Zero-shot Cross-lingual Automated Essay Scoring

Xia Li and Junyi He*
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China
{xiali,junyihe}@gdufs.edu.cn

Abstract

Due to the difficulty of creating high-quality labelled training data for different languages, the low-resource problem is crucial yet challenging for automated essay scoring (AES). However, little attention has been paid to addressing this challenge. In this paper, we propose a novel zero-shot cross-lingual scoring method from the perspectives of pretrained multilingual representation and writing quality alignment to score essays in unseen languages. Specifically, we adopt multilingual pretrained language models as the encoder backbone to deeply and comprehensively represent multilingual essays. Motivated by the fact that the scoring knowledge for evaluating writing quality is comparable across different languages, we introduce an innovative strategy for aligning essays in a language-independent manner. The proposed strategy aims to capture shared knowledge from diverse languages, thereby enhancing the representation of essays written in unseen languages with respect to their quality. We include essay datasets in six languages (Czech, German, English, Spanish, Italian and Portuguese) to establish extensive experiments, and the results demonstrate that our method achieves state-of-the-art cross-lingual scoring performance.

Keywords: cross-lingual automated essay scoring, writing quality alignment, multilingual representation

1. Introduction

Automated essay scoring (AES) is the task of automatically scoring essays according to writing quality evaluated by language processing techniques (Page, 1966; Foltz et al., 1999; Attali and Burstein, 2004; Dong et al., 2017; Chen and Li, 2023). It is widely studied in the fields of language education (Ke and Ng, 2019), language testing (Shin and Gierl, 2020; Latifi and Gierl, 2020) and writing assessment (McNamara et al., 2015).

To automatically score essays in a specific target language, a general approach is to train a monolingual scoring model using essays in the same language manually rated by experienced human raters (Dong et al., 2017; Hirao et al., 2020; Yupei and Renfen, 2021; He et al., 2022; Chen and Li, 2023). However, the creation of high-quality rated (labelled) training data requires professional grading knowledge and considerable time (Shermis, 2014; Mathias and Bhattacharyya, 2018; Ke and Ng, 2019), resulting in the insufficiency of AES datasets for many languages. This low-resource problem in monolingual AES highlights the necessity of zero-shot cross-lingual AES, which aims at training a scoring model with rated essays in several source languages and using it to score unrated essays in an unseen target language (Figure 1). In this manner, cross-lingual AES alleviates data insufficiency by reducing the requirement for labelled data in low-resource target languages.

Despite its importance, cross-lingual AES has received little research attention. Existing cross-lingual AES studies mainly focus on feature-based and translation-based statistical methods. Specifically, Vajjala and Rama (2018), del Río (2019) and Arhiliuc et al. (2020) use word count, part-of-speech (POS) n-grams, dependency n-grams and other hand-crafted features to train a monolingual scoring model, and directly use it to score essays in unseen languages. Horbach et al. (2018) propose using machine translation techniques to translate multilingual essays into a common language, and training a scoring model on the translated essays using token and character n-grams in a monolingual manner. While being simple and intuitive, these approaches have several limitations. First, the extraction of hand-crafted features necessitates the use of language-specific textual

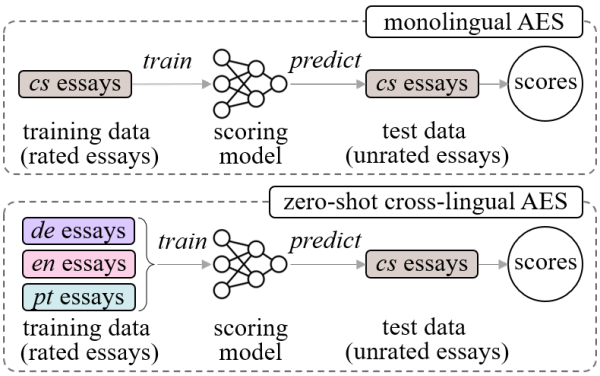

Figure 1: Illustration of monolingual and zero-shot cross-lingual AES, where cs, de, en, and pt denote the Czech, German, English, and Portuguese.

* Corresponding author.

LREC-COLING 2024, pages 17819–17832

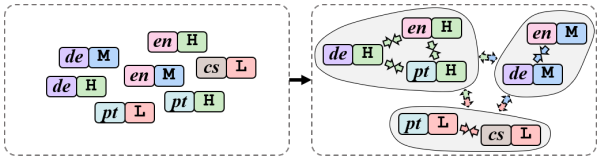

Figure 2: Our method captures language-independent scoring knowledge by aligning writing quality, where L, M and H denote low, medium and high writing quality.

analysis tools, such as text parsers and spell checkers. However, these tools may not be readily available for certain languages, particularly for low-resource languages. Second, POS n-grams, dependency n-grams and other hand-crafted features merely represent essay characteristics from limited perspectives, resulting in the lack of rich multi-dimensional textual information for evaluating essay quality, such as semantic coherence and contextual nuance. Third, due to language discrepancy and machine translation noises, the translation-based approach may alter specific elements of the original essays, such as tone and sentence structure. Hence, the translated essays may not faithfully reflect the original writing quality, leading to unreliable scoring results.

We believe that cross-lingual AES presents at least two core challenges. The first challenge lies in the capability to profoundly and comprehensively represent multilingual essays. This capability is required for understanding the lexical and semantic patterns of different languages. It is also necessary for evaluating word choice, sentence linking and other scoring perspectives demanding a mastery of the essay language. The second challenge is to learn rich cross-lingual shared knowledge and transfer it to score essays in unseen languages. Specifically, we argue that some essay grading knowledge of human raters is comparable and transferable across languages despite differences in essay languages. For instance, regardless of language, an essay with rigorous argumentation and a complete structure tends to receive a higher score, while an essay with frequent spelling errors and poor lexical diversity may receive a lower score.

