Pioneering Granularity: Advancing Native Language Identification in Ultra-Short EAP Texts

Zhendong Du, Kenji Hashimoto
Graduate School of Information, Production and Systems
Waseda University, Fukuoka, Japan 808-0135

***Abstract*—This study addresses the challenge of Native Language Identification (NLI) in ultra-short English for Academic Purposes (EAP) texts by proposing an innovative two-stage recognition method. Conventional views suggest that ultra-short texts lack sufficient linguistic features for effective NLI. However, we have found that even in such brief texts, subtle linguistic cues—such as syntactic structures, lexical choices, and grammatical errors—can still reveal the author's native language background. Our approach involves fine-tuning the granularity of first language (L1) labels and refining deep learning models to more accurately capture the subtle differences in second language (L2) English texts written by individuals from similar cultural backgrounds. To validate the effectiveness of this method, we designed and conducted a series of scientific experiments using advanced Natural Language Processing (NLP) techniques. The results demonstrate that models adjusted for granular L1 distinctions exhibit greater sensitivity and accuracy in identifying language variations caused by nuanced cultural differences. Furthermore, this method is not only applicable to ultra-short texts but can also be extended to texts of varying lengths, offering new perspectives and tools for handling diverse language inputs. By integrating in-depth linguistic analysis with advanced computational techniques, our research opens up new possibilities for enhancing the performance and adaptability of NLI models in complex linguistic environments. It also provides fresh insights for future efforts aimed at optimizing the capture of linguistic features.**

***Keywords*—Native language identification; English for academic purposes; natural language processing**

I. Introduction

NLI is a technique used to determine an author's native language by evaluating their written text. This method plays a crucial role in L2 writing research [1]. Traditionally, it has been assumed that effective NLI requires longer texts [2], as extended content provides a broader range of stylistic and linguistic features, making it easier to infer the author's native language with greater accuracy. This belief stems from the linguistic diversity and variability seen across different languages, as individuals' written expressions are shaped by their unique cultural and educational backgrounds [3]. However, these influences are often difficult to capture in shorter texts, leading to the widespread misconception that extremely short texts are unsuitable for NLI.

This assumption, however, overlooks certain fundamental aspects of linguistics. Even in very short texts, specific linguistic features—such as grammatical structures, lexical choices, and common errors—can reveal clues about the author's native language. For instance, writers from different language backgrounds often follow identifiable patterns when using prepositions, articles, or complex sentence structures [4]. These patterns may still be present even in short text segments. Moreover, mistakes such as misusing certain tenses or irregular verb forms can provide valuable insights for NLI.

In many practical applications, it is common to encounter texts at the sentence level or even shorter [5]. This reality has driven the exploration of NLI techniques capable of handling ultra-short texts, particularly in the context of EAP. In EAP writing, each word choice and grammatical structure may reflect the author's linguistic habits and native language influences [6]. Thus, even brief texts, such as titles, abstracts, or notes, can contain enough information to facilitate effective NLI.

Recognizing that traditional NLI approaches often overlook the subtle linguistic cues in ultra-short texts, this study proposes an innovative two-stage NLI method. By fine-tuning the granularity of NLI classification labels, we enhance the system's ability to analyze very short texts. This approach challenges conventional wisdom and proves particularly useful for academic English, enabling accurate identification of subtle differences in expression that arise from cultural and linguistic background variations.

Moreover, our method emphasizes the importance of linguistic analysis to uncover and utilize nuanced differences, such as slight variations in grammatical structures or lexical preferences. These aspects are often neglected in traditional NLI methodologies. To validate the effectiveness of our approach, we conducted a series of targeted experiments using advanced machine learning models. The results demonstrate that our method not only excels in identifying native languages from ultra-short texts but also offers more precise language analysis for applications like academic writing assistants and Grammatical Error Correction (GEC) tools. These applications can provide more tailored writing suggestions and GEC by accurately identifying the author's native language, ultimately improving both the quality and efficiency of writing. This study's success paves the way for further integration of linguistic research and natural language processing technologies.

