IndoCL: Benchmarking Indonesian Language Development Assessment

Nankai Lin[1], Hongyan Wu[2], Weixiong Zheng[1],
Xingming Liao[1], Shengyi Jiang[3], Aimin Yang[1,4] and **Lixian Xiao[5,✉]**
[1] School of Computer Science and Technology, Guangdong University of Technology
[2] College of Computer, National University of Defense Technology
[3] School of Information Technology and Engineering, Guangzhou College of Commerce
[4] School of Computer Science and Intelligence Education, Lingnan Normal University
[5] Faculty of Asian Languages and Cultures, Guangdong University of Foreign Studies

Abstract

Recently, the field of language acquisition (LA) has significantly benefited from natural language processing technologies. A crucial task in LA involves tracking the evolution of language learners' competence, namely language development assessment (LDA). However, the majority of LDA research focuses on high-resource languages, with limited attention directed toward low-resource languages. Moreover, existing methodologies primarily depend on linguistic rules and language characteristics, with a limited exploration of exploiting pre-trained language models (PLMs) for LDA. In this paper, we construct the IndoCL corpus (**Indo**nesian **C**orpus of L2 **L**earners), which comprises compositions written by undergraduate students majoring in Indonesian language. Moreover, we propose a model for LDA tasks, which automatically extracts language-independent features, relieving laborious computation and reliance on specific language. The proposed model uses sequential information attention and similarity representation learning to capture the differences and common information from the first-written and second-written essays, respectively. It has demonstrated remarkable performance on both our self-constructed corpus and publicly available corpora. Our work could serve as a novel benchmark for Indonesian LDA tasks. We also explore the feasibility of using existing large-scale language models (LLMs) for LDA tasks. The results show significant potential for improving LLM performance in LDA tasks.[1]

1 Introduction

The advancement of artificial intelligence has significantly enhanced the evolution of educational technology and the utilization of computer-assisted learning (Huang and Wei, 2022; Phan et al., 2023). Recent interest has surged in applying natural language processing (NLP) and machine learning (ML) to evaluate language development in both first (L1) and second (L2) language acquisition. The goal is to analyze learners' linguistic attributes and the progression of their language skills across various modalities and stages (Crossley, 2020).

Language Development Assessment (LDA) (Sagae, 2021; Wu et al., 2023) is a critical task in the field of language acquisition that focuses on evaluating the progress of language learners over time. As shown in Figure 1, a student wrote two paragraphs at two different time periods, denoted as Text A and Text B. If the model determines that text B was written after text A, it is considered that text B is better written than text A, indicating an improvement in the student's writing skills. Conversely, if the model believes that the writing order is that the student wrote text B before text A, then text B is considered inferior to text A, indicating a decline in the student's writing skills.

For LDA, the majority of research has been focused on high-resource language. Weiss and Meurers (2019) and Kerz et al. (2020) conducted investigations into the development of writing skills among German-speaking students from elementary to secondary school levels. Miaschi et al. (2020, 2021a) proposed a method to track the evolution of written language competence in L2 Spanish and L1 Italian learners. To the best of our knowledge, few existing LDA studies focus on low-resource languages, such as Indonesian, and there is a lack of available LDA corpora for these languages.

Early research in LDA has explored enhancing model performance through text feature extraction (Barbini et al., 2023; Stemle et al., 2023). Due to the complexity and tediousness of the feature extraction process, some research (Miaschi et al., 2021b) adopts the fully data-driven approach for LDA tasks. This type of approach tracks language development by automatically extracting linguis-

[1]Our code and corpus can be obtained from https://github.com/GKLMIP/IndoCL.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4873–4885

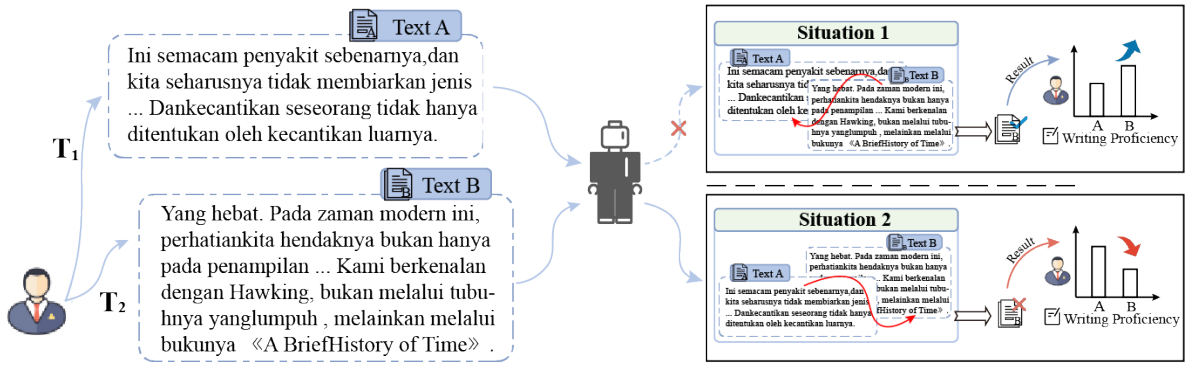

Figure 1: Illustration of the language development assessment task that focuses on evaluating the progress of language learners over time. The red curve indicates the order of text predicted by the model.

tic features. Recent advances in PLMs demonstrate strong capabilities in generating fluent text and providing useful representations of textual patterns. Nevertheless, a limited amount of research has concentrated on employing PLMs for LDA tasks. Wu et al. (2023) introduced a novel sequential information attention mechanism to analyze the interaction between essay pairs. ExtremITA (Alzetta et al., 2023) employed two PLMs trained for undergoing joint fine-tuning using prompting techniques. However, an important issue to be addressed in this regard is how to effectively mine language-independent text features, and extract the differences and common information between the first-written and second-written essays.