To address these challenges, we propose a novel zero-shot cross-lingual essay scoring method from the perspectives of pretrained multilingual representation and writing quality alignment [1]. For the first challenge, we propose to use multilingual pretrained language models (multilingual PLMs. e.g., Devlin et al. 2019; Conneau and Lample 2019; Conneau et al. 2020) as the encoder backbone for deeply and comprehensively representing essays in different languages. Multilingual PLMs acquire a strong multilingual understanding ability from the pretraining on massive multilingual corpora, and are capable of effectively tokenizing and representing multilingual essays in a unified manner (Pires et al., 2019; Wu and Dredze, 2019). For the second challenge, we introduce a novel strategy to align the writing quality of multilingual essays in a language-independent manner. As shown in Figure 2, essays with identical writing quality but in different languages initially have different representations depending on the language. For example, the representation of the high-quality German essay $[de|\mathtt{H}]$ is far from that of the similarly high-quality Portuguese essay $[pt|\mathtt{H}]$. As human raters' essay grading knowledge for writing quality is comparable across languages (e.g., high-quality essays usually have coherent sentences, diverse vocabulary, and a rigorous essay structure), we use contrastive learning to pull essays with the same writing quality together and push essays with different quality apart. For example, after the alignment, the high-quality German essay $[de|\mathtt{H}]$ is close to the high-quality Portuguese essay $[pt|\mathtt{H}]$. In this way, essays in different languages are aligned with respect to their writing quality, and the model is expected to learn quality-aware, language-independent scoring knowledge, thereby enhancing the representation of essays in unseen languages. To summarize, the main contributions of our work are as follows:

- To the best of our knowledge, this is the first attempt in zero-shot cross-lingual AES to explore the learning of shared cross-lingual knowledge by introducing a writing-quality aligning strategy and using multilingual PLMs as a shared encoder.
- We conduct zero-shot cross-lingual scoring experiments with essays in up to six languages (Czech, English, German, Italian, Portuguese, and Spanish), and establish comprehensive statistical and neural baselines. Experimental results show that our approach achieves state-of-the-art, demonstrating its cross-lingual scoring effectiveness.
- We also analyze the impact of source language diversity and language similarity on cross-lingual AES.

2. Related Work

2.1. Automated Essay Scoring

Monolingual AES. Existing AES studies have largely focused on English scoring (Page, 1966,

[1]The source code is available at https://github.com/gdufsnlp/XAES.

1994; Mohler and Mihalcea, 2009; Persing and Ng, 2013; Phandi et al., 2015; Taghipour and Ng, 2016; Jin et al., 2018; Uto et al., 2020; Ridley et al., 2021; Wang et al., 2022; Jiang et al., 2023). Additionally, there are also AES investigations on other languages, including Chinese (Song et al., 2020; He et al., 2022), Estonian (Vajjala and Lõo, 2013, 2014), German (Zesch et al., 2015; Horbach et al., 2017), Norwegian (Johan Berggren et al., 2019), Portuguese (Amorim and Veloso, 2017; Amorim et al., 2018), and Swedish (Pilán et al., 2016).

Despite ongoing efforts, existing non-English AES datasets fall short of their English counterparts in size and quality, with many not accessible to the public. Furthermore, AES datasets are still lacking in many languages, including widely used ones like Russian and Urdu. This scarcity of resources underlines the critical role of cross-lingual AES in addressing these limitations.

Moreover, while cross-lingual AES appears similar to cross-prompt AES (an AES research topic aiming at training models to score essays in response to unseen prompts, e.g., Jin et al. 2018; Li et al. 2020; Cao et al. 2020; Ridley et al. 2020, 2021; Chen and Li 2023) in the generalization of scoring ability from existing domains to new domains, cross-lingual AES focuses more on language commonalities rather than prompt commonalities. This distinction underscores the fact that strategies effective for cross-prompt scoring might not be directly applicable to cross-lingual AES, further emphasizing the unique need for approaches specifically designed for cross-lingual contexts.

Cross-lingual AES. Despite the importance of cross-lingual AES, few studies have investigated this task. Existing cross-lingual scoring attempts (Vajjala and Rama, 2018; Horbach et al., 2018; del Río, 2019; Arhiliuc et al., 2020) concentrate on feature-based and translation-based statistical methods. For example, Vajjala and Rama (2018) use word count, POS n-grams and other linguistic features to train a German scoring model, and apply it to score essays in Czech and Italian. Similar to them, del Río (2019) conduct cross-lingual experiments between Spanish and Portuguese with bag-of-words, lexical and other hand-crafted features. Horbach et al. (2018) use translation to convert essays for training and testing into the same language, and use token- and character-level n-grams to train a scoring model.

While straightforward, existing cross-lingual AES approaches are limited by the need for language-specific tools, ineffective essay representation, altering of essay content, little awareness of shared scoring knowledge across languages, and the lack of experiments using essays in multiple languages for model training. In this study, we attempt to address these issues.

2.2. Multilingual Pretrained Language Models

Multilingual pretrained language models (multilingual PLMs), such as multilingual BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-RoBERTa (XLM-R, Conneau et al. 2020) are deep neural networks that acquire the ability to comprehend multiple languages through self-supervised pretraining on massive multilingual corpora. They have exhibited their efficacy in addressing various cross-lingual tasks, such as machine translation (Liu et al., 2020) and natural language inference (Qi et al., 2022). In this study, we explore their ability and effectiveness in cross-lingual AES.

