BERT_4EVER at LangLearn: Language Development Assessment Model based on Sequential Information Attention Mechanism

Hongyan Wu[1], Nankai Lin[2,*], Shengyi Jiang[1,*] and Lixian Xiao[3]

[1]School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, PR China
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, PR China
[3]Faculty of Asian Languages and Cultures, Guangdong University of Foreign Studies, Guangzhou, Guangdong, PR China

Abstract

In recent years, investigations into language acquisition have greatly benefited from the utilization of natural language processing technologies, particularly in analyzing extensive corpora consisting of authentic texts produced by learners across the realms of first and second language acquisition. A crucial task in this domain involves the assessment of language learners' language ability development. The "Language Learning Development" task featured in EVALITA 2023 [1] marks a significant milestone as the inaugural shared task focused on automated language development assessment, which entails predicting the relative order of two essays written by the same student. We introduce a novel attention mechanism, namely sequential information attention mechanism, with the primary objective of exploiting information interaction between sequence texts. Experimental results on the COWS dataset show the effectiveness of our proposed sequential information attention mechanism, showcasing its substantial impact on model performance during the final evaluation phase.

Keywords

Language Development Assessment, Sequential Information Attention Mechanism, BERT,

1. Introduction

Recently, there has been a surge of interest in harnessing the potential of natural language processing (NLP) tools and machine learning techniques to explore the realm of language development, both in first (L1) and second language (L2) acquisition scenarios. The primary objective revolves around comprehensively characterizing the linguistic attributes of learners and the dynamic evolution of their language ability across different modalities and stages of acquisition. The utilization of learner corpora and the enhanced dependability of linguistic features extracted through computational tools and machine learning techniques have significantly advanced our comprehension of the linguistic properties exhibited by language learners. The empirical evidence has shed light on the temporal dynamics and the evolution of these language properties as learners progress in language ability [2].

A significant focus of scholarly inquiry has been directed towards the exploration of various avenues for advancing the field of language development research. One prominent line of investigation has concentrated on the formulation and refinement of automated methodologies capable of effectively dealing with intricate metrics associated with the multifaceted process of language development. By doing so, these methodologies aim to alleviate the arduous and time-consuming manual computation that domain experts traditionally undertake in their analyses [3, 4].

In parallel, another compelling track of research has embarked upon the more demanding endeavor of employing entirely data-driven approaches [5] within the domain of language development. This approach capitalizes on the wealth of information contained within textual data, leveraging diverse linguistic features that can be automatically extracted from such data sources. By harnessing the power of machine learning and computational techniques, researchers have sought to develop robust models capable of automatically assigning a learner's language production to a specific developmental level.

The data-driven approach holds significant promise, as it offers the potential to enhance the efficiency and accuracy of language development assessments. By training models on extensive datasets containing diverse samples of language production, researchers aim to uncover patterns and associations that can inform the classification of learners into distinct developmental levels. This approach not only holds the potential to expedite the process of evaluating language development but also offers valuable insights into the underlying mechanisms and

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
2754976781@qq.com (H. Wu); neakail@outlook.com (N. Lin); 200511402@oamail.gdufs.edu.cn (S. Jiang); 173829137@qq.com (L. Xiao)
0000-0003-2838-8273 (N. Lin)

progression of language acquisition.

The "Language Learning Development" task featured in EVALITA 2023 is concerned with predicting the chronological sequence of essays produced by the same student over different periods. We introduce a novel attention mechanism, namely sequential information attention mechanism (SIAM), intending to exploit information interaction between sequence texts. We submitted three results in total, namely the fine-tuned BERT model (Run 2), the fine-tuned BERT model with SIAM (Run 3), and the fusion of the results of the previous two models (Run 1). Experimental results demonstrate that our proposed sequential information attention mechanism has a remarkable impact on model performance during the final evaluation phase.

