Cross-Lingual Content Scoring

Andrea Horbach, Sebastian Stennmanns, Torsten Zesch
Language Technology Lab, Department of Computer Science and Applied Cognitive Science,
University of Duisburg-Essen, Germany
{andrea.horbach|torsten.zesch}@uni-due.de
sebastian.stennmanns@stud.uni-due.de

Abstract

We investigate the feasibility of cross-lingual content scoring, a scenario where training and test data in an automatic scoring task are from two different languages. Cross-lingual scoring can contribute to educational equality by allowing answers in multiple languages. Training a model in one language and applying it to another language might also help to overcome data sparsity issues by re-using trained models from other languages. As there is no suitable dataset available for this new task, we create a comparable bi-lingual corpus by extending the English ASAP dataset with German answers. Our experiments with cross-lingual scoring based on machine-translating either training or test data show a considerable drop in scoring quality.

1 Introduction

Automatically scoring the content of student answers is a well-established research field (see, e.g., Sukkarieh and Blackmore (2009); Ziai et al. (2012); Higgins et al. (2014)). However, content scoring is usually restricted to training a model on labeled answers in one language and then applying it to unseen student answers in the same language. In this paper, we examine how well the scoring models transfer when being applied cross-lingually, i.e., whether data in one language can be used for training a model to score data in another language.

The motivation for our study is two-fold: First, cross-lingual scoring can contribute to **educational equality**. In a realistic educational setting, scores assigned to an answer given in the language of instruction can discriminate against non-native students who might conceptually understand the topic in question, but are unable to express their understanding in that language. 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. Such a setting would of course 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 for the student. In such a scenario, automatic scoring of answers in different languages can help to treat students equally.

Second, cross-lingual scoring can help to **overcome data sparsity**. Existing short-answer datasets have mainly been collected in English. If a researcher or practitioner wants to work on a different language, little annotated data is available. Cross-lingual approaches can help in such a scenario to re-use trained models from different languages or to combine data from several languages to train a new model.

In our study, we investigate whether cross-lingual scoring is possible using state-of-the-art machine translation techniques. We translate either training or test data from one language to another, such that both training and test data are available in the same language. We then build prompt-specific models for each prompt and compare the performance to a monolingual approach. Figure 1 illustrates the different approaches.

It is likely that machine translation will negatively impact scoring quality due to translation errors. Additionally, student answers often contain language errors that might further decrease translation quality. However, translation might also have a positive effect on automatic scoring in case of typos being corrected during translation (e.g. *sep<u>e</u>rate* correctly translated as *getrennt*).

Datasets in more than one language might also differ depending on different teaching or learning traditions in the environments where they are collected, so that a new dataset collection has to be carefully planned to control such influence factors.

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 410–419
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