II. Related Work

A. Overview

NLI aims to infer an author's native language based on the text written in a target language. This task holds significant importance in the field of NLP due to its various applications. For instance, NLI can be leveraged to improve language teaching methods, enhance the quality of machine translation, and bolster security monitoring capabilities. By understanding an

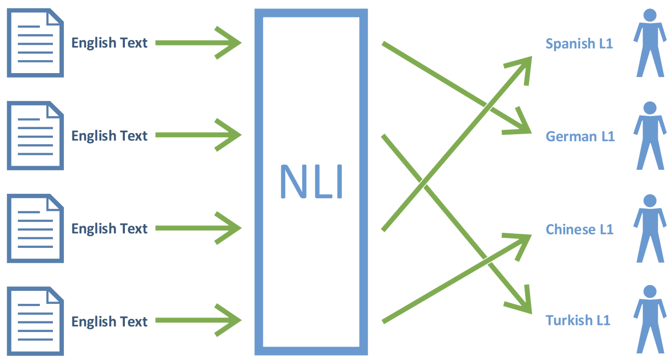

Fig. 1. The general concept of an NLI system is depicted in this figure. Image adapted from [7].

author's native language, we can more accurately identify their linguistic habits and potential errors [8], thereby providing valuable support for downstream tasks [9]. Fig. 1 illustrates the concept of NLI.

B. English NLI

English NLI is the most extensively studied area within NLI. This prominence is due to the widespread use and influence of English as the primary international language. In English NLI, researchers focus on extracting linguistic features from English texts that reflect the author's native language characteristics. Most existing studies have concentrated on long texts, such as essays [10], articles [11], and speeches [12]. These long texts provide a wealth of data, enabling models to capture the linguistic habits and preferences of speakers from different native language backgrounds. Common analytical dimensions include lexical usage frequency [13], syntactic structures [14], and pragmatic features [15].

Academic English NLI is a specialized subfield of NLI, aiming to infer the author's native language background through the analysis of academic English texts. Due to the relatively uniform style and conventions of academic English, extracting native language features poses a greater challenge [16]. Nevertheless, authors from different native language backgrounds exhibit variations in lexical choices [17], syntactic complexity [18], and argumentation styles within academic writing. Some studies utilize corpora comprising academic papers, research reports, and students' academic writings to analyze features such as discipline-specific terminology [20], the frequency of passive voice usage [21], and the overall text organization structure [22].

C. Short Text NLI

It is important to note that current research primarily focuses on NLI for long texts [23], while there is skepticism regarding the feasibility of effective NLI for very short texts, such as at the sentence level. This skepticism stems from the limited linguistic features available in ultra-short texts, making it challenging for models to capture stable native language traits [24]. Despite this, NLI for short texts is essential for specific NLP downstream tasks, such as English GEC or writing assistance tools. Unfortunately, due to the prevailing view that NLI has a minimum text length requirement, researchers have not yet explored this area adequately.

We argue that the conclusion regarding the limitations of NLI for short texts arises because prior studies did not consider different approaches for varying text lengths. Long texts, such as paragraphs, contain a wealth of linguistic features, and methods developed for them are often inappropriately applied to sentence-level texts. Hence, the current state of research in this area requires further investigation and reevaluation.

III. Goal

This study aims to address the gap in NLI research for very short texts by proposing a systematic and scientific solution. To ensure the quantifiability of the research, the focus is specifically on EAP, with the text length constrained to the sentence level.

IV. Method

NLI is a typical text classification task where the core challenge lies in enabling the NLI system to effectively learn to balance different classification labels [25]. Traditional classification methods often struggle when the NLI system is presented with sentence-level texts that have minimal distinguishing features. However, even at the sentence level, there are usually subtle differences present [26]; the key challenge is how to identify and leverage these differences.

A recent study explored this by extracting sentences from academic papers and translating them into different languages using a specific method [27]. These sentences were then translated back into English by English L2 speakers whose native languages matched the target language. The researchers found that texts produced by L2 speakers from similar cultural backgrounds exhibited certain similarities, which were evident in aspects such as grammatical errors and linguistic style. Fig. 2 illustrates part of the study's findings, showing the distribution of grammatical errors. Some researchers may view these results as reinforcing the idea that such linguistic similarities make sentence-level NLI even less feasible. However, we take an opposing stance.