2.3. Contrastive Learning

Contrastive learning (Chen et al., 2020; He et al., 2020; Gao et al., 2021) is a representation learning technique for obtaining class-discriminative representations. Its core idea is to pull semantically similar examples (i.e., positive examples) together in the embedding space, and push semantically dissimilar examples (i.e., negative examples) apart. Its effectiveness has been highlighted in aspect-based sentiment analysis (Li et al., 2021), pretrained language model fine-tuning (Gunel et al., 2021), long context question answering (Caciularu et al., 2022) and other tasks. In this study, we use contrastive learning to distinguish multilingual essays of different writing quality to learn quality-aware cross-lingual shared scoring knowledge.

3. Method

3.1. Task Definition

In zero-shot cross-lingual AES, the scoring model is trained with rated essays in several source languages: $D^{train} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is an essay in a specific source language, y_i is the essay score, and N is the number of source essays. The trained model is tested on unrated essays in a target language not seen during training to perform cross-lingual scoring.

The model architecture is demonstrated in Figure 3. Specifically, each essay $\boldsymbol{x}_i$ is encoded by the essay encoder, producing an essay representation $\boldsymbol{r}_i$. Then, $\boldsymbol{r}_i$ is fed into an essay scorer for score prediction, and simultaneously fed into a cross-lingual contrastive learning (XCL) module for aligning writing quality across languages.

3.2. Essay Encoder

The essay encoder encodes the essay $\boldsymbol{x}_i$ into a latent essay representation $\boldsymbol{r}_i$. There are two typical essay representation approaches in AES studies.

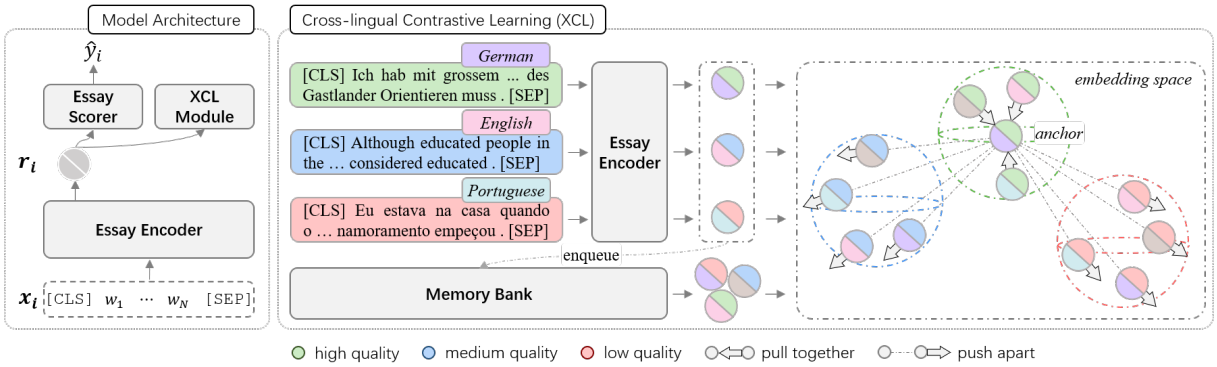

Figure 3: Demonstration of the proposed method. The left part presents the model architecture. The essay $\boldsymbol{x}_i$ is encoded by the essay encoder into an essay representation $\boldsymbol{r}_i$, which is fed into the essay scorer for score prediction, and into the XCL module for writing quality alignment. The right part presents the proposed XCL method, which pulls positive examples (essay representations of the same writing quality as the anchor) towards the anchor in the embedding space, and pushes negative examples (essay representations of different writing quality from the anchor) away from the anchor.

First, essays are split into words/subwords with a monolingual vocabulary, and encoded by a hierarchical network (Dong and Zhang, 2016; Dong et al., 2017; Dasgupta et al., 2018). However, this approach requires a language-dependent monolingual vocabulary, which is inappropriate to extend to cross-lingual scoring. Second, essays are converted into POS and dependency tags before hierarchical encoding (Jin et al., 2018; Ridley et al., 2021; Chen and Li, 2023), which can be extended to cross-lingual scoring by utilizing universal POS or dependency tags of UDPipe (Straka and Straková, 2017). However, this approach only represents essays from a single shallow dimension, e.g., semantic dimension from POS tags, and cannot provide multi-dimensional deep textual information for scoring. To deeply and comprehensively represent multilingual essays, we propose to use multilingual PLMs as the essay encoder. They employ a shared multilingual subword vocabulary, and utilize a universal encoder architecture to capture language-agnostic or cross-lingual deep textual patterns (Pires et al., 2019; Wu and Dredze, 2019), enabling them to understand and represent essays in diverse languages.

Specifically, the input essay $\boldsymbol{x}_i$ is first tokenized into a subword sequence $\boldsymbol{x}'_i = (\texttt{[CLS]}, w_1, w_2, \cdots, \texttt{[SEP]})$ using the multilingual subword vocabulary of the multilingual PLM, where `[CLS]` and `[SEP]` denote the start and end of the sequence, and w_i denotes the i-th subword. After tokenization, the multilingual PLM encoder $mPLM(\cdot)$ encodes the subword sequence $\boldsymbol{x}'_i$ and generates a sequence of hidden representations $\boldsymbol{H}_i = (\boldsymbol{h}_{\texttt{[CLS]}}, \boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{\texttt{[SEP]}})$ as in Equation 1. The first hidden representation $\boldsymbol{h}_{\texttt{[CLS]}}$ ($\boldsymbol{H}_i[0]$), which is treated as the global representation of the input sequence (Devlin et al., 2019), is used as the representation $\boldsymbol{r}_i$ of the essay (Equation 2).

$$\boldsymbol{H}_i = mPLM(\boldsymbol{x}'_i) \quad (1)$$

$$\boldsymbol{r}_i = \boldsymbol{H}_i[0] \quad (2)$$

3.3. Cross-lingual Contrastive Learning

As mentioned above, we assume that certain knowledge for scoring essays is language-independent (e.g., essay structure, lexical diversity), and such knowledge can be aligned across languages for better cross-lingual scoring performance. Driven by this motivation, we propose to align the writing quality of essays with contrastive learning. Specifically, with each essay being the anchor, the positive examples are essays with the same writing quality as the anchor, and the negative examples are essays with different writing quality from the anchor. In contrastive learning, the positive examples (same quality) are pulled toward the anchor, and the negative examples (different quality) are pushed apart from the anchor. As the essays (i.e., positive and negative examples) are multilingual, the model can learn to align writing quality across languages, through which the language-independent shared scoring knowledge is learned.