To address these gaps, we construct the IndoCL corpus (Indonesian Corpus of L2 Learners), which is a longitudinal corpus of essays written by L2 Indonesian students in the second and third year of the undergraduate period. Furthermore, we propose a novel model for LDA tasks, which effectively extracts the differences and common information between first-written essay and second-written essay through **sequential information attention** and **similarity representation learning** respectively, gaining substantial improvement over existing models on our self-constructed corpus. Simultaneously, our proposed model exhibits remarkable merit in alleviating the laborious computation and preventing reliance on specific language. Experimental results on publicly available corpora reveal that our model achieves promising performance. The main contributions of the paper are:

(1) We construct the IndoCL corpus, a dataset for Indonesian LDA tasks, which addresses the lack of resources for Indonesian LDA tasks.

(2) We present an innovative model designed for LDA tasks, which simultaneously captures the difference and common information between essay pairs and achieves prominent performance.

(3) Our method automatically extracts language-independent features for LDA tasks, demonstrating the effectiveness of reducing laborious computation and proving applicable across languages.

2 Related Work

Language acquisition (LA) is the field of study concerned with understanding how individuals acquire first or second language competencies. In recent years, with the rapid development of natural language processing (NLP) technologies, the field of LA has significantly benefited (Chaudhary et al., 2023; Evanson et al., 2023; Oba et al., 2023; Yadavalli et al., 2023). Previous studies on LDA tasks can be broadly divided into two distinct groups: one group focuses on constructing a LDA model that is based on linguistic features while the other focuses on constructing a LDA model that is based on neural networks. To address the difficulty of identifying unique linguistic features, **linguistic-features-based methods** extract a broad range of linguistic-level features for training supervised classification systems on authentic learner data from multiple languages (Hancke and Meurers, 2013; Vajjala and Lõo, 2014; Pilán and Volodina, 2018; Miaschi et al., 2020, 2021b; Bulté and Housen, 2014; Cui and Sachan, 2023). Recent studies on neural network-based language modeling (e Silva et al., 2023; Arehalli and Linzen, 2024; McCoy et al., 2020) have shown that certain neural architectures can understand syntactic details from text without being directly programmed to do so. Some studies employ **neural-networks-based methods** to track and model students' writing skills (Sagae, 2021;

Wu et al., 2023; Barbini et al., 2023). The more detailed related work is presented in Appendix A.

Year	2021	2022	2023	Total
Training set	7	7	7	21
Validation set	2	3	2	7
Test set	2	3	2	7
Total	11	13	11	35

Table 1: Students distribution of each group.

3 Indonesian Corpus of L2 Learners

To ensure that our trained model evaluates language proficiency based on the content of the text, without reliance on domain-specific information, we have constructed a general domain dataset for Indonesian. The corpus has a diverse range of essay types, including narrative, descriptive, argumentative, and expository essays, which encompass a diverse range of essay topics, including but not limited to food, environment, and lifestyle. These topics are selected to cover a wide range of everyday and relatable topics that are likely to elicit meaningful and context-rich responses from the students.

Data Collection and Grouping. We track the L2 language writing of students majoring in Indonesian at a university and collect the essays of 35 students in three different Indonesian classes, spanning the years 2021, 2022, and 2023, respectively. The students' first language is Chinese, and they are all undergraduate students majoring in Indonesian language. Professional Indonesian teachers provide writing skills training to students, and the essays are written as part of writing ability tests administered at intervals of 1 to 2 months. 24 students participated in 2021 and 2022, each writing 7 essays. In 2023, 11 students contributed to the dataset, each producing 4 essays. The entire corpus contains a total of 212 essays. In order to protect the privacy of essay writers, we have masked personal information such as name and contact information that appear in the essay. Owing to the different writing styles among the students, we firstly group the students to avoid information leaks in terms of writing style information in the model, that is, the same student's essays will not appear in different data sets. Subsequently, we divide the 35 students into training, validation, and test sets in the ratio of 6:2:2, which is shown in 1.

Data Preparation. For each student's essay, we arrange them in order of writing time. Subsequently, for the essay with the length being too long, it is divided into sub-paragraphs up to 150 tokens in length. After segmentation, 212 essays are divided into 733 sub-paragraphs. Thus, we construct a sub-paragraph set $PC = \{(c_1, t_1), (c_2, t_2), \ldots, (c_n, t_n)\}$ for each student, where n denotes the number of paragraphs written by the student. For the paragraph i, c_i is the text content and t_i represents the writing time of the paragraph. It is worth noting that sub-paragraphs from the same sample are considered texts from the same period, not first-written and second-written texts from different periods.

Sample Construction. Suppose there are two essays c_i and c_j from the same author, with essay c_i written before essay c_j. In the language learning development task, when constructing a positive sample (c_i, c_j), the label corresponding to the sample is 1, and when constructing a negative sample (c_j, c_i), the label is 0. Since the order of the two essays determines the label of the sample, we can generate different samples by changing the order of the essays in the sample.