2. Related Work

The existing research on the language development assessment task is mainly divided into two types, one focuses on the construction of the language ability development assessment model based on language features, and the other is concerned with the construction of a language ability development assessment model based on neural networks.

Given the inherent challenge of establishing a unique indicator of linguistic complexity within the domain of second language (L2) development, a diverse range of features spanning various linguistic levels have been employed as inputs for supervised classification systems. These systems are trained on genuine learner data pertaining to different L2 languages. Notable examples include the works of Hancke and Meurers [6] as well as Vajjala and Lõo [7], which respectively investigated L2 German and L2 Estonian. Pilán and Volodina [8] provided a comprehensive analysis of predictive features extracted from both receptive and productive texts within the context of Swedish L2 acquisition. Miaschi et al. [9] used various linguistic features automatically extracted from students' written expressions to track the evolution of written language abilities of second-language Spanish learners. Furthermore, Miaschi et al. [5] proposed a natural language processing-based style measure to track the evolution of Italian L1 learners' written language competence, which relied on capturing a range of linguistically motivated features in terms of text style. In a study conducted by Bulté and Housen [10], the objective was to determine the nature and extent of English L2 writing proficiency development among 45 adult ESL learners throughout the duration of an intensive short-term academic English language program. The investigation employed quantitative measures that specifically targeted various aspects of lexical and syntactic complexity exhibited in the learners' writing performance. Additionally, the study aimed to establish a comparison between the scores obtained from these measures and the subjective ratings provided for the overall writing quality of the learners.

Recent work on the application of neural networks to language modeling has shown that models based on certain neural architectures can capture syntactic information from utterances and sentences even without explicit syntactic goals. Sagae [11] conducted a study to determine whether a fully data-driven model of language development, utilizing a recurrent neural network encoder to encode utterances, could track changes in children's language over the course of their language development in a comparable manner to the leverage of expertly established language assessment metrics for language-specific information.

The untapped potential of neural networks in language development assessment tasks necessitates further exploration, as the application of pre-trained models in this context has not been investigated.

3. Method

3.1. Overview

Figure 1 provides a comprehensive depiction of our methodology. Initially, we concatenate the historical text and current text, and utilize the pre-trained model BERT to encode them. Subsequently, we employ the sequential information attention mechanism to capture the interaction of information within the sequence of text, thereby updating the representation of the historical text to obtain an improved global representation. Ultimately, we combine the enhanced global representation of the historical text with the original sentence representation for the final assessment of language development.

3.2. Text Representation

Aiming to effectively capture the intricate semantic information embedded within the text, we employ a non-autoregressive pre-trained model BERT [12] renowned for its remarkable performance in generating text-based semantic representations for sentence encoding. BERT possesses abundant linguistic, syntactic, and lexical knowledge, which is acquired through unsupervised training on a substantial corpus during the pre-training phase. The fundamental architecture of the model encompasses a multi-layer bidirectional Transformer encoder [13], facilitating global information processing and extraction. Given a historical text $T_a = \{w_a^1, w_a^2, w_a^3, ..., w_a^n\}$ and a current text $T_b = \{w_b^1, w_b^2, w_b^3, ..., w_b^m\}$, two special tokens $[CLS]$ and $[SEP]$ of BERT are utilized to stitch them together, forming the text input $T = \{[CLS], w_a^1, w_a^2, w_a^3, ..., w_a^n, [SEP], w_b^1, w_b^2, w_b^3, ..., w_b^m, [SEP]\}$