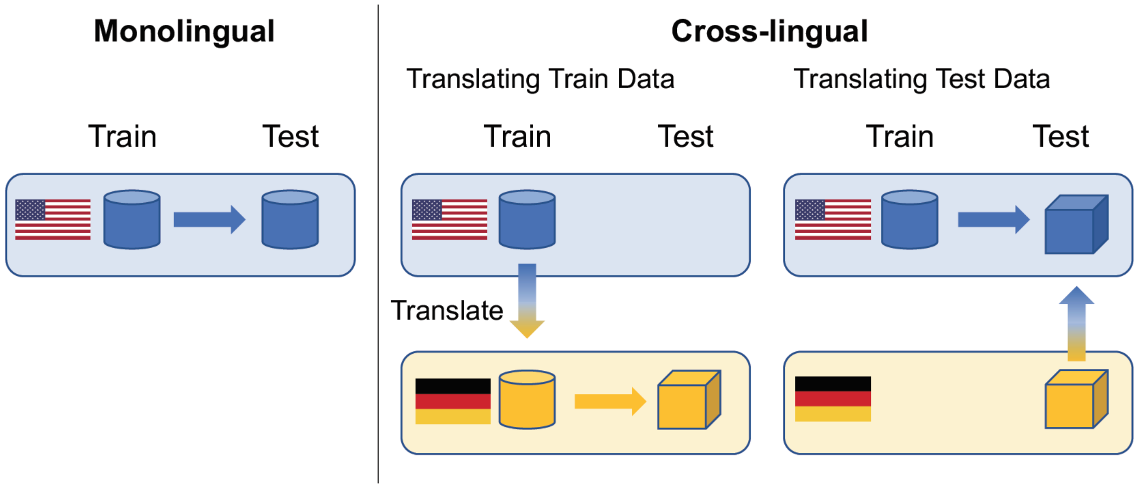

Figure 1: Monolingual vs. cross-lingual scoring

To the best of our knowledge, we are the first to investigate the feasibility of cross-lingual scoring. As our approach relies heavily on the availability of machine-translation methods, we also assess whether state-of-the-art machine translation methods perform well enough to be used in automatic scoring. To evaluate cross-lingual scoring in a realistic scenario, we collect and release a new dataset ASAP-DE that consists of three prompts from the ASAP corpus for which we collect answers in German.[1] In our experiments, we find that cross-lingual scoring using machine-translation is feasible, but –unsurprisingly– at the cost of a decrease in performance. Preliminary analyses showed that his performance drop varies across prompts and is only in part due to artifacts of machine translation, but it rather results from differences between the two datasets involved.

2 Pilot Study

Machine translation nowadays has good quality in general, but we need to assess its performance with respect to the language used in content scoring datasets. In contrast to standard newspaper data, answers in such datasets have been written by non-professional writers, so they may contain typos and ungrammaticalities. These datasets can thus be harder to translate than newspaper text.

To examine the impact of these issues, we conduct monolingual scoring experiments with the English ASAP dataset. We translate both the training and the test section of the ASAP data into a different language and build and train a model in that language. For the moment, we do not change the score an answer receives after translation because we assume that translating an answer preserves its meaning. We will revisit this issue later.

We automatically translate the English ASAP dataset using two different translation frameworks: Google Translate API[2] and DeepL[3]. As target languages, we use German as a closely related language and Russian as a more distantly related language. Table 2 shows the results of a state-of-the-art scoring system applied in this setting measured as quadratically weighted kappa.. We can see that there is a performance drop when translating to a different language, but that the change is within a reasonable margin, such that we can assume that machine translation is good enough for our purposes. We select Google Translate for all further experiments in this paper, as it produces on average better results than DeepL.

Influence of Spelling Errors Translating misspelled words is especially challenging, and we expect two different types of influence on the results. There could be a normalizing effect when wrong forms are translated into correct ones, or a noise-introducing effect when a wrong form from one language leaks into the other. We observe both effects in the data. First, some errors are corrected by the translation, mainly for very common misspellings, which a machine translation system might have encountered during training, such as

[1] https://github.com/ltl-ude/crosslingual

[2] https://cloud.google.com/translate/

[3] https://www.deepl.com/home

Domain	Prompt ID	En	De^T Google	De^T DeepL	Ru^T Google
Science	1	.72	.69	.64	.66
	2	.68	.64	.52	.57
	10	.65	.66	.67	.64
Biology	5	.75	.70	.70	.71
	6	.80	.75	.73	.79
ELA	3	.59	.60	.54	.59
	4	.66	.67	.54	.60
	7	.62	.50	.50	.55
	8	.51	.53	.50	.53
	9	.75	.75	.70	.71

Table 1: Monolingual scoring

Domain	Prompt ID	En	De^T Google	De^T DeepL	Ru^T Google
Science	1	.69	.69	.63	.65
	2	.69	.64	.52	.57
	10	.66	.64	.68	.64
Biology	5	.75	.71	.69	.70
	6	.81	.76	.74	.79
ELA	3	.59	.58	.53	.58
	4	.66	.67	.54	.60
	7	.62	.47	.50	.54
	8	.51	.53	.47	.54
	9	.75	.76	.69	.71

Table 2: Monolingual scoring

seperate instead of *separate*, which are both correctly translated to the German *getrennt*. Second, the less frequent misspellings are often preserved, although nouns are capitalized and inflected in German. An example would be the phrase *the temperature of vineger* which is translated to *die Temperatur des Vinegers* using the correct German inflected form, but not translating the word to the correct German *Essig*. Another is the misspelled word *diffrence* which is translated to the similarly misspelled form *Diffrenz* (instead of *Differenz*), i.e., the affix *-ence* is correctly translated while keeping the misspelled stem of the word.