For students in the training set, we traverse the samples in the paragraph set to construct text pairs (c_i, c_j). Specifically, provided that the paragraph c_i is written before the paragraph c_j, the text pair is retained and is labeled "1". All the retained text pairs form the positive sample set $P = (q_1, k_1, 1), ..., (q_m, k_m, 1)$, where m is the number of samples contained in the positive sample set, and q_i, k_i are the first-written paragraph text and second-written paragraph text, respectively. However, for the model training, solely positive samples are unreasonable. Therefore, we further construct negative samples for training. For each sample $(q_i, k_i, 1)$ in the positive sample set, the corresponding negative sample $(k_i, q_i, 0)$ can be obtained after being inverted. We enumerate all samples in the positive sample set to construct a corresponding negative sample set $N = \{(k_1, q_1, 0), ..., (k_m, q_m, 0)\}$, where m denotes the number of samples contained in the negative sample set. Ultimately, a positive sample set and a negative sample set are merged to construct a training set $T = \{(q_1, k_1, 1), (k_1, q_1, 0), ..., (q_m, k_m, 1), (k_m, q_m, 0)\}$, where the number of samples contained in the training set is $2m$. Intuitively, the training set is

balanced, with an equal number of positive samples and negative samples.

Datasets	Positive	Negatives	Total
Training set	1573	1573	3146
Validation set	279	279	558
Test set	310	310	620

Table 2: Datasets distribution of Indonesian. The numbers in the table are the number of samples included in each dataset. Each sample consists of two essays.

For the students in the validation set and the test set, we adopt the same enumeration method to generate positive samples and negative samples. For the training set, it is expected that as many samples as possible are used for training. Whereas for the validation set and test set, we aim to ensure that no highly similar samples are in the same set. In the case of sample pairs (q_i, k_i) and (k_i, q_i), the primary distinction lies in the sequence of the two essays, with other textual characteristics remaining notably similar. Including such closely aligned pairs in both the test and validation sets could inadvertently inflate the perceived efficacy of a model's language assessment capabilities. Specifically, if the model possesses limited proficiency in language evaluation, it might achieve a boost in accuracy by merely assigning identical labels to these similar pairs. This phenomenon risks leading to an overestimation of the model's true evaluative capacity. Our primary objective in ensuring that the validation and test sets do not contain highly similar samples is to better evaluate the model's generalization capabilities. Hence, for the positive sample $(q_i, k_i, 1)$ and negative sample $(k_i, q_i, 0)$, we only randomly retain one of the samples, constructing the validation set $V = \{(q_1, k_1, 1), (k_2, q_2, 0), ..., (q_{l-1}, k_{l-1}, 1), (k_l, q_l, 0)\}$ and test set $T = \{(q_1, k_1, 1), (k_2, q_2, 0), ..., (q_{g-1}, k_{g-1}, 1), (k_g, q_g, 0)\}$ respectively, where l and g denote the number of samples contained in the validation set and the test set, respectively. After processing, the data distribution of our constructed datasets is shown in Table 2.

Example. We provide the example from the IndoCL dataset in Table 3. Indonesian experts annotated the good and bad expressions in the text. The sample labeled "0" has more bad expressions in Text B than in Text A, justifying the label.

However, it is crucial to efficiently extract language-independent text features and identify the differences and shared information between the first-written and second-written essays.

4 Model

Our proposed framework is shown in Figure 2. Initially, in the **text representation module**, we concatenate the first-written essay and second-written essay in a text pair to form a text sequence and encode it employing the PLM. Subsequently, the **sequential information attention mechanism** is leveraged to capture the information interaction in the text sequence, so as to obtain the enhanced global representation of the first-written essay and second-written essay respectively. In the **similarity representation learning module**, the information similarity between the enhanced first-written text representation and the second-written text representation is calculated via exploiting a learnable similarity vector. Finally, the **language development assessment** module combines the enhanced global first-written text representation, the enhanced global second-written text representation, the raw text representation, and the vector-based similarity representation to conduct a language development assessment. It is worth mentioning that our framework focuses on capturing both the differences and common information between two essays for language development assessment without the need to extract specific linguistic features, demonstrating extensive applicability. Each module is comprehensively presented in detail below.

4.1 Text Representation

We utilize the pre-trained model BERT for text representation. Given a first-written text $T_a = \{w_a^1, w_a^2, w_a^3, ..., w_a^n\}$ and a second-written text $T_b = \{w_b^1, w_b^2, w_b^3, ..., w_b^m\}$, two specific tokens "[CLS]" and "[SEP]" of BERT are employed to concatenate two texts, forming the input sequence $T = \{[CLS], w_a^1, w_a^2, w_a^3, ..., w_a^n, [SEP], w_b^1, w_b^2, w_b^3, ..., w_b^m, [SEP]\}$, where "[CLS]" indicates the beginning of a sentence and "[SEP]" indicates the end of a sentence. Moreover, n and m denote the length of the two texts respectively. We feed the text sequence to the PLM BERT (Devlin et al., 2019) to generate the semantic representation as:

$$H = Encoder(T), \tag{1}$$

where $H = \{h^{[CLS]}, h_a^1, h_a^2, h_a^3, ..., h_a^n, h^{[SEP]}, h_b^1, h_b^2, h_b^3, ..., h_b^m, h^{[SEP]}\} \in R^{(n+m+3)\cdot z}$, and z represents the dimension of semantic representation.