Writing Quality Derivation. Motivated by Jin et al. (2018), we divide the essay writing quality into three levels: low, medium, and high. Since essays with higher scores are of higher quality and those with lower scores are of lower quality, we derive the writing quality of each essay from its essay score. The essays with scores in the lowest 20% of the score range are of low quality, the highest

20% are of high quality, and the remaining (20%-80%) are of medium quality.

Contrastive Loss. For each essay $\boldsymbol{x}_i$ being the anchor, the cross-lingual contrastive loss (L_{XCL}) is calculated as in Equation 3. $\boldsymbol{r}_i$ denotes the essay representation for $\boldsymbol{x}_i$. $\boldsymbol{r}_p$ denotes a positive example, and P_i denotes the set of positive examples for $\boldsymbol{r}_i$. $\boldsymbol{r}_c$ denotes a contrastive example (either a positive example or a negative example), and C_i denotes the set of contrastive examples for $\boldsymbol{r}_i$. $s(\cdot)$ denotes a similarity function for measuring the similarity between the anchor and a contrastive example. Following existing contrastive learning studies (Chen et al., 2020; He et al., 2020; Gao et al., 2021; Li et al., 2021; Gunel et al., 2021; Caciularu et al., 2022), we apply cosine similarity as $s(\cdot)$. τ denotes the contrastive temperature, a hyper-parameter that controls the difficulty of distinguishing positive and negative examples (Wang and Liu, 2021).

$$L_{XCL} = -\frac{1}{|P_i|}\sum_{\boldsymbol{r}_p}^{|P_i|}\log\frac{\exp(s(\boldsymbol{r}_i,\boldsymbol{r}_p)/\tau)}{\sum_{\boldsymbol{r}_c}^{|C_i|}\exp(s(\boldsymbol{r}_i,\boldsymbol{r}_c)/\tau)} \quad (3)$$

Source of Contrastive Examples. In general, the contrastive examples of an anchor are the remaining examples within the same mini-batch B (Khosla et al., 2020; Li et al., 2021; Chen and Li, 2023), which are referred to as in-batch contrastive examples. However, some studies have found that sufficient contrastive examples are helpful for improving the effectiveness of contrastive learning (He et al., 2020; Tan et al., 2022). To increase the number of contrastive examples and avoid repetitive encoding, we additionally introduce a memory bank (Wu et al., 2018) $Q = \{\boldsymbol{r}_i, q_i\}_{i=1}^{S}$ with size S to reuse the examples from previous batches for contrastive learning. Specifically, after each iteration, all essay representations of the current mini-batch and their corresponding writing quality labels are enqueued into the memory bank. If the memory bank is full, the latest S examples will be kept and the earliest examples will be discarded. As a result, the contrastive examples of each essay anchor $\boldsymbol{x}_i$ are the combination of in-batch examples and all examples from the memory bank. Consequently, $P_i = \{\boldsymbol{r}_j | \boldsymbol{r}_j \in B \cup Q, q_j = q_i\}$, and $C_i = \{\boldsymbol{r}_j | \boldsymbol{r}_j \in B \cup Q\}$.

3.4. Essay Scorer and Training Loss

We use a fully connected layer as the essay scorer to transform the essay representation $\boldsymbol{r}_i$ to the predicted essay score $\hat{y}_i$ (Equation 4), where $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable parameters.

$$\hat{y}_i = \boldsymbol{W}\cdot\boldsymbol{r}_i + \boldsymbol{b} \quad (4)$$

We use mean squared error L_{MSE} as the scoring loss (Equation 5). The overall loss is presented in Equation 6, where λ is a hyper-parameter that balances the contribution of the two losses.

$$L_{MSE} = (\hat{y}_i - y_i)^2 \quad (5)$$

$$L = \lambda\cdot L_{MSE} + (1-\lambda)\cdot L_{XCL} \quad (6)$$

4. Experimental Settings

4.1. Datasets and Evaluation Metric

We include four datasets that contain essays in six languages for cross-lingual scoring experiments[2], including MERLIN (Boyd et al., 2014) for Czech, German and Italian, Write & Improve (Yannakoudakis et al., 2018) for English, CEDEL2 (Lozano, 2022) for Spanish, and COPLE2 (Mendes et al., 2016) for Portuguese. Dataset statistics are presented in Table 1. Essays in all included datasets are produced by second-language learners. Essays in MERLIN, Write & Improve, and COPLE2 are manually assigned with CEFR (Common European Framework of Reference) levels: A1 (lowest), A2, B1, B2, C1 and C2 (highest). Essays in CEDEL2 are manually assigned with University of Wisconsin placement test scores in the range of 0 to 43. For consistency, we map them into the CEFR levels according to the guidelines (Lozano, 2022).

We discard unrated essays in the MERLIN and COPLE2 datasets, as they are unlabelled and cannot be used for training. We also discard spoken essays in the CEDEL2 and COPLE2 datasets, as we focus on scoring written essays. We convert the CEFR levels into 1-6, and use a train/dev split of 3/1 for model training.