We propose that, by dynamically adjusting the granularity of classification labels during the training of an NLI system, and allowing the system to learn the features of English texts in stages, it becomes possible to capture those subtle features more effectively. This approach may enhance the system's ability to identify nuanced linguistic characteristics that are otherwise overlooked in conventional classification methods.

Fig. 2 reveals some intriguing patterns in the distribution of grammatical error types among certain L2 learners. Notably, learners from China, Japan, and South Korea exhibit highly similar distributions of grammatical errors. According to the original authors, this similarity can be attributed to the cultural and linguistic ties shared by these three countries, which result in analogous challenges in learning English as a second language.

Our proposed approach involves training the model to first learn the shared linguistic features present in English texts produced by L2 learners from these three countries. Following this, the model will focus on the subtle, unique characteristics

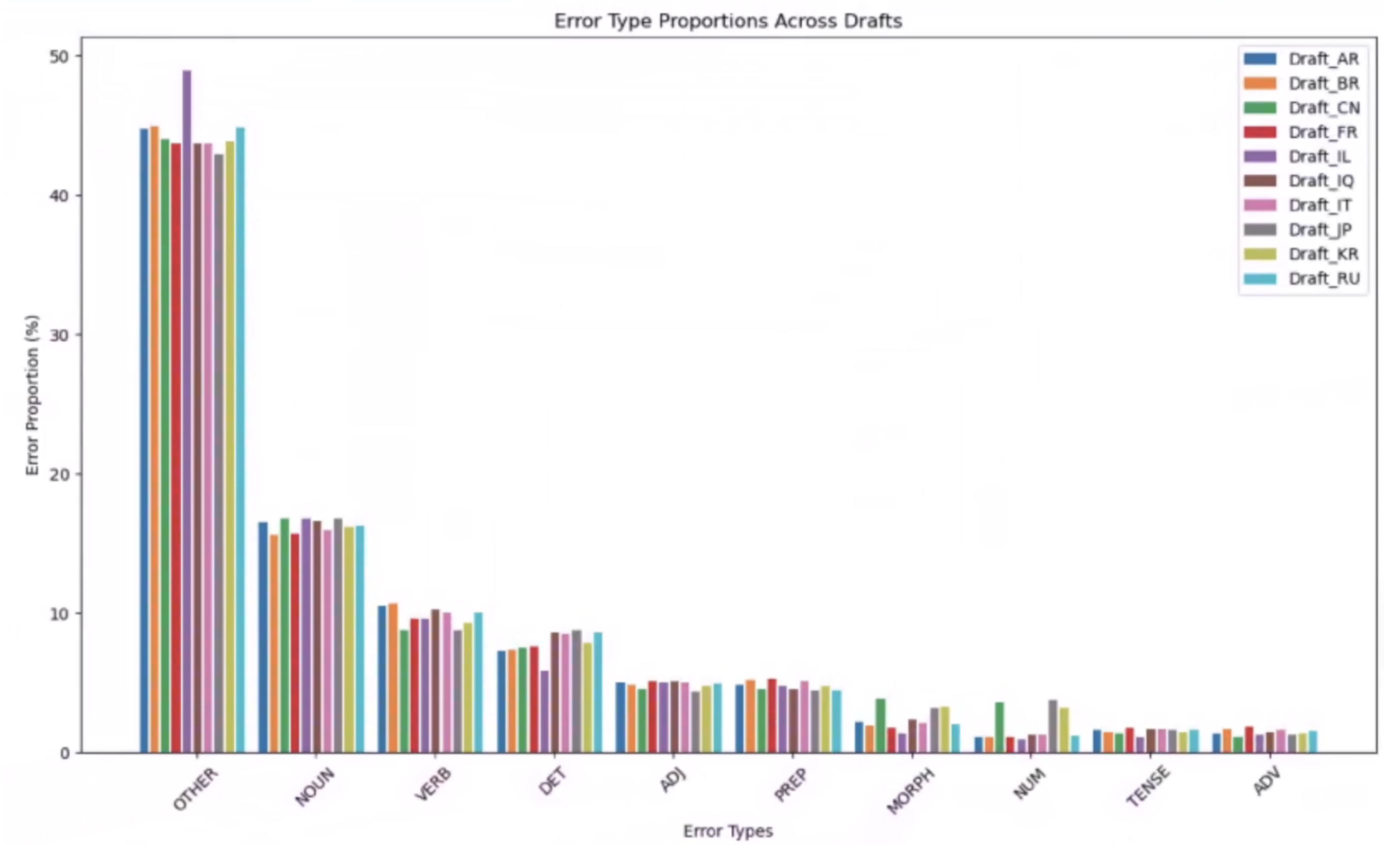

Fig. 2. Distribution of grammatical error types across different L2 transcribed texts. Different colors represent the native languages of various L2 speakers. Image adapted from [27].

of English usage specific to each country. Although similar patterns of error distribution may exist among L2 learners from other countries, our study is constrained by the availability of existing research data. Therefore, we have chosen to focus on learners from China, Japan, and South Korea. The specific methodology is detailed in Fig. 3.

V. Experiment

A. Data

In our study, we have selected the validation set of the TCNAEC [27] dataset as the training data and the test set as the testing data. To the best of our knowledge, there are currently no other publicly available datasets that meet the specific requirements of our task. While there are some English datasets consisting of