Text A	Text B	Label
Ini semacam penyakit sebenarnya, dan kita seharusnya tidak membiarkan jenis diskriminasi ini berkembang, jadi apa yang bisa kita lakukan adalah tidak menilai orang lain berdasarkan penampilan mereka. Apa yang seseorang terlihat seperti, tidak pernah ditentukan oleh dirinya sendiri. Dan kecantikan seseorang tidak hanya ditentukan oleh kecantikan luarnya. ···	Yang hebat. Pada zaman modern ini, perhatian kita hendaknya bukan hanya pada penampilan fisik, melainkan juga pada "memperkuat hati". Wajah yang cantik akan memuaskan mata, tapi memperkuat hati akan membuat Anda berhasil. Kedua, kita hendaknya belajar untuk mengenali kecantikan batin orang lain. Kami berkenalan dengan Hawking, bukan melalui tubuhnya yang lumpuh, melainkan melalui bukunya 《A Brief History of Time》. ···	0

Table 3: Examples of the IndoCL dataset. The text with a red background is deemed great expressions by Indonesian experts, while the text with a green background is deemed bad.

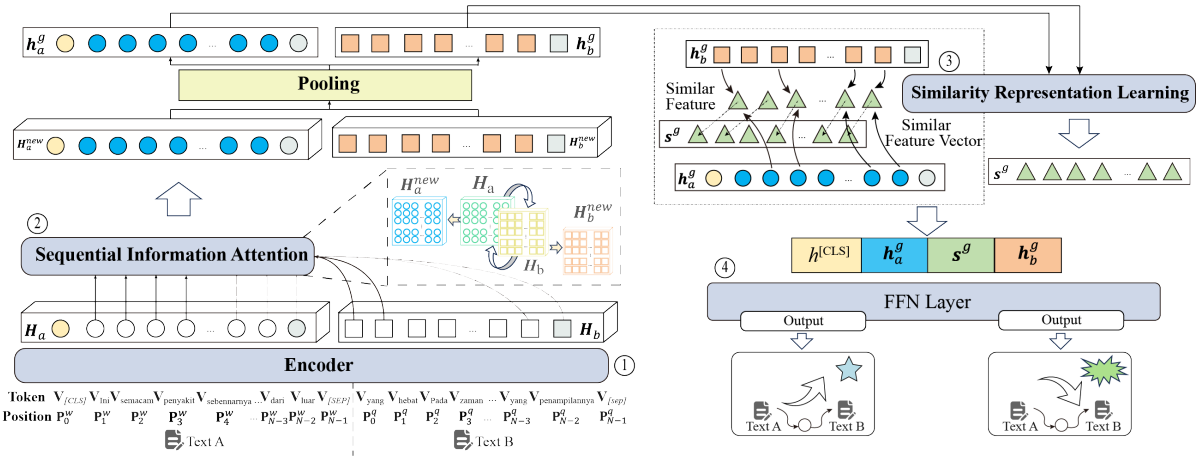

Figure 2: The architecture of our proposed model. ①, ②, ③ and ④ are text representation module, sequential information attention mechanism module, similarity representation learning module, and language development assessment module respectively. Initially, the text representation module encodes input text using the PLM, leveraging the sequential information attention mechanism to capture information interaction and obtain enhanced global representations. Similarity representation learning calculates the similarity between the enhanced representations, and the language development assessment module combines multiple representations to assess language development.

$Encoder(\cdot)$ is the pre-trained model. The semantic representation of text sequence T_a and text sequence T_b are formulated as follows:

$$H_a = \{h^{[CLS]}, h_a^1, h_a^2, h_a^3, \ldots, h_a^n, h^{[SEP]}\}, \quad (2)$$

$$H_b = \{h_b^1, h_b^2, h_b^3, \ldots, h_b^m, h^{[SEP]}\}. \quad (3)$$

4.2 Sequential Information Attention Mechanism

Inspired by Wu et al. (2023), we exploit a sequential information attention mechanism to capture information interaction between two sequence essays, enhancing the first-written text representation and the second-written text representation. Initially, we compute the attention weight w_1 of the second-written text T_b on the first-written text T_a and the attention weight w_2 of the first-written text T_a on the second-written text T_b respectively:

$$w_1 = softmax(H_a \cdot (H_b)^T), \quad (4)$$

$$w_2 = softmax(H_b \cdot (H_a)^T). \quad (5)$$

Then the semantic representation of the first-written text and the second-written text are updated with the attention weight w_1 and the attention weight w_2 to capture the difference between the two sequence texts:

$$H_a^{new} = w_1 \cdot H_a, \tag{6}$$

$$H_b^{new} = w_2 \cdot H_b. \tag{7}$$

Ultimately, for the updated H_a^{new} and H_b^{new}, we perform the average pooling operation to obtain an enhanced global first-written text representation and the second-written text representation:

$$h_a^g = pooling(H_a^{new}), \tag{8}$$

$$h_b^g = pooling(H_b^{new}), \tag{9}$$

where $pooling(\cdot)$ refers to global average pooling.

4.3 Similarity Representation Learning

Unlike the sequential attention mechanism that captures the differences between the two essays, similarity representation learning focuses on the common information of the two sequence essays. Furthermore, the similarity representation learning we adopt is different from the traditional methods directly calculating the similarity score with cosine distance or Euclidean distance between two feature vectors, which aims to learn an abstract vector-based similarity representation rather than a scalar-based similarity to capture more detailed similarity information between two feature representations. Based on the enhanced global first-written text representation and the second-written text representation, a learnable similarity vector is employed to characterize the information similarity between the two sequence essays:

$$s^g = \frac{W_g|h_b^g - h_a^g|^2}{||W_g|h_b^g - h_a^g|^2||_2}, \tag{10}$$

where W_g represents a learnable parameter matrix for deriving the similarity vector, $|\cdot|^2$ and $||\cdot||_2$ denote element-wise squaring and the L2 norm.

