and smaller amounts of in-domain target language data. This setup yielded overall the best results coming close to performance on large in-domain setups in English.

We postulated that differences in behaviour between datasets can have two reasons: language differences and differences between learner populations. One aspect of such population effects is the difference between real learners, such as high school students, and paid crowd workers. In addition, other aspects we could not address here might play a role, such as the cultural background of a learner population. To understand these influences better, more research, ideally under more controlled conditions would be necessary.

Ethical Considerations

The benefits and risks of automated scoring in general have been extensively debated in the literature (see, for example Loukina et al. (2019)). One important argument in favor of the use of automated scoring is that normally all learner answers are scored by the same algorithm. If we use crosslingual scoring where one scoring model per language is trained, learners will be scored by different models depending on their choice of language and thus be subjected to different model biases. A more advisable scenario would thus be the automatic translation of all learner answers into the same language in which a model has been trained, but again with the caveat that machine translation can have different quality for different source languages and learners writing in a less-resourced minority language might be discriminated against by the automated scoring process. Similarly, multilingual transformer models work differently well for different languages and language combination. Thus we would consider crosslingual scoring at the moment not suitable for high-stakes assessment.

Acknowledgements This work was partially conducted at “CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics” of the FernUniversität in Hagen, Germany, and partially within the KI-Starter project “Explaining AI Predictions of Semantic Relationships” funded by the Ministry of Culture and Science, Nordrhein-Westfalen, Germany.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The data used in this paper is publicly available under https://github.com/andreahorbach/CrosslingualScoring/.

Declarations

Conflicts of Interest The authors have no relevant financial or non-financial interests to disclose.

References

Abdul Salam, M., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for arabic short answer questions using optimized deep learning model. *Plos one, 17*(8), e0272269.

Andersen, N., Zehner, F., & Goldhammer, F. (2023). Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning, 39*(3), 841–854.

Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and hebrew nlp for automated assessment of open-ended questions in biology. *International journal of artificial intelligence in education, 33*(1), 1–34.

Artetxe, M., Goswami, V., Bhosale, S., Fan, A., & Zettlemoyer, L. (2023). Revisiting machine translation for cross-lingual classification. arXiv preprint. arXiv:2305.14240

Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. *In Proceedings of the third workshop on innovative use of NLP for building educational applications* (pp. 107– 115)

Ballesteros, L., & Croft, B. (1996). Dictionary methods for cross-lingual information retrieval. *International conference on database and expert systems applications* (pp. 791–801)

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics, 1*, 391–402.

Bexte, M., Horbach, A., & Zesch, T. (2022). Similarity-based content scoringhow 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)

Bexte, M., Horbach, A., & Zesch, T. (2023). Similarity-based content scoring- a more classroom-suitable alternative to instance-based scoring? *Findings of the association for computational linguistics: Acl 2023* (pp. 1892–1903). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.119

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. *LREC* (pp. 1281–1288)

Camus, L., & Filighera, A. (2020). Investigating transformers for automatic short answer grading. *International conference on artificial intelligence in education* (pp. 43–48)

Çınar, A., Ince, E., Gezer, M., & Yılmaz, Ö. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and information technologies, 25*(5), 3821–3844.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. Retrieved from http://epm.sagepub.com/content/20/1/37.short

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of quantitative linguistics, 17*(2), 94–100.

Cui, Y., Zhu, J., Yang, L., Fang, X., Chen, X., Wang, Y., & Yang, E. (2022). CTAP for chinese: a linguistic complexity feature automatic calculation platform. *In Proceedings of the thirteenth Language Resources and Evaluation Conference* (pp. 5525–5538)

Devlin, J., Chang, M. -W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186)

Ding, Y., Horbach, A., Wang, H., Song, X., & Zesch, T. (2020). Chinese Content Scoring: Open-Access Datasets and Features on Different Segmentation Levels. *In Proceedings of the 1st conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th international joint conference on Natural Language Processing (AACL-IJCNLP 2020)*

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., & Dang, H. T. (2013). *Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge* (Tech. Rep.). North Texas State University Denton

Funayama, H., Asazuma, Y., Matsubayashi, Y., Mizumoto, T., & Inui, K. (2023). Reducing the cost: Cross-prompt pre-netuning for short answer scoring. *International conference on artificial intelligence in education* (pp. 78–89)