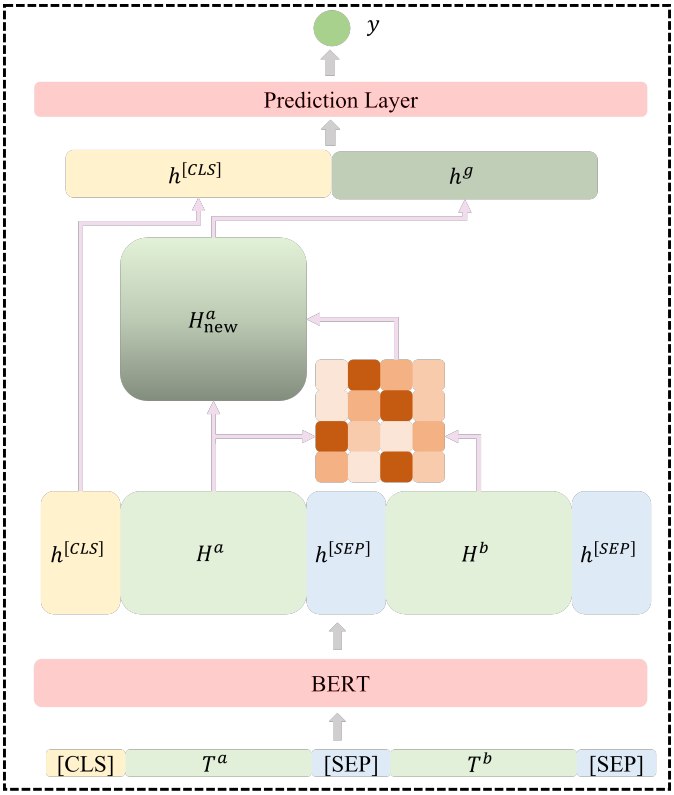

Figure 1: Model Framework Diagram.

to the pre-trained model BERT, where n and m represent the length of the two texts respectively. The semantic representation H corresponding to T encoded by the BERT pre-trained model is:

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

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

Then the semantic representation of text T_a and text T_b is respectively:

$$H^a = \{h_a^1, h_a^2, h_a^3, ..., h_a^n\} \tag{2}$$

$$H^b = \{h_b^1, h_b^2, h_b^3, ..., h_b^m\} \tag{3}$$

3.3. Sequential Information Attention Mechanism

We present an innovative attention mechanism known as the sequential information attention mechanism (SIAM), specifically designed to exploit information interaction between sequence texts. Initially, we calculate the attention weight of the current text T_b to the historical text T_a:

$$w = softmax(H^a \cdot (H^b)^T) \tag{4}$$

Historical text H^a_{new} is updated with attention weights to capture differences between the current text and the historical text:

$$H^a_{new} = w \cdot H^a \tag{5}$$

For the updated H^a_{new}, we use the average pooling operation to obtain an enhanced global representation of historical text:

$$h^g = pooling(H^a_{new}) \tag{6}$$

3.4. Language Development Assessment

We concatenate the enhanced global representation h^g of historical text with the original sentence representation $h^{[CLS]}$ to obtain a text representation for classification:

$$h^c = concat(h^{[CLS]}, h^g) \tag{7}$$

where $h^{[CLS]}$ represents the semantic representation associated with the token "[CLS]" within the given sentence. The representation h^c is utilized as the sentence's overall feature representation, which is subsequently fed into a linear classifier with a softmax function. The predicted probabilities language development assessment of are:

$$y = softmax(W^T \cdot h^c + b) \tag{8}$$

where W and b are learnable parameters. The cross-entropy loss is employed to calculate the loss that penalizes the predicted class probability based on how far it is from the actual expected value. The cross-entropy loss function of language development assessment L_{ce} is defined as:

$$L_{ce} = -\sum_{j=1}^{2} e_j \cdot log(y_j) \tag{9}$$

where e is the one-hot encoding of the text's actual expected value. When $e = 1, 0$ means that the writing time of the text T_b is before the text T_a; otherwise, when $e = 0, 1$ means that the writing time of the text T_b is after the text T_a.