Influence of Translation on Human Scores So far, we have simply assumed that the machine translation process is good enough that it does not affect the score assigned to an answer. We examine whether this assumption is valid by re-scoring a small sample of 50 answers each from ASAP prompts 1, 2 and 10, which have been machine translated to German. The annotator also scored the original English data (with some delay time in between to avoid memory effects) so that we can compare scores in different languages assigned by the same annotator. We found the annotation to be consistent between different language versions. (Quadratically weighted kappa (Cohen, 1968) of between .75 and .94 for the agreement of the same annotator between the original version and the one translated using google translate. Inter-annotator agreement on this sample between the original annotation and our annotator is between .66 and .84.) If machine translation introduced a lot of noise, one would expect scores to differ more between the two versions. One would especially assume that translated answers might make less sense, and would therefore receive lower scores, but we do not see such a phenomenon in the data.

3 Collecting a Cross-lingual Dataset

For our cross-lingual experiments, we need a dataset that contains answers to the same prompt in at least two different languages. As no such dataset is publicly available so far, we decided to create and release a new dataset.

3.1 Selecting a Source Dataset

We decided to extend an existing monolingual dataset instead of collecting a new dataset from scratch, as it provides the advantage that larger amounts of data are already available in one language. The majority of datasets is available in English, so this is a realistic option for the source language. We use German as the target language due to familiarity with the language, as we need to be able to manually score the new dataset. Also, the expected translation quality between English and German is rather high providing a good test case for the feasibility of the approach in general.

There is a set of publicly available English datasets that we could base our experiments on: The ASAP-2 short answer scoring dataset [4], the Powergrading dataset by (Basu et al., 2013), the computer science dataset by (Mohler and Mihalcea, 2009), and the SemEval2013 dataset (Dzikovska et al., 2013). When deciding for a dataset, we took the following criteria into account: First, all necessary **prompt material has to be completely available**, including reading texts or connected images. This requirement rules out the SemEval2013 data, where the prompt contains pictures and graphs (such a drawing of a electrical circuit) that are necessary to answer the questions but that are not included in the dataset.

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

Second, the prompts should be **language and culture-independent** so that speakers of a different language or from a different culture have similar chances to answer the questions correctly. This requirement rules out the Powergrading data, as this dataset contains solely questions from US immigration tests like, *If both the President and the Vice President can no longer serve, who becomes President?* German participants are rather unlikely to correctly answer those questions.

Third, the prompts should be **curriculum-independent**, i.e., they should not be based on a specific university course, as we expect answers in those settings to be heavily influenced by what exactly was taught in the corresponding course. Thus, we excluded the computer science dataset, which was targeted at students from a specific computer science class. (In addition, the number of only 30 answers per prompt is relatively small.)

Last, in order to be able to score the newly collected data, **scoring guidelines** for the original dataset have to be **available** and we must be able to apply them with a reasonable inter-annotator-agreement.

Re-scoring Study The ASAP dataset is the only dataset fulfilling the first two requirements and seems relatively curriculum-independent as well. We tested in an annotation study, whether we are able to apply the available annotation guidelines. We selected one prompt for each of the three domains covered by the dataset (science, biology, English Language Arts (ELA)). Two German native speakers with a good command in English annotated a subset of 50 answers for each prompt. For the science prompt, the pairwise inter-annotator agreement between our two annotators and the original English annotators, measured by quadratically weighted kappa, was between .70 and .79 for the science prompt, between .60 and .78 for biology, and between .26 and .63 for ELA. IAA between the two German annotators lies in similar regions. The agreement between the two original annotations was .95 for science, .98 for biology and .77 for ELA. Based on these numbers, we deemed ELA prompts unsuitable for re-collection.

3.2 Dataset Collection & Annotation

As described above, we find the science and biology prompts from ASAP to be suitable for the re-collection process. An exploratory data collection for the three science and two biology prompts revealed that the knowledge tested in the biology prompts was more course-specific than we thought and most participants were unable to answer these questions. Therefore, we restricted ourselves to the three science prompts, which we translated into German. We collect answers from the crowdsourcing platform CrowdFlower,[5] as well as by directly asking colleagues and students, with the majority of answer (>90%) originating from CrowdFlower. We excluded answers in any language different from German and obvious non-answers, such as copying the prompt.[6] Overall, we collect a total of 301 answers per prompt. Table 3 compares the resulting German dataset with the original English one.