We use quadratic weighted kappa (QWK) for evaluating cross-lingual scoring performance, which has been widely adopted as a general evaluation metric in the AES literature of both English and other languages (Taghipour and Ng, 2016; Dong et al., 2017; Hirao et al., 2020; Yupei and Renfen, 2021; Ridley et al., 2021; Chen and Li, 2023) for measuring scoring agreement between the human rater and the scoring model. Its value typically ranges from 0 to 1 and a higher value indicates a higher human-model scoring agreement.

4.2. Baseline Models

We include existing statistical cross-lingual scoring methods as baselines for comparison (Vajjala and

[2]While there are AES studies for other languages, such as Chinese and Japanese, the datasets used are not publicly available. Therefore, we do not choose them for experiments.

Dataset	Language	Code	N_{raw}	$N_{filtered}$	Scores					
					1	2	3	4	5	6
MERLIN	Czech	cs	441	439	1	188	165	81	4	0
MERLIN	German	de	1,033	1,033	57	306	331	293	42	4
MERLIN	Italian	it	813	806	29	381	394	2	0	0
Write & Improve	English	en	3,300	3,300	585	845	631	469	483	287
CEDEL2	Spanish	es	3,034	3,007	92	526	543	672	756	418
COPLE2	Portuguese	pt	1,658	1,566	301	378	377	314	196	0

Table 1: Dataset statistics. Code represents language code. N_{raw} denotes the essay number before filtering (removing unrated and spoken essays), and $N_{filtered}$ denotes the number of essays after filtering. The last six columns present the number of essays for each score.

Multilingual PLM	N_{langs}	Case Sensitivity
mBERT-cased	104	Sensitive
mBERT-uncased	102	Insensitive
XLM-R	100	Insensitive

Table 2: Summary of multilingual PLMs. N_{langs} denotes the number of languages supported (comprehended) by each PLM.

Rama, 2018; Horbach et al., 2018). Additionally, we extend the CNN-LSTM hierarchical attention network (Dong et al., 2017) to cross-lingual scoring for implementing neural baselines.

Universal Features. Vajjala and Rama (2018) use word count, POS n-grams, dependency n-grams and domain features (e.g., document length and lexical richness features), which they call universal features, to train models with linear regression, linear support vector regression and random forest regression.

Essay Translation. Horbach et al. (2018) either translate the source essays to the target language or translate the target essays to the source language, and extract token-level and character-level n-grams for model training. Since there are multiple source languages for target-to-source translation in our setting, we choose the one with the most essays for translation.

Hierarchical Encoding. For neural baselines, we adopt the CNN-LSTM hierarchical attention network (Dong et al., 2017) as the essay encoder, which has been proved effective by previous English AES studies (e.g., Dasgupta et al. 2018; Ridley et al. 2020, 2021; Chen and Li 2023). The original model uses a monolingual vocabulary for text tokenization, which is inappropriate for multilingual essay representation. Hence, we extend this model to cross-lingual scoring by respectively constructing the vocabulary with POS tags, relation tags and dependency tags.

4.3. Implementation Details

Implementation of Our Method. We use three multilingual PLMs for multilingual essay representation (Table 2), including mBERT-cased [3] (Devlin et al., 2019), mBERT-uncased [4] (Devlin et al., 2019) and XLM-R [5] (Conneau et al., 2020). We use AdamW (Loshchilov and Hutter, 2017) to update model parameters with a learning rate of 5e-5 for 50 epochs. The maximum sequence length is 512. The batch size is 16. Early stopping is applied and the training is stopped when the development QWK is not improved for 5 consecutive epochs. For contrastive learning, the temperature τ is 0.1, the bank size S is 128, and the balancing factor λ is 0.9. According to the aforementioned writing quality derivation in Section 3.3 and the essay score range of the used datasets, the writing quality q_i of each score y_i is determined as follows:

- $y_i \in \{1, 2\}$: q_i = low
- $y_i \in \{3, 4\}$: q_i = medium
- $y_i \in \{5, 6\}$: q_i = high

For our models and all baselines, we average the results of three runs with different random seeds to reduce randomness.

Implementation of the Baselines. For Universal Features (Vajjala and Rama, 2018), we use their published code to extract hand-crafted features. Similar to them, we cannot find a proper Czech spell checker, so we discard spelling number from the domain features for all languages to ensure a fair comparison. For Essay Translation (Horbach et al., 2018), we use Google Translate [6] to translate multilingual essays into a common language. For Hierarchical Encoding (Dong et al., 2017), we use UDPipe (Straka and Straková, 2017) to extract POS, relation, and dependency