4. Experiments

4.1. Experimental Setup

All experimental procedures are conducted utilizing the NVIDIA A30 24-GB GPU. We utilize pytorch [14] and transformers [15] to build our models. Considering the similarity between the two languages, we only use the Italian BERT model (dbmdz/bert-base-italian-uncased), as we think it also contains a small amount of Spanish information. The feed-forward layer is initialized using weights drawn from a truncated normal distribution with a standard deviation of 2e-2, while the bias is initialized to zero. A fixed initial learning rate of 5e-5 is consistently applied across all experiments. The maximum sequence length is set to 512, representing the prescribed constraint on the number of tokens within a sentence. To optimize training, a warmup proportion of 1e-3 is implemented. The training episodes span 10 epochs with a batch size of 4.

4.2. Datasets

The datasets provided by EVALITA 2023 "Language Learning Development" task come from two samples, CItA [16] and COWS-L2H [9], where the number of training sets is 2394 and 1009 respectively. We perform data augmentation based on the datasets. Specifically, if essay 1 in sample 1 appears before essay 2, we describe it as (A_1,A_2,1), where '1' denotes the positive sample. While essay 2 in sample 2 appears before essay 3, we describe it as (A_2,A_3,1). Then we construct sample 3 based on the above two samples. In terms of sample 3, essay 1 appears before essay 3, which is defined as (A_1,A_3,1). In addition, we expand the negative samples based on the above positive samples, namely (A_2,A_1,0), (A_3,A_2,0), and (A_3,A_1,0), where '0' represents the the negative sample. The scales the augmented datasets for CItA and COWS-L2H are 5056 and 2042, respectively. In the training set, each positive sample can match a corresponding negative sample, so the number of positive and negative samples in the dataset is consistent. Ultimately, two datasets are combined to get a new training set.

In order to ensure the rationality of our strategy evaluation, we employ a 5-fold cross-validation methodology, which involves dividing the datasets into five distinct subsets to construct an ensemble model that exhibits enhanced generalization capabilities. More precisely, four of these subsets are assigned for training purposes, while the remaining subset is utilized for verification. The effective evaluation results of our strategies are derived by averaging the outcomes obtained from the five models.

4.3. Submission

We submit three results in total, namely the fine-tuned BERT model (Run 2), the fine-tuned BERT model with SIAM (Run 3), and the merge method (Run 1). The fine-tuned BERT model is to fine-tune directly on the BERT model based on the dataset. Concretely, the model in section 3 removes the sequential information attention mechanism. The merge method is the fusion of output probabilities of the fine-tuned BERT model (Run 2) and the fine-tuned BERT model with SIAM (Run 3).

4.4. Experimental results

Experimental results in the evaluation phase are shown in Table 1.

It can be seen that on the CItA test set, the BERT model achieves the best performance, F_{binary}, F_{macro} and the accuracy are 0.9338, 0.9315 and 0.9316 respectively, while the BERT model with SIAM has slightly declined. We deem that the impunity can be attributed to our methods being trained on two corpora simultaneously, to some extent, the information of the two corpora affects each other, sacrificing the performance of the CItA dataset in exchange for the improvement of the COWS dataset. Concerning the merge method, regardless of the CItA test set, the COWS test set or the combined test set, the strategy of model fusion is powerless, which has not brought effective improvement.

Our proposed sequential information attention mechanism has demonstrated substantial improvements in both the COWS test set and the combined test set. Specifically,

Table 1
Main Result.

Dataset	Metrics	BERT (Run 2)	BERT with SIAM (Run 3)	Merge (Run 1)
CItA	F_{binary}	0.9338	0.9260	0.9270
	F_{macro}	0.9315	0.9251	0.9250
	Acc	0.9316	0.9251	0.9251
COWS	F_{binary}	0.6201	0.6628	0.6306
	F_{macro}	0.6091	0.6391	0.6150
	Acc	0.6094	0.6406	0.6156
All	F_{binary}	0.7740	0.7883	0.7747
	F_{macro}	0.7669	0.7796	0.7669
	Acc	0.7671	0.7799	0.7671

on the COWS test set, the BERT model with SIAM outperforms the BERT model by 0.0427, 0.0300, and 0.0313 in the three indicators of F_{binary}, F_{macro}, and accuracy, respectively. Likewise, on the combined test set, the BERT model with SIAM gains consistent improvement of 0.0143, 0.0126, and 0.0128 in three metrics based on the BERT model.