	ASAP	**ASAP-DE**
Language	English	German
#Prompts	10	3
#Answers / prompt	>2000	300
Domains	Science ELA Biology	Science

Table 3: Dataset statistics

All answers have been annotated by two German annotators (one being one of the authors of this paper). We found an inter-annotator agreement per prompt between .58 and .84 quadratically weighted kappa. Figure 2 shows some exemplary answers from Prompt 1 both for the original English and the newly collected German dataset.

3.3 Dataset Analysis

We provide a corpus analysis to get further insights into the differences between the two language versions of the dataset.

Label distribution A first indicator as to whether the two language versions are comparable is the label distribution as shown in Table 3. We see that the distribution in the German dataset is skewed towards lower scores, which could be an artifact of our assessment situation. While we tried to avoid questions answerable only by a certain group of learners, it might still be that the original English test taker population was either better prepared or more motivated to answer the ques-

[5] https://www.crowdflower.com

[6] We needed to do so because of a relatively high number of such non-answers. However, we kept other non-answers such as *"Ich weiß es nicht"* (*I don't know.*)

English

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

Learner Answers:

- **3 points:** Some additional information you will need are the material. You also need to know the size of the contaneir to measure how the acid rain effected it. You need to know how much vineager is used for each sample. Another thing that would help is to know how big the sample stones are by measureing the best possible way.
- **1 point:** After reading the expirement, I realized that the additional information you need to replicate the expireiment is one, the amant of vinegar you poured in each container, two, label the containers before you start yar expirement and three, write a conclusion to make sure yar results are accurate.
- **0 points:** The student should list what rock is better and what rock is the worse in the procedure.

German

Question: Nachdem Sie die Prozedur der Gruppe gelesen haben, beschreiben Sie, welche zusätzlichen Informationen nötig sind, um das Experiment zu wiederholen. Geben Sie mindestens drei benötigte Informationen an.

Learner Answers:

- **3 points:** Es fehlt der Säuregehalt des Essigs. Die Menge Essig die verwendet wurde. Und welche Holzart da Holzsorten unterschiedliche Säureresistenz aufweist.
- **2 points:** Wie viel Essig wurde verwendet? Aus welchem Material waren die Behälter? Wurden die Behälter verschlossen?
- **0 points:** Wir müssen wissen, wie viel Wasser wir sammeln müssen, um die Probe zu machen

Figure 2: Exemplary answers for prompt 1 from the English and the German datasets.

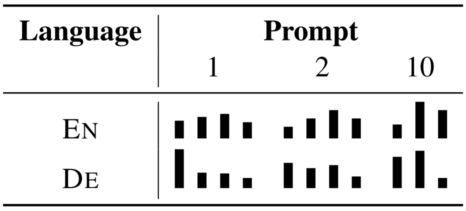

Figure 3: Label distribution for each prompt in the German and English version of the data.

tions correctly than the crowd-workers providing the German answers.

Average Length Figure 4 shows that answers in the English dataset are considerably longer than in the German one. This difference can be due to two parameters. One is the learner population from which the data is collected, the other is idiosyncrasies of the language itself. To differentiate between the influence of these two effects as far as possible, we also run our comparisons on versions of each dataset that have been automatically translated into the other language (EN^T and DE^T). Thus, comparing the English dataset to DE^T should only display effects of having different datasets, not different languages, while comparing the English dataset to EN^T should show differences between languages but is the same data.

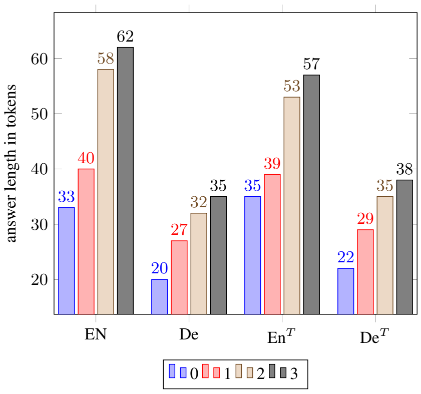

Figure 4: Answer length in tokens averaged over all answers with a certain score.