paragraph collections that can be segmented into sentences, they lack the necessary labeled test data. Specifically, our task requires both the training and testing datasets to include English texts produced by L2 speakers from China, Japan, and South Korea. TCNAEC is the only dataset that satisfies this condition. It comprises 10 categories labeled as AR (Argentina), BR (Brazil), CN (China), FR (France), IL (Israel), IQ (Iraq), IT (Italy), JP (Japan), KR (South Korea), and RU (Russia), with each label containing 1,000 validation entries and 1,000 test entries.

B. Model

In this study, we selected the RoBERTa-large [28] model as the pre-trained foundation for our experiments. RoBERTa-large was chosen due to its robust text classification capabilities, which makes it well-suited to validate the proposed approach. The experiment was divided into two phases, each with distinct objectives and configurations.

In the first experiment, we followed the conventional method where data, balanced across original labels, was input into the model. This model is referred to as RoBERTa-10[1]. For the second experiment, we implemented our proposed method, where the three labels—CN, JA, and KR—were initially merged into a single label, CJK. This produced an intermediary model, referred to as NLI Model 1. Subsequently, using the same settings as in the first experiment, the data was reclassified into CN, JA, and KR labels, resulting in NLI Model 2, named RoBERTa-8to10[2].

The two experiments employed distinct hyperparameter configurations across different phases of training. In the first experiment, parameters were set according to configuration two (shown in Table II). For the second experiment, the first phase used configuration one (Table I), while the second phase used configuration two. The rationale behind using different parameter configurations for each phase was to address the

[1]For more details about the **RoBERTa-10**, visit **this link**.

[2]For more details about the **RoBERTa-8to10**, visit **this link**.