[3] https://huggingface.co/bert-base-multilingual-cased

[4] https://huggingface.co/bert-base-multilingual-uncased

[5] https://huggingface.co/xlm-roberta-base

[6] https://translate.google.com

Model	cs	de	en	es	it	pt	Avg
Universal Features (Vajjala and Rama, 2018)							
LR + word count	0.381	0.520	0.017	0.129	0.235	0.275	0.259
LR + domain	0.334	0.439	0.132	0.152	0.297	0.299	0.275
LR + POS n-grams	0.149	0.289	0.170	0.305	0.264	0.350	0.254
LR + dependency n-grams	0.472	0.300	0.167	0.395	0.268	0.299	0.317
SVR + word count	0.416	0.527	0.091	0.106	0.263	0.137	0.257
SVR + domain	0.380	0.515	0.151	0.143	0.303	0.365	0.310
SVR + POS n-grams	0.157	0.302	0.187	0.307	0.271	0.356	0.263
SVR + dependency n-grams	0.482	0.317	0.203	0.393	0.276	0.306	0.329
RFR + word count	0.191	0.064	0.282	0.227	0.117	0.091	0.162
RFR + domain	0.318	0.250	0.033	0.235	0.256	0.141	0.206
RFR + POS n-grams	0.668	0.575	0.344	0.332	0.599	0.377	0.482
RFR + dependency n-grams	0.663	0.518	0.389	0.347	0.582	0.421	0.487
Essay Translation (Horbach et al., 2018)							
LR + token&char n-grams ($sl \rightarrow tl$)	0.165	0.089	0.000	0.014	0.055	0.209	0.089
LR + token&char n-grams ($tl \rightarrow sl$)	0.502	0.567	0.329	0.421	0.390	0.451	0.443
SVR + token&char n-grams ($sl \rightarrow tl$)	0.169	0.102	0.000	0.012	0.055	0.214	0.092
SVR + token&char n-grams ($tl \rightarrow sl$)	0.514	0.598	0.344	0.419	0.432	0.456	0.461
RFR + token&char n-grams ($sl \rightarrow tl$)	0.666	0.630	0.421	0.414	0.610	**0.547**	0.548
RFR + token&char n-grams ($tl \rightarrow sl$)	**0.670**	0.653	0.387	0.386	0.579	0.535	0.535
Hierarchical Encoding (Dong et al., 2017)							
HAN + POS	0.618	0.561	0.391	0.369	0.500	0.415	0.476
HAN + relation	0.579	0.552	0.401	0.357	0.483	0.410	0.464
HAN + dependency	0.642	0.497	0.407	0.359	0.498	0.483	0.481
Our Proposed Method							
mBERT-cased	0.418	0.746	0.345	0.361	**0.613**	0.383	0.478
XLM-R	0.635	0.660	0.390	0.527	0.471	0.506	0.531
mBERT-uncased	0.566	0.726	0.419	0.544	0.574	0.453	0.547
mBERT-uncased + XCL	0.558	**0.789**	**0.438**	**0.565**	0.539	0.520	**0.568**

Table 3: QWKs of zero-shot cross-lingual scoring. LR: linear regression; SVR: linear support vector regression; RFR: random forest regression; HAN: CNN-LSTM hierarchical attention network (Dong et al., 2017). $sl \rightarrow tl$: translating source essays to the target language; $tl \rightarrow sl$: translating target essays to a source language. The best results are in bold and underlined.

(i.e., triplets of POS, relation and head node POS) tags for constructing model vocabularies. The batch size is set to 16. Other training settings are maintained consistent with Dong et al. (2017).

5. Results and Analysis

5.1. Main Results

Table 3 presents the zero-shot cross-lingual scoring results, where essays for one language (e.g., Czech) are set aside for testing, and essays of the remaining languages (e.g., German, English, Spanish, Italian, and Portuguese) are used for model training.

Overall Analysis. The results show that our proposed method outperforms all baseline models. Specifically, our proposed "mBERT-uncased + XCL" produces an average QWK of 0.568, which is higher than the best-performing statistical baseline "RFR + token&char n-grams ($sl \rightarrow tl$)" (0.548) by 2%, and is higher than the best-performing neural baseline "HAN + dependency" (0.481) by a large margin of 8.7%. Our model also produces the highest QWK on German (0.789), English (0.438) and Spanish (0.565) zero-shot cross-lingual scoring. This observation demonstrates the effectiveness of our approach on cross-lingual AES.

It is worth noting that although the "RFR + token&char n-grams ($sl \rightarrow tl$)" method from Horbach et al. (2018) produces the second-best average QWK (0.548), it requires translating source essays into the target language for model development, leading to time-consuming and inefficient model re-training for newly introduced languages. Besides, their approach may lead to unrealistic and unreliable scoring results due to the inability of machine translation to faithfully represent original essay writing quality. On the contrary, our method does not require model re-training for new target languages and does not alter essay content, but still outperforms their method by 2%.

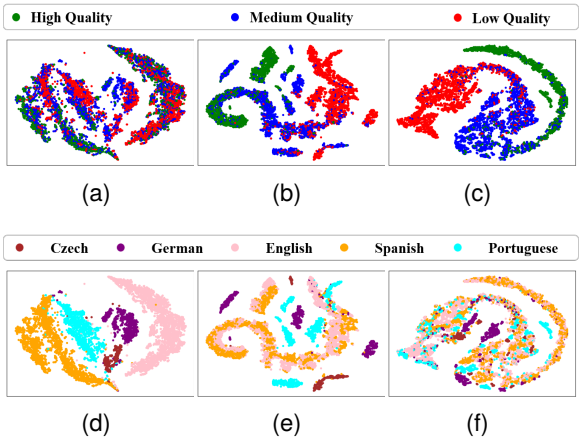

Figure 4: Visualization of essay representations in terms of writing quality (upper) and language (lower) with Italian being the target language. (a) and (d): without fine-tuning; (b) and (e): fine-tuning without the proposed XCL; (c) and (f): fine-tuning with the proposed XCL. Red, blue and green colours represent low, medium, and high-quality essays. Brown, purple, pink, orange and cyan represent essays in Czech, German, English, Spanish and Portuguese.

Effectiveness of Writing Quality Alignment. After incorporating the proposed XCL technique into the best-performing multilingual PLM mBERT-uncased, the performance improves from 0.547 to 0.568 with a performance gain of 2.1%, indicating the effectiveness of our method. Notably, the QWK on Portuguese improves the most from 0.453 to 0.520 by a large margin of 6.7%.

Multilingual PLM Performance Comparison. The results also show that different multilingual PLMs lead to varied cross-lingual scoring performance. Specifically, mBERT-cased produces the lowest average QWK (0.478), while mBERT-uncased produces the highest (0.547).