Galhardi, L., Barbosa, C. R., de Souza, R. C. T., & Brancher, J. D. (2018). Portuguese automatic short answer grading. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (vol. 29, pp. 1373)

Gomes Rocha, F., Rodriguez, G., Andrade, E. E. F., Guimarães, A., Gonçalves, V., & Sabino, R. F. (2021). Supervised machine learning for automatic assessment of free-text answers. *Advances in soft computing: 20th mexican international conference on artificial intelligence, micai 2021, Mexico City, Mexico, October 25–30, 2021, proceedings, part ii 20* (pp. 3–12)

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. *In Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 81–85)

Herwanto, G. B., Sari, Y., Prastowo, B. N., Bustoni, I. A., & Hidayatulloh, I. (2018). Ukara: A fast and simple automatic short answer scoring system for Bahasa Indonesia. *ICEAP, 2019*(2), 48–53.

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., others (2014). Is getting the right answer just about choosing the right words? the role of syntactically-informed features in short answer scoring. arXiv preprint. arXiv:1403.0801

Horbach, A., Stennmanns, S., & Zesch, T. (2018). Cross-lingual Content Scoring. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 410–419). New Orleans, LA, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W18-0550

Horbach, A., & Zesch, T. (2019). The influence of variance in learner answers on automatic content scoring. *Frontiers in education* (vol. 4, pp. 28)

Klebanov, B. B., & Madnani, N. (2021). Automated essay scoring. *Synthesis Lectures on Human Language Technologies, 14*(5), 1–314.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012* (pp. 1459–1474)

Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019). Get it scored using autosas — An automated system for scoring short answers. Proceedings of the AAAI conference on artificial intelligence (pp. 9662–9669)

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., & Walter, M. (2008). *Das Lernerkorpus Falko. Deutsch als Fremdsprache, 45*(2), 67.

Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications* (pp. 1–10)

Megyesi, B., Rudebeck, L., & Volodina, E. (2021). SweLL pseudonymization guidelines

Mendes, A., Antunes, S., Jansseen, M., & Gonçalves, A. (2016). The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings of the tenth language resources and evaluation conference–LREC'16* (pp. 3207–3214)

Meurers, D., Ziai, R., Ott, N., & Kopp, J. (2011). Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 workshop on textual entailment* (pp. 1–9)

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th conference of the European Chapter of the Association for Computational Linguistics* (pp. 567–575). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1609067.1609130

Nguimkeng, P. J. (2021). *Cross-lingual content scoring with a focus on French* (Bachelor's Thesis). University of Duisburg-Essen

Ouahrani, L., & Bennouar, D. (2020). AR-ASAG an Arabic dataset for automatic short answer grading evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 2634–2643)

Pado, U., & Kiefer, C. (2015). Short answer grading: When sorting helps and when it doesn't. In *Proceedings of the fourth workshop on NLP for computer-assisted language learning* (pp. 42–50)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Peña-López, I., et al. (2012). PISA 2012 assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*

Prettenhofer, P., & Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1118–1127)

Sawatzki, J., Schlippe, T., Benner-Wickner, M. (2021). Deep learning techniques for automatic short answer grading: Predicting scores for english and german answers. *International conference on artificial intelligence in education technology* (pp. 65–75)

Schlippe, T., & Sawatzki, J. (2022). Cross-lingual automatic short answer grading. *Artificial intelligence in education: Emerging technologies, models and applications* (pp. 117–129). Springer

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.

Shi, L., Mihalcea, R., Tian, M. (2010). Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 conference on empirical methods in Natural Language Processing* (pp. 1057–1067)

Vajjala, S., & Rama, T. (2018). Experiments with universal CEFR classification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 147–153)

Volodina, E., Pilán, I., Alfter, D., et al. (2016). Classification of Swedish learner essays by CEFR levels. *CALL communities and culture-short papers from EUROCALL, 2016*, 456–461.

Walia, T. S., Josan, G. S., & Singh, A. (2019). An efficient automated answer scoring system for punjabi language. *Egyptian Informatics Journal, 20*(2), 89–96.

Weegar, R., & Idestam-Almquist, P. (2023). Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education*, 1–27

Wijaya, M.C. (2021). Automatic short answer grading system in indonesian language using bert machine learning. *Revue d'Intelligence Artificielle, 35*(6)

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and psychological measurement, 76*(2), 280–303.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.