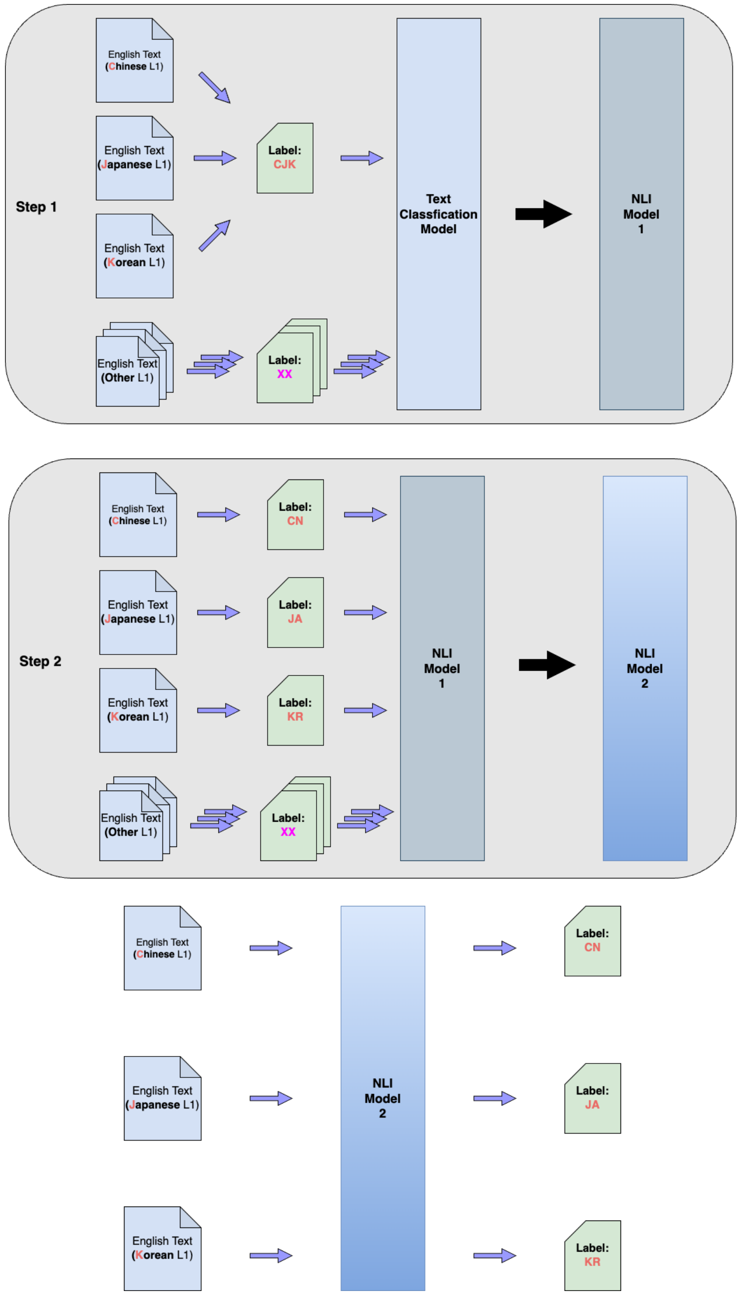

Fig. 3. Method details.
Step 1 Consolidates English text from Chinese, Japanese, and Korean L1 into the ***CJK*** label and feeds it, along with other labels, into a Text Classification Model to create **NLI Model 1**.
Step 2 Re-divides the ***CJK*** label into ***CN***, ***JA***, ***KR*** labels, which, alongside other labels, are fed into **NLI Model 1** to generate **NLI Model 2**. **NLI Model 2** can then more accurately identify and categorize English text from Chinese, Japanese, and Korean L1.

varying demands of the tasks and differences in the distribution of data labels.

Step 1: Establishing Baseline Classification Capabilities: The objective of the first phase was to establish NLI Model 1, which was designed to handle tasks involving broader categorical distinctions. Specifically, the model's primary goal was to classify English text originating from Chinese, Japanese, and Korean into the unified CJK label, alongside other distinct labels. This strategy aimed to capture key distinguishing features across broader language groups while ignoring minor linguistic variations, thereby forming a foundational understanding and classification ability for the main language groups.

To accommodate this requirement, we opted for a relatively large batch size (32) and a high learning rate (0.0001) to facilitate rapid convergence in the early stages of training. A cosine learning rate scheduler with restarts was employed to help the model effectively navigate potential local minima during training. Additionally, we adjusted the class weights to reflect label imbalances, particularly assigning a lower weight (0.333) to the CJK class to prevent the model from disproportionately favoring this category, which combined texts from three different languages.

Step 2: Enhancing Language Recognition Precision: In the second phase, the focus shifted towards more fine-grained language recognition. The objective of RoBERTa-8to10 in this phase was to differentiate between CN, JA, and KR labels, building on the language recognition capabilities developed in the first phase. To capture the nuanced differences between these languages, we adopted a smaller batch size (16) and a lower learning rate (0.00005), encouraging the model to become more sensitive to subtle details and achieve higher accuracy. To further prevent overfitting and ensure better generalization across fine-grained categories, we increased the dropout rate to 0.5.