The performance discrepancy can be explained from two perspectives. First, lower-casing essays may be helpful for cross-lingual scoring, as the case-insensitive mBERT-uncased (0.547) performs better than the case-sensitive mBERT-cased (0.478). Second, pretraining corpora, subword tokenization methods and pretraining objectives may affect cross-lingual scoring performance, as mBERT-uncased and XLM-R are different in these aspects. Hence, these two perspectives should be taken into consideration when choosing multilingual PLMs for multilingual essay representation in cross-lingual AES.

5.2. Visualization Analysis

To demonstrate the effectiveness of writing quality alignment by contrastive learning, we use the t-SNE algorithm (van der Maaten and Hinton, 2008) to visualize essay representations in terms of writing quality and language, as shown in Figure 4.

Without fine-tuning (Figure 4(a), 4(d)), the essays are gathered by language rather than writing quality, showing that the vanilla mBERT-uncased PLM can distinguish different languages, but is not able to distinguish writing quality. When fine-tuned without XCL (Figure 4(b), 4(e)), the essays tend to be gathered according to writing quality, but there are multiple clusters for the same writing quality. This suggests that naive fine-tuning enables mBERT to recognize writing quality to a certain degree, but the recognition ability is not sufficient. When fine-tuned with XCL (Figure 4(c), 4(f)), the essays are clearly clustered into three groups according to writing quality rather than according to language, suggesting that the proposed XCL method indeed enables mBERT to align writing quality across languages.

N_{level}	cs	de	en	es	it	pt	Avg
2	0.547	0.780	0.410	**0.567**	**0.577**	0.490	0.562
3	**0.558**	**0.789**	**0.438**	0.565	0.539	**0.520**	**0.568**
6	0.553	0.755	0.382	0.540	0.550	0.477	0.543

Table 4: Cross-lingual scoring performance with different degrees of writing quality granularity (i.e., the number of writing quality levels, as denoted by N_{level}). The best results are bolded.

5.3. Impact of Writing Quality Granularity

In this section, we investigate the impact of writing quality granularity on cross-lingual scoring performance. Specifically, we train scoring models with three degrees of writing quality granularity for cross-lingual contrastive learning, including a 2-level scale (low quality for the lowest 50% scores in the score range and high quality for the highest 50% scores), a three-level scale (low quality for the lowest 20% scores, high quality for the highest 20% scores, and medium quality for the remaining 60%), and a 6-level scale (consistent with the 1-6 scores of the essays in each dataset).

The results show that the 2-level scale and the 3-level scale produce similar average QWKs (0.562 and 0.568). However, the 6-level scale produces a much lower average QWK of 0.543. This shows that the more coarse-grained scales of 2-3 levels result in better cross-lingual scoring performance than the more fine-grained 6-level scale. The possible reason is that a too small granularity may result in smaller differences in adjacent levels, which increases the difficulty for the model to distinguish different degrees of writing quality in contrastive learning, leading to worse performance.

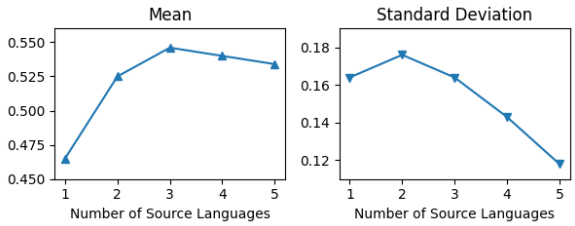

Figure 5: Mean and standard deviation of QWKs in different source language diversity (indicated by the number of source languages).

Source		Target	QWK
Spanish (R)	Italian (R)		0.468
German (G)	Italian (R)	Portuguese (R)	0.363
German (G)	English (G)		**0.529**
Portuguese (R)	Spanish (R)		0.589
Portuguese (R)	English (G)	Italian (R)	**0.594**
German (G)	English (G)		0.562
Italian (R)	Portuguese (R)		0.307
Italian (R)	English (G)	Spanish (R)	0.413
German (G)	English (G)		**0.582**

Table 5: Cross-lingual scoring performance on three target languages: Portuguese, Italian and Spanish. For each target language, its source languages have varying degrees of similarity to it. R and G denote Romance and Germanic languages.

5.4. Impact of Source Language Diversity

To investigate the impact of source language diversity on cross-lingual scoring performance, we conduct a source language increment experiment that compares the performance of models trained with different numbers of source languages [7].

Specifically, taking each of the six languages as the target language, we gradually increase its source languages from 1 to 5, and calculate the evaluated QWK for each number of source languages. Subsequently, we calculate the mean and standard deviation of the QWK scores for the six languages at each number of source languages, as shown in Figure 5. With more source languages, the mean increases and the standard deviation decreases, indicating that source language diversity improves overall cross-lingual scoring performance and performance consistency across languages. Furthermore, increasing the source languages from 1 to 2 leads to a huge performance boost, indicating that more diverse source languages enable our model to effectively learn cross-lingual scoring knowledge that cannot be acquired from monolingual source data.

5.5. Impact of Language Similarity

There are differences in similarity between languages. As shown in the following example, Portuguese (a Romance language) is very similar to Spanish (another Romance language) in terms of vocabulary and sentence structure, but is very different from German (a Germanic language).

- Portuguese: Este aluno está aprendendo?
- Spanish: ¿Este alumno está aprendiendo?
- German: Lernt dieser Schüler?

In this section, we investigate the impact of language similarity on cross-lingual scoring (Table 5).

[7] In this section and Section 5.5, we ensure a consistent number of essays in each source language to allow for fair comparisons.

The results show that higher similarity between source and target languages does not necessarily yield better cross-lingual scoring performance. For instance, when Spanish and Italian (two Romance languages) are the source languages, and Portuguese (another Romance language) is the target language, the QWK is 0.468. However, when German and English (two Germanic languages that are dissimilar to Portuguese) are the source languages, the performance increases to 0.529.