Figure 4 shows that the difference in length observed between English and German is not an effect of the different languages, but of the different datasets. Additionally, we observe in both datasets that answers with a higher score tend to be longer than incorrect answers.

Linguistic diversity Next, we look at the linguistic diversity in both datasets. We compute the type-token-ratio (TTR) for each dataset, by randomly sampling chunks of 100 tokens and averaging over the individual values to avoid effect of different corpus sizes, shown in Figure 5. The two main findings from this analysis are: First,

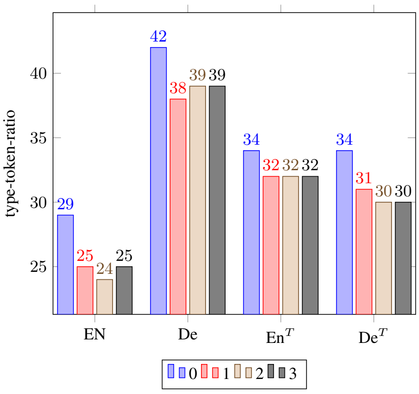

Figure 5: Type-token-ratio for the four datasets computed for all labels with a particular score.

Compared datasets	Prompt 1	2	10
EN (train) / EN (test)	.68	.68	.62
EN (train) / DE^T	.45	.49	.45
EN^T (train) / DE	.43	.48	.43

Table 4: Lexical overlap measured on the type level for the top 1000 unigrams for each prompt.

TTR is slightly higher for low scoring answers. This fits the observation that there are often more ways to get an answer wrong than ways to formulate the correct answer. (Note that annotators ignore spelling errors when scoring an answer, and we found that low-scoring answers do not contain more errors than high-scoring answers.) Second, TTR is higher for the original German than for the original English dataset. This is in part due to the language difference. German has a much richer morphology than English, and translating data from German to English reduces TTR while translating from English to German increases it.

Some part of the difference, however, cannot be explained by the different languages and must come from the learner population, which is more homogeneous in the English version (high-school students) as compared to the German version (crowd-workers).

Vocabulary overlap Here, we compute the overlap between the vocabulary used in the English data and the vocabulary of the German dataset. Table 4 shows the comparison measuring the overlap of types on the unigram level. As a baseline, we compute the overlap between training and test data from the English dataset. Next we compare the English training data with the German dataset by either translating the English or the German data to the respective other language. We find a much lower lexical overlap across all prompts.

We therefore expect a decrease in performance when using n-grams as features in a cross-lingual setup compared to the monolingual case.

Summary Overall we observe differences between the datasets in terms of answer length, label distribution, linguistic diversity and used vocabulary. They can only be partially explained by the language difference and seem to be mostly due to differences between the datasets themselves or rather between the learner populations that produced theses answers. In the next section, we examine the effect these differences have on automatic scoring.

4 Cross-lingual Scoring Experiments

After finding in the previous monolingual pilot study that machine translation quality is good enough for our purposes, we now present in this section our cross-lingual experiments. We assume that training data in one language is used to score test data in another language by means of translating either the test or the training data.

4.1 Experimental setup

For our scoring experiments, we use a standard supervised machine learning setup with Weka's SVM classifier in standard configuration as classification backbone, implemented using free-text scoring toolkit ESCRITO (Zesch and Horbach, 2018). We use token uni-, bi- and trigrams as well as character bi- to five-grams as features and evaluate our results using accuracy and quadratically weighted Kappa (Cohen, 1968).

The English ASAP dataset comes with an established split into train and test data, which we reuse. The German dataset is very small in direct comparison, so that we cannot use a fixed split into training and test data. Therefore, we use 10-fold cross-validation for the German dataset.

Experimental conditions We conduct experiments falling into four groups:

(1) for the **baseline** experiments, we train and test models on monolingual datasets and use either the English or German dataset exclusively. These two datasets have very different sizes. For the original English data, we have over 2000 answers per prompt. For the re-collected German set, we

only have 300 answers per prompt, 270 of which are used for training in our cross-validation setup. This difference in size might also reflect in different performances. To eliminate such effects, we conduct experiments on the English training data in a variant that uses only 270 training items, sampled from the training data section. For comparison, we also conduct the baseline experiment on the full English train data (EN_{all}). To avoid sampling artifacts, we repeat the experiment 100 times with different splits and report the average of all runs.