5. Conclusion

The "Language Learning Development" task revolves around accurately predicting the sequential order of two essays authored by a single student. In the study, we first attempt to tackle the task leveraging a high-performing pre-trained language model, demonstrating the strong potential of pre-trained language models to solve the language development assessment task. Moreover, we present a novel attention mechanism, known as sequential information attention, designed to effectively capture and leverage the interaction of information within sequential texts. In the final evaluation stage, experimental results reveal the effectiveness of our proposed method, substantiating that sequential information attention contributes to tracking the evolution of language competence.

In the future, we will further try to focus on neural networks to extract language features suitable for language development assessment tasks, so as to further improve the performance of the model, driving advancements in the field of language development assessment.

Acknowledgments

This work was supported by the Guangdong Philosophy and Social Science Foundation (No. GD20CWY10), the National Social Science Fund of China (No. 22BTQ045), and the Science and Technology Program of Guangzhou (No.202002030227).

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] S. A. Crossley, Linguistic features in writing quality and development: An overview, Journal of Writing Research 11 (2020) 415–443. URL: https://jowr.org/index.php/jowr/article/view/582. doi:10.17239/jowr-2020.11.03.01.

[3] K. Sagae, A. Lavie, B. MacWhinney, Automatic measurement of syntactic development in child language, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 197–204. URL: https://aclanthology.org/P05-1025. doi:10.3115/1219840.1219865.

[4] X. Lu, Automatic measurement of syntactic complexity in child language acquisition, International Journal of Corpus Linguistics 14 (2009) 3–28. doi:10.1075/ijcl.14.1.02lu.

[5] A. Miaschi, D. Brunato, F. Dell'Orletta, A nlp-based stylometric approach for tracking the evolution of l1 written language competence, Journal of Writing Research 13 (2021) 71–105. URL: https://www.jowr.org/index.php/jowr/article/view/778. doi:10.17239/jowr-2021.13.01.03.

[6] J. Hancke, D. Meurers, Exploring cefr classification for german based on rich linguistic modeling, 2013, pp. 54–56.

[7] S. Vajjala, K. Lõo, Automatic CEFR level prediction for Estonian learner text, in: Proceedings of the third workshop on NLP for computer-assisted language learning, LiU Electronic Press, Uppsala, Swe-

den, 2014, pp. 113–127. URL: https://aclanthology.org/W14-3509.

[8] I. Pilán, E. Volodina, 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, Association for Computational Linguistics, Santa Fe, New-Mexico, 2018, pp. 49–58. URL: https://aclanthology.org/W18-4606.

[9] A. Miaschi, S. Davidson, D. Brunato, F. Dell'Orletta, K. Sagae, C. H. Sanchez-Gutierrez, G. Venturi, 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, Association for Computational Linguistics, Seattle, WA, USA → Online, 2020, pp. 92–101. URL: https://aclanthology.org/2020.bea-1.9. doi:10.18653/v1/2020.bea-1.9.

[10] B. Bulté, A. Housen, Conceptualizing and measuring short-term changes in l2 writing complexity, Journal of Second Language Writing 26 (2014) 42–65. URL: https://www.sciencedirect.com/science/article/pii/S1060374314000666. doi:https://doi.org/10.1016/j.jslw.2014.09.005, comparing perspectives on L2 writing: Multiple analyses of a common corpus.

[11] K. Sagae, Tracking child language development with neural network language models, Frontiers in Psychology 12 (2021) 674402. doi:10.3389/fpsyg.2021.674402.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[16] A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, 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), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 88–95. URL: https://aclanthology.org/L16-1014.