4.4 Language Development Assessment

The enhanced global first-written text representation h_a^g, the enhanced global second-written text representation h_b^g, raw text representation $h^{[CLS]}$ and vector-based similarity representation s^g between two sequence essays are concatenated to obtain a text representation h^c for classification:

$$h^c = concat(h^{[CLS]}, h_a^g, h_b^g, s^g), \tag{11}$$

where $h^{[CLS]}$ indicates raw text representation corresponding to the token [CLS] and h^c is an overall enhanced feature representation of the text sequence. Subsequently, the enhanced feature representation is input into a linear classifier employing a softmax function, which is formulated as follows:

$$p = softmax(W^T \cdot h^c + b), \tag{12}$$

where W and b denote learnable parameters, and $p = [p_1, p_2]$ represents the predicted probability for language learning development. p_1 is the probability that the text T_b is written before the text T_a, and p_2 represents the probability that the text T_b is written after the text T_a. Cross-entropy loss is used to compute the discrepancy between the predicted class probability and the actual expected value. More precisely, the cross-entropy loss function is defined as follows:

$$L_{ce} = -\sum_{j=1}^{2} y_j \cdot \log(p_j), \tag{13}$$

where y is a one-hot encoding of the actual expected value. Specifically, $y = [1, 0]$ indicates that the text T_b is written before the text T_a, and vice versa concerning $y = [0, 1]$.

5 Experiments and Analysis

5.1 Datasets

We conducted experiments on the self-constructed Indonesian corpus. Furthermore, we also performed language development assessment on the Italian and Spanish corpora provided by LangLearn shared task, namely CItA (Barbagli et al., 2016) and COWS-L2H (Miaschi et al., 2020). We utilized the corpus that was processed by Wu et al. (2023). Furthermore, we adopted the 5-fold cross-validation approach, similar to Wu et al. (2023), dividing the datasets into five distinct parts to build an ensemble model with enhanced generalization abilities. This method entailed using four subsets for training and one for validation, with the model's performance systematically evaluated by averaging the results from all five models.

5.2 Evaluation Metrics

For the evaluation of the experimental results in Italian and Spanish, in addition to using the evaluation metrics accuracy Acc and macro-average F value F_m of the LangLearn shared task, we also employ the evaluation metric binary F value F_b commonly used in binary classification tasks. For the Indonesian language, we report more evaluation metrics to conduct a more detailed evaluation

Method	Acc	P_b	R_b	F_b	P_m	R_m	F_m
BERT (Devlin et al., 2019; Wu et al., 2023)	88.87	84.38	95.50	89.59	89.58	88.85	88.82
SIAM (Wu et al., 2023)	89.84	84.44	**97.75**	90.61	90.88	89.81	89.77
ChatGPT	56.06	54.42	75.48	63.24	57.10	56.03	54.31
Our method	**91.61**	**89.12**	94.86	**91.90**	**91.79**	**91.60**	**91.60**

Table 4: Experimental results for the Indonesian language. BERT, SIAM, and ChatGPT are the comparison methods. Highlighted metrics represent the best performing models.

Method	Acc	P_b	R_b	F_b	P_m	R_m	F_m
Our method	**91.61**	89.12	94.86	**91.90**	**91.79**	**91.60**	**91.60**
w/o SRL	90.00	**90.03**	90.03	90.03	90.00	90.00	90.00
w/o SIAM	90.00	85.27	**96.78**	90.66	90.76	89.98	89.95

Table 5: Ablation results for Indonesian language. **"SRL"** stands for similarity representation learning module, and **"SIAM"** stands for sequential information attention mechanism module.

and analysis. Therefore, the models' performance achieved on the Indonesian test sets has been independently evaluated using metrics such as accuracy Acc, P_b, R_b, F_b, P_m, R_m, and F_m.

5.3 Experimental Settings

Experiments utilize an NVIDIA A5000 24-GB GPU with PyTorch and Transformers for model development. The feed-forward layer's weights are initialized from a truncated normal distribution (standard deviation 2e-2), and biases are set to 0. A consistent initial learning rate of 5e-5 and a maximum sequence length of 512 are applied. Training optimization includes a warmup proportion of 1e-3, spanning 10 epochs with a batch size of 4. For different languages, we select different pre-trained language models to conduct the experiments. Specifically, we select bert-base-indonesian[2] for the training of Indonesian models. For Italian and Spanish, we choose bert-base-italian-uncased[3] and bert-base-spanish-wwm-uncased[4] respectively.

5.4 Comparison Methods

We use six baselines ChatGPT, BERT (Devlin et al., 2019; Wu et al., 2023), SIAM (Wu et al., 2023), IUSS-NeTS (Barbini et al., 2023), bot.zen (Stemle et al., 2023) and ExtremITA (Alzetta et al., 2023) to verify the effectiveness of our proposed method. The more detailed descriptions of comparison methods are provided in Appendix B.

[2]https://huggingface.co/cahya/bert-base-indonesian-1.5G

[3]https://huggingface.co/dbmdz/bert-base-italian-uncased

[4]https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased

5.5 Experiments on Indonesian Dataset

As the results shown in Table 4, our proposed method achieves state-of-the-art performance compared to existing methods on a self-constructed Indonesian corpus, with 91.61, 91.90 and 91.60 in Acc, F_b and F_m respectively, demonstrating the effectiveness of our method. To be more specific, it can be seen that our method outperforms the BERT model by 2.74, 2.31, and 2.78 in the three metrics of Acc, F_b and F_m respectively. Likewise, the results reveal that our method simultaneously capturing the differences and common information gains consistent improvements compared with SIAM solely concentrating on the differences between essay pairs. It is noteworthy that when employed in language development assessment tasks, the ChatGPT reveals a significant performance gap compared to existing models, which poses a challenge to the language model's capability to understand and generate low-resource language.