(2) In the **monolingual** condition, we translate both the training and the test data, similar to our experiments in the pilot study, but using data sampling that makes sure that training data sizes are comparable. Differences to the baseline are thus only due to the machine translation process.

(3) In the **translate train** experiments, we combine the original English test data with the German training data automatically translated to English, as well as original German test data with the English training data translated to German.

(4) In the **translate test** condition, we use test data translated into the other language with the original test data from that language. In these last two conditions, differences to the baseline result either from machine translation or from differences inherent to the datasets.

			QWK			
	Train	**Test**	**1**	**2**	**10**	**∅**
	EN_{all}	EN	.72	.68	.65	.68
baselines	EN	EN	.64	.56	.64	.61
	DE	DE	.78	.61	.63	.67
translate both	EN^T	EN^T	.63	.49	.64	.58
	DE^T	DE^T	.84	.62	.54	.66
translate train	EN^T	DE	.49	.08	.46	.34
	DE^T	EN	.41	.39	.39	.40
translate test	EN	DE^T	.35	.08	.43	.29
	DE	EN^T	.26	.35	.33	.31

Table 5: Content scoring performance measured in quadratically weighted kappa for different cross-lingual setups.

4.2 Results

Table 5 shows our results measured in quadratically weighted kappa. When looking at the baselines first, we see that automatic scoring in this monolingual case works comparably well for English and German. This shows that our manual scoring of the German data set is reliable enough to learn a competitive model. In the second monolingual case, when we translate both training and test data, we only observe moderate losses or for some prompts even small improvements compared to the original language version.

		QWK			
Train	**Test**	**1**	**2**	**10**	**∅**
EN	EN	.64	.56	.64	.61
EN	EN^{2T}	.50	.40	.55	.48
EN^{2T}	EN	.64	.52	.62	.60

Table 6: Double translation in monolingual setting

When turning towards the cross-lingual results, where we either only translate train or test data, the picture looks quite different: in all four conditions, scoring performance is considerably lower compared to the monolingual settings. The loss is especially pronounced for prompt 2. This difference between prompts cannot be explained by our corpus analysis in Section 3, especially the vocabulary overlap between English and German datasets, which were in the same range for all three prompts (and even slightly higher for prompt 2 than for the other two prompts).

Differences between Prompts To investigate the apparent differences in similarity between training and test data for the individual prompts further, we analyze the data using language models. We build a trigram language model per prompt for the English data using the SRILM toolkit (Stolcke, 2002) and measure the perplexity of translated German answers under that language model. We find the perplexity of answers to prompt 2 to be higher than answers to prompt 1 and 10, indicating that German answers to prompt 2 fit the model of the English answer worse than the other prompts. Considering that using n-grams as classification features as well as for language models are quite related tasks, these results are not surprising but do not provide a full explanation to our observations. Further investigations into the differences between prompts are definitely necessary

4.3 Follow-up Experiment: The Influence of Machine Translation

As discussed in the introduction, the difference between the baseline and cross-lingual scoring performance can originate from two sources: different learner populations and effects of machine translation. In order to assess the individual contributions of these two factors, we propose a variant of our experiment that operates on only one dataset but still uses machine translation on either the test **or** the training data, so that the delta in performance is due to translation and not to different learner populations. We achieve this by *double-translating* the training or test data of the English ASAP dataset, i.e., we have the data automatically translated from English to German and then back to English (marked as EN^{2T}). Table 6 shows the performance in comparison to the monolingual baseline experiments where we see that double-translating the test data decreases performance considerably while –surprisingly– double-translating the training data leaves performance unaffected.

A naive approach to factor out artifacts from translationese, while keeping effects stemming from the differences between the datasets, would be to use translated datasets in the cross-lingual case both for training and testing, i.e., we double-translate one dataset and translate the other one only once. In this setup, shown in Table 7, performance benefits only slightly, if at all, from double-translation (with the exception of double translated train data in prompt 1).

Consider the following example of an answer from the original English dataset:

> *(A) Plastic type B was the superior in both trial 1 and trial 2. (B) Record the weight that was put on to show how much effected each plastic. Also conducting more trials (...)*

After translating the answer automatically to German and back to English it looks like this:

> *Type B plastic was the supervisor in both Trial 1 and Trial 2. (B) Write down the weight that was put on to show how much each one has made plastic. Also do more experiments (...)*

Apart from obvious translation errors (*superior–supervisor*), we see a simplifying effect of translation: *record–write down*, *effect–make*, and *conduct–do*. Such simplifications might on the one hand normalize over different paraphrases of the same content, but could on the other hand also remove meaningful differences between correct and incorrect answers.