The limited impact of language similarity could be explained by the multi-perspective characteristics of AES for evaluating writing quality. Specifically, higher source-target language similarity principally lies in the similarity in vocabulary, sentence structure and other linguistic components that benefit the model to assess quality from the perspective of language. However, essay scoring involves multiple perspectives beyond language, such as content, coherence and composition structure. Therefore, a higher language similarity may not help the scoring of these perspectives, leading to the limited impact of language similarity on cross-lingual scoring.

6. Conclusion

In this paper, we propose a novel method for zero-shot cross-lingual automated essay scoring. It enables the model to effectively represent multilingual essays, and to learn cross-lingual scoring knowledge by aligning writing quality. We conduct comprehensive experiments on six languages with diverse baselines, and demonstrate the effectiveness of our method.

Apart from writing quality, finer-grained scoring perspectives such as structure completeness could be aligned for better cross-lingual scoring performance, which we leave for future work. Besides, we only used essays in Indo-European languages due to data constraints. We plan to include more diverse language families in the future.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China [grant number: 61976062].

8. Bibliographical References

Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, New Orleans, Louisiana. Association for Computational Linguistics.

Evelin Amorim and Adriano Veloso. 2017. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Cristina Arhiliuc, Jelena Mitrović, and Michael Granitzer. 2020. Language proficiency scoring. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5624–5630, Marseille, France. European Language Resources Association.

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v.2.0. *ETS Research Report Series*, 2004(2):i–21.

Avi Caciularu, Ido Dagan, Jacob Goldberger, and Arman Cohan. 2022. Long context question answering via supervised contrastive learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2872–2879, Seattle, United States. Association for Computational Linguistics.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1011–1020, New York, NY, USA. Association for Computing Machinery.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–799, Berlin, Germany. Association for Computational Linguistics.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, Melbourne, Australia. Association for Computational Linguistics.

Iria del Río. 2019. Linguistic features and proficiency classification in L2 Spanish and L2Portuguese. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 31–40, Turku, Finland. LiU Electronic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Peter W. Foltz, Darrell Laham, and Thomas K Landauer. 1999. Automated essay scoring: Applications to educational technology. In *Proceedings of EdMedia + Innovate Learning 1999*, pages 939–944, Seattle, WA USA. Association for the Advancement of Computing in Education (AACE).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2022. Automated Chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3007–3016, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, and Mamoru Komachi. 2020. Automated essay scoring system for nonnative Japanese learners. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1250–1257, Marseille, France. European Language Resources Association.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2017. Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366, Copenhagen, Denmark. Association for Computational Linguistics.

Andrea Horbach, Sebastian Stennmanns, and Torsten Zesch. 2018. Cross-lingual content scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 410–419, New Orleans, Louisiana. Association for Computational Linguistics.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, Toronto, Canada. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. 2019. Regression or classification? automated essay scoring for Norwegian. In *Pro-*

ceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 92–102, Florence, Italy. Association for Computational Linguistics.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Syed Latifi and Mark Gierl. 2020. Automated scoring of junior and senior high essays using coh-metrix features: Implications for large-scale language testing. *Language Testing*, 38(1):62–85.

Xia Li, Minping Chen, and Jian-Yun Nie. 2020. Sednn: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems*, 210:106491.

Xia Li, Minping Chen, Jianyun Nie, Zhenxing Liu, Ziheng Feng, and Yingdan Cai. 2018. Coherence-based automated essay scoring using self-attention. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 386–397, Cham. Springer International Publishing.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Danielle S. McNamara, Scott A. Crossley, Rod D. Roscoe, Laura K. Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

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 ACL (EACL 2009)*, pages 567–575, Athens, Greece. Association for Computational Linguistics.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Ellis Batten Page. 1994. Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education*, 62(2):127–142.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *CoRR*, abs/2008.01441.

Lawrence Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *Journal of Technology, Learning, and Assessment*, 1.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1049–1054, Denver, Colorado. Association for Computational Linguistics.

Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76.

Jinnie Shin and Mark J. Gierl. 2020. More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, 38(2):247–272.

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. 2020. Multi-stage pre-training for automated Chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6723–6733, Online. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Domain generalization for text classification with memory-based supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6916–6926, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sowmya Vajjala and Kaidi Lõo. 2013. Role of morpho-syntactic features in Estonian proficiency classification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72, Atlanta, Georgia. Association for Computational Linguistics.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova, and Mamoru Komachi. 2020. Cross-lingual transfer learning for grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4704–4715, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wang Yupei and Hu Renfen. 2021. A prompt-independent and interpretable automated essay scoring method for Chinese second language writing. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1202–1217, Huhhot, China. Chinese Information Processing Society of China.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado. Association for Computational Linguistics.

9. Language Resource References

Boyd, Adriane and Hana, Jirka and Nicolas, Lionel and Meurers, Detmar and Wisniewski, Katrin and Abel, Andrea and Schöne, Karin and Štindlová, Barbora and Vettori, Chiara. 2014. *The MERLIN corpus: Learner language and the CEFR*. European Language Resources Association (ELRA). PID https://www.merlin-platform.eu.

Lozano, Cristóbal. 2022. *CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research*. Sage Publications Sage UK: London, England. PID http://cedel2.learnercorpora.com.

Mendes, Amália and Antunes, Sandra and Janssen, Maarten and Gonçalves, Anabela. 2016. *The COPLE2 corpus: a learner corpus for Portuguese*. European Language Resources Association (ELRA), ISLRN 642-718-655-040-0.

Yannakoudakis, Helen and Andersen, Øistein E and Geranpayeh, Ardeshir and Briscoe, Ted and Nicholls, Diane. 2018. *Developing an automated writing placement system for ESL learners*. Taylor & Francis. PID https://www.cl.cam.ac.uk/research/nl/bea2019st.