5.6 Ablation Study

To investigate the contribution of key components of our model, we ablate different modules in turn and report experimental results on Indonesian datasets. The results in Table 5 show that the model's performance suffers a drop by 1.61, 1.87 and 1.60 in Acc, F_b and F_m respectively when removing the SRL. Similarly, discarding the SIAM results in a considerable performance decline in Acc, F_b and F_m by 1.61, 1.24 and 1.65, respectively. Experimental Results demonstrate the potential of our model in capturing linguistic information. Moreover, the results suggest that focusing on both

Method	Italian			Spanish			Average		
	Acc	F_b	F_m	Acc	F_b	F_m	Acc	F_b	F_m
BERT (Devlin et al., 2019; Wu et al., 2023)	**93.16**	**93.15**	**93.38**	60.94	62.01	60.91	77.05	77.58	77.15
SIAM (Wu et al., 2023)	92.51	92.51	92.60	64.06	66.28	63.91	78.29	79.40	78.26
IUSS-NeTS (Barbini et al., 2023)	64.50	-	67.30	**75.30**	-	**75.20**	69.90	-	71.25
bot.zen (Stemle et al., 2023)	83.40	-	83.70	49.70	-	51.70	66.55	-	67.70
ExtremITA (LLaMA) (Alzetta et al., 2023)	59.60	-	61.30	57.50	-	55.30	58.55	-	58.30
ExtremITA (IT5) (Alzetta et al., 2023)	60.60	-	41.00	50.60	-	16.00	55.60	-	28.50
ChatGPT	52.61	39.83	50.37	51.10	59.59	48.84	51.86	49.71	49.61
Our method	92.51	92.65	92.51	72.81	**72.03**	72.79	**82.66**	**82.34**	**82.65**

Table 6: Experimental results for Italian language and Spanish language. **"Average"** is the average metrics of the models in Italian language and Spanish language.

the differences and common information between essay pairs is conducive to prominent performance in language development assessment.

5.7 Experiments on Other Datasets

We further perform language development assessment tasks on public datasets to explore the performance of our model in Italian and Spanish, thereby demonstrating that our method possesses good generalization capabilities. Table 6 reports the results of our proposed model and comparison models for Italian Language and Spanish Language.

In terms of Italian, the BERT model achieves the most promising performance, which highlights the potential of PLMs. Although the performance of our proposed model is not as encouraging as the BERT model across metrics, its performance is still competitive compared with feature-based methods and large language model-based methods.

On Spanish datasets, IUSS-NeTS employing explicit features that measure raw text properties achieves a prominent performance, significantly surpassing the base model BERT. Whereas the feature-based model bot.zen being poor at computing features capturing text complexity for Spanish, resulting in lower scores on the Spanish corpus. Although our model does not exceed the IUSS-NeTS model, its performance gains consistent and favorable improvements compared with other methods. Concerning large language model-based methods, they are powerless in language development assessment for low-resource language, which sheds light on the necessity of promoting language comprehension capability for low-resource language.

Although our model doesn't outperform BERT or IUSS-NeTS in Table 5 in Italian and Spanish datasets, our method achieves significant results on the Indonesian dataset. Furthermore, on the Italian and Spanish datasets, the average accuracy and macro F1 of our method are 82.66 and 82.65, which are higher than BERT's 77.05 and 77.15 and IUSS-NeTS's 69.90 and 71.25. This shows that even if our method does not exceed a certain method in the open-source corpus, it has better performance and stability in terms of overall performance.

Notably, our model performs significantly better on Spanish written by L2 learners than on Italian written by L1 learners, which is attributed to greater variation in a second language in terms of style within a shorter time period compared to a first language. Consequently, the method SIAM capturing the differences between text pairs achieves superior performance compared to the base model BERT. Our proposed model simultaneously focuses on the differences and common information between essay pairs to yield further improvement.

6 Conslusion

In this paper, we introduce the ICL corpus, which, to the best of our knowledge, is the first corpus tailored for Indonesian LDA tasks. Moreover, we present a model capable of automating the extrac-

tion of language-independent features for LDA tasks, which yields improved performance on both our self-constructed corpus and publicly available corpora. Our work could serve as a novel benchmark for Indonesian LDA tasks. In addition, our research reveals that the performance of LLMs on LDA tasks still has considerable room for improvement. In the future, we will further expand our corpus to encompass additional languages. Meanwhile, we will explore innovative ways to enhance the capabilities of LLMs on LDA tasks.

Acknowledgements

This work was supported by the Guangdong Philosophy and Social Science Foundation Regular Project (No. GD20CWY10).

Limitations

In this section, we discuss the limitations of this work. Firstly, we exclusively assessed ChatGPT's performance in language development assessment, potentially overlooking other LLMs that could excel in this task. Secondly, we utilized OpenAI's LLMs API without fine-tuning the LLM, raising the possibility that fine-tuned the LLM may be better suited for this task. Thirdly, due to the limited size of the original corpus, we divided the students' essays according to length to ensure that the scale of the corpus can meet the training and testing of the model, albeit at the expense of sample fluency.