			QWK			
	Train	**Test**	**1**	**2**	**10**	**∅**
translate train	EN^{T}	DE	.49	.08	.46	.34
	EN^{T}	DE^{2T}	.49	.07	.46	.34
	DE^{T}	EN	.41	.39	.39	.40
	DE^{T}	EN^{2T}	.43	.36	.44	.41
translate test	EN	DE^{T}	.35	.08	.43	.29
	EN^{2T}	DE^{T}	.55	.03	.46	.35
	DE	EN^{T}	.26	.35	.33	.31
	DE^{2T}	EN^{T}	.41	.38	.32	.37

Table 7: Double translation in cross-lingual setting

5 Related Work

To the best of our knowledge, there are no previous approaches to cross-lingual scoring in the educational domain. However, cross-lingual NLP approaches have been successfully used for a variety of tasks, including information retrieval (Oard and Diekema, 1998), sentiment analysis (Mihalcea et al., 2007) and textual similarity (Mohammad et al., 2007; Potthast et al., 2008). While in some of these approaches, dictionaries are used as the bridge the gap between languages to translate search queries (e.g. by Ballesteros and Croft (1996) for cross-lingual information retrieval) or translate features in a learned model ((Shi et al., 2010)), many approaches rely on having similar training data in both languages, often by means of parallel or comparable corpora (Gliozzo and Strapparava, 2006). If such corpora are not available, as is the case for our scenario, leveraging machine translation to create training data for handling a new language or to transfer test data into a language for which training data exists has been explored for example by Fortuna and Shawe-Taylor, while other approaches use cross-lingual word embeddings (Klementiev et al., 2012).

6 Conclusion

In this paper we showed the general feasibility of cross-lingual short-answer scoring. We also identified a number of challenges: One is that artifacts from machine translation seem to produce a language that is substantially different from genuine text, and that this translationese poses a problem, as highlighted by our experiments with double-translated items. Second, the two datasets bear differences that go beyond differences in language. In a real-life application scenario, this problem might be less severe, e.g. in a class where everyone received the same instructions and just answers an exam in different languages, where answers can be expected to be more consistent than the two versions of the ASAP corpus in our experiments.

For future work, we want to explore more sophisticated approaches going beyond our straightforward procedure of automatically translating test or training data, such as translation of word features or using cross-lingual embeddings in a neural network approach as well as extending our experiments to a broader variety of data.

Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research under grant no. FKZ 01PL16075.

References

Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *International Conference on Database and Expert Systems Applications*. Springer, pages 791–801.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics (TACL)* 1:391–402.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213.

Myroslava O. Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. **SEM 2013: The First Joint Conference on Lexical and Computational Semantics* .

Blaz Fortuna and John Shawe-Taylor. ???? The use of machine translation tools for cross-lingual text mining. In *Proceedings of the Workshop on Learning with Multiple Views*.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 553–560.

Derrick Higgins, Chris Brew, Michael Heilman, Ramon Ziai, Lei Chen, Aoife Cahill, Michael Flor, Nitin Madnani, Joel Tetreault, Daniel Blanchard, et al. 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* .

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012* pages 1459–1474.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. pages 976–983.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Michael Mohler and Rada Mihalcea. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '09, pages 567–575. http://dl.acm.org/citation.cfm?id=1609067.1609130.

Douglas W Oard and Anne R Diekema. 1998. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)* 33:223–56.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A wikipedia-based multilingual retrieval model. In *European conference on information retrieval*. Springer, pages 522–530.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 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*. Association for Computational Linguistics, pages 1057–1067.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *In proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*. pages 901–904.

Jana Zuheir Sukkarieh and John Blackmore. 2009. C-rater: Automatic content scoring for short constructed responses. In *FLAIRS Conference*. pages 290–295.

Torsten Zesch and Andrea Horbach. 2018. ESCRITO An NLP-Enhanced Educational Scoring Toolkit. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, pages 190–200.