Ethics Statement

In conducting this research, we place paramount importance on ethical considerations, ensuring that our methodologies, data handling, and technological applications meet rigorous ethical standards. All students whose writings contribute to our corpus provided informed consent, fully understanding the scope and purpose of their contributions. We have implemented robust measures to safeguard their privacy, ensuring that personal identifiers are removed and data is handled in accordance with data protection regulations. The corpus in our dataset comes from second language learners who use a majority dialect. When a language development assessment model is trained based on our corpus, it poses a significant risk of harming individuals who speak the minority dialect. If the model is used in educational settings to guide curriculum development, instruction, or assessment, its bias could lead to inappropriate educational strategies for minority dialect speakers. This could exacerbate existing educational disparities and deny these students equitable opportunities for learning and development. As for the AI assistant, we utilize ChatGPT to identify textual errors and polish paper.

References

Chiara Alzetta, Dominique Brunato, F Delll'Orletta, Alessio Miaschi, Kenji Sagae, Claudia H Sánchez-Gutiérrez, and Giulia Venturi. 2023. Langlearn at evalita 2023: Overview of the language learning development task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*.

Suhas Arehalli and Tal Linzen. 2024. Neural Networks as Cognitive Models of the Processing of Syntactic Constraints. *Open Mind*, 8:558–614.

Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. CItA: an L1 Italian learners corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 88–95, Portorož, Slovenia. European Language Resources Association (ELRA).

Matilde Barbini, Emma Zanoli, Cristiano Chesi, et al. 2023. Iuss-nets at langlearn: The role of morphosyntactic features in language development assessment. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org*.

Bram Bulté and Alex Housen. 2014. Conceptualizing and measuring short-term changes in l2 writing complexity. *Journal of Second Language Writing*, 26:42–65. Comparing perspectives on L2 writing: Multiple analyses of a common corpus.

Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2023. Teacher perception of automatically extracted grammar concepts for L2 language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3776–3793, Singapore. Association for Computational Linguistics.

Scott A. Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10184–10198, Toronto, Canada. Association for Computational Linguistics.

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.

Lenardo Chaves e Silva, Álvaro Alvares de Carvalho César Sobrinho, Thiago Damasceno Cordeiro, Rafael Ferreira Melo, Ig Ibert Bittencourt, Leonardo Brandão Marques, Diego Dermeval Medeiros da Cunha Matos, Alan Pedro da Silva, and Seiji Isotani. 2023. Applications of convolutional neural networks in education: A systematic literature review. *Expert Systems with Applications*, 231:120621.

Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.

Julia Hancke and Detmar Meurers. 2013. Exploring cefr classification for german based on rich linguistic modeling. *Learner Corpus Research*, pages 54–56.

Qing Huang and Lu Wei. 2022. Explaining education-based difference in systematic processing of covid-19 information: Insights into global recovery from infodemic. *Information Processing and Management*, 59(4):102989.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 2020. Becoming linguistically mature: Modeling English and German children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–74, Seattle, WA, USA → Online. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Alessio Miaschi, Dominique Brunato, and Felice Dell'Orletta. 2021a. A nlp-based stylometric approach for tracking the evolution of l1 written language competence. *Journal of Writing Research*, 13(1):71–105.

Alessio Miaschi, Dominique Brunato, and Felice Dell'Orletta. 2021b. A nlp-based stylometric approach for tracking the evolution of l1 written language competence. *Journal of Writing Research*, 13(1):71–105.

Alessio Miaschi, Sam Davidson, Dominique Brunato, Felice Dell'Orletta, Kenji Sagae, Claudia Helena Sanchez-Gutierrez, and Giulia Venturi. 2020. Tracking the evolution of written language competence in L2 Spanish learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–101, Seattle, WA, USA → Online. Association for Computational Linguistics.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Minh Phan, Arno De Caigny, and Kristof Coussement. 2023. A decision support framework to incorporate textual data for early student dropout prediction in higher education. *Decision Support Systems*, 168:113940.

Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico. Association for Computational Linguistics.

Kenji Sagae. 2021. Tracking child language development with neural network language models. *Frontiers in Psychology*, 12:674402.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: An italian instruction-tuned llama. *arXiv preprint arXiv:2307.16456*.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.

Egon W Stemle, Martina Tebaldini, Francesca Bonanni, Filippo Pellegrino, Paolo Brasolin, Greta H Franzini, Jennifer-Carmen Frey, Olga Lopopolo, Stefania Spina, et al. 2023. bot. zen at langlearn: regressing towards interpretability. In *CEUR WORKSHOP PROCEEDINGS*, pages 1–5. CEUR-WS. org.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Zarah Weiss and Detmar Meurers. 2019. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the Fourteenth*

Workshop on Innovative Use of NLP for Building Educational Applications, pages 380–393, Florence, Italy. Association for Computational Linguistics.

Hongyan Wu, Nankai Lin, Shengyi Jiang, and Lixian Xiao. 2023. Bert_4ever at langlearn: Language development assessment model based on sequential information attention mechanism. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org*.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

A Related Work

A.1 Language Acquisition

Language acquisition (LA) is the field of study concerned with understanding how individuals acquire first or second language competencies. In recent years, with the rapid development of natural language processing (NLP) technologies, the field of LA has significantly benefited. Chaudhary et al. (2023) examined the role of automated tools in enhancing the creation and visualization of grammar descriptions for languages like Kannada and Marathi, improving the accessibility of educational materials and facilitating more effective language learning strategies. Evanson et al. (2023) investigated the learning trajectories of GPT-2 models and human children, demonstrating that both acquire linguistic skills in a systematic order, offering insights into the computational principles of language acquisition. Additionally, Oba et al. (2023) explored second language acquisition in neural language models, revealing that L1 pretraining accelerates L2 linguistic generalization and that language transfer configurations significantly impact their generalizations. Furthermore, Yadavalli et al. (2023) studied the effects of positive and negative cross-linguistic transfer in second language acquisition, highlighting that language family distance predicts more negative transfer and that conversational speech data facilitates language acquisition more effectively than scripted speech data.

A.2 Language Development Assessment

Previous studies on LDA tasks can be broadly divided into two distinct groups: one group focuses on constructing a LDA model that is based on linguistic features while the other focuses on constructing a LDA model that is based on neural networks.

Linguistic-features-based Methods. Language possesses many complex features, and this approach typically involves developing techniques to extract these complex features to construct the model. Hancke and Meurers (2013) tracked the progression of writing complexity and accuracy in German students from first through eighth grade, focusing on early academic language development. They strengthened the applicability and reliability of their models by incorporating a comprehensive set of complexity features across various writing subjects. Vajjala and Lõo (2014) focused on automating the prediction of learner language proficiency in Estonian, using morphological and POS tag extraction from learner texts. Pilán and Volodina (2018) undertook a comprehensive study of features predicting Swedish L2 learning from texts. Miaschi et al. (2020) analyzed automatically extracted linguistic features from essays to track L2 Spanish learners' written skill development. Miaschi et al. (2021b) developed a method to measure the evolution of written competence in Italian L1 learners, focusing on stylistic features. Bulté and Housen (2014) researched English L2 writing proficiency growth among 45 adult ESL learners in a short, intensive academic program, employing quantitative analyses of lexical and syntactic complexity in their writing. This study also compared these objective measures to subjective assessments of writing quality. Cui and Sachan (2023) focused on tracking the evolution of writing complexity and accuracy in German learners from elementary through middle school, highlighting the importance of linguistic complexity features in early academic language development.

Neural-networks-based Methods. Recent studies on neural network-based language modeling (e Silva et al., 2023; Arehalli and Linzen, 2024; McCoy et al., 2020) have shown that certain neural architectures can understand syntactic details from text without being directly programmed to do so. Sagae (2021) explored if a recurrent neural network, focused on data-driven language development, could track children's language progression as effectively as established language assessment metrics. Additionally, the innovative sequential information attention mechanism in SIAM (Wu et al., 2023) captured interactions between essay pairs by integrating attention weights from the most recent essay, using the "[CLS]" token and average pooling for pair representation. Barbini et al. (2023) calculated the surprisal-based metrics by extracting token probabilities from pre-trained language-specific BERT models.

However, it is crucial to efficiently extract language-independent text features and identify the differences and shared information between the first-written and second-written essays.

B Comparison Methods

We use six baselines ChatGPT, BERT (Devlin et al., 2019; Wu et al., 2023), SIAM (Wu et al., 2023), IUSS-NeTS (Barbini et al., 2023), bot.zen (Stemle

Prompt Template for Language Development Assessment

'role': 'system', 'content': You are an assessment tool that can evaluate the progression of students' {language} language development by comparing their writing proficiency.'

'role': 'user', 'content': Your task involves analyzing a pair of {language} essays authored by the same student and determining their chronological order. The input format consists of two essays labeled as "essay 1" and "essay 2". Your objective is to judge the relative order of two essays.

The format for input is as follows: (essay 1 | essay 2)

Your output should be presented in JSON format, with an example as follows: {'Is essay 1 written before essay 2?': true}

{language} sentence: ({essay 1} | {essay 2})

JSON object:'

Figure 3: An illustration of the prompt template used for ChatGPT. {language} is the language of the input sample. {essay 1} and {essay 2} respectively represent two different essays in an essay pair.

et al., 2023) and ExtremITA (Alzetta et al., 2023) to verify the effectiveness of our proposed method.

ChatGPT[5]: We investigate the performance of ChatGPT using OpenAI's official API. The model selected for evaluation was gpt-3.5-turbo, renowned for being advanced and tailored for chat-based tasks. The prompt template of ChatGPT is shown in Figure 3.

BERT (Devlin et al., 2019; Wu et al., 2023): BERT (Bidirectional Encoder Representations from Transformers) introduces a pre-trained deep learning architecture capable of understanding context and semantics in text more effectively than previous models. BERT utilizes bidirectional attention mechanisms to capture the relationships between words in both directions, empowering BERT to outperform in a wide range of NLP tasks, such as text classification, question answering, and language understanding.

SIAM (Wu et al., 2023): SIAM employs a new sequential information attention mechanism to analyze interactions between essay pairs. This approach integrates attention weights from the latest essay into the pair representation, using the "[CLS]" token and average pooling for effective processing.

IUSS-NeTS (Barbini et al., 2023): The IUSS-Nets utilize linguistic features including part-of-speech category density, syntactic constituent frequency, and mean sentence length, extracted via the Common Text Analysis Platform (CTAP). Moreover, it employs surprisal metrics based on token probabilities from pre-trained, language-specific BERT models.

bot.zen (Stemle et al., 2023): The bot.zen approaches the LDA task as a regression problem, aiming to identify the learning process stage of a student's essay. Datasets are pre-processed to sequence the essays accurately, and predictions rely on an ensemble of decision tree algorithms. The model is trained on 125 normalized features capturing lexical and morpho-syntactic properties.

ExtremITA (Alzetta et al., 2023): ExtremITA uses two models in a multi-task learning framework: an encoder-decoder from IT5-small (Sarti and Nissim, 2022) and a decoder from Camoscio (Santilli and Rodolà, 2023), the Italian LLaMA variant (Touvron et al., 2023). Both models are jointly fine-tuned with prompting techniques.

[5]https://chat.openai.com