By employing this step, targeted training strategy with parameter adjustments tailored to each phase's needs, RoBERTa-8to10 was able to improve classification accuracy and sensitivity when distinguishing between Chinese, Japanese, and Korean texts. This approach successfully met the distinct objectives of each phase.

TABLE I. HYPERPARAMETER SETTING 1

Parameter	Value
Mixed Precision	none
Optimizer	adamw_torch
Scheduler	cosine_with_restarts
Batch Size	32
Epochs	20
Gradient Accumulation	1
Learning Rate	0.0001
Maximum Sequence Length	128
Dropout Rate	0.3
Class Weight	AR: 1 BR: 1 CJK: 0.333 FR: 1 IL: 1 IQ: 1 IT: 1 RU: 1

TABLE II. HYPERPARAMETER SETTING 2

Parameter	Value
Mixed Precision	none
Optimizer	adamw_torch
Scheduler	cosine_with_restarts
Batch Size	16
Epochs	30
Gradient Accumulation	1
Learning Rate (lr)	0.00005
Maximum Sequence Length	128
Dropout Rate	0.5
Class Weight	balanced

VI. RESULTS

A. Analysis

We compared the classification performance of two models, RoBERTa-10 and RoBERTa-8to10, through metric evaluations (Table III) and confusion matrix analyses (Fig. 4) to explore the impact of a phased classification strategy on model performance.

TABLE III. MODEL PERFORMANCE METRICS

Model	Metric	CN	JP	KR
RoBERTa-10	Precision	28.02	35.29	10.12
	Recall	**78.60**	1.20	2.50
	$F_{0.5}$	32.16	5.28	6.29
	Average $F_{0.5}$		14.58	
RoBERTa-8to10	Precision	**33.06**	**42.37**	**16.85**
	Recall	59.50	**2.50**	**16.80**
	$F_{0.5}$	**36.28**	**10.11**	**16.84**
	Average $F_{0.5}$		**21.08**	

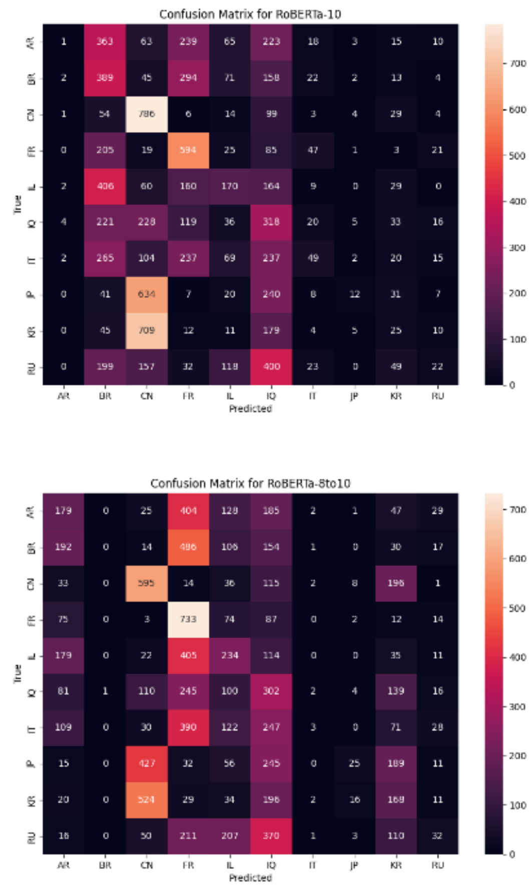

Fig. 4. Confusion matrix for models.

In the RoBERTa-10 model, all 10 labels were used simultaneously for training and prediction. The metric evaluations show that the model achieved a relatively high recall rate (78.60) for the CN label, but its precision was low (0.2802), resulting in a relatively low $F_{0.5}$ (32.16). This suggests that although the model successfully captured a significant portion of CN-labeled samples, its accuracy was insufficient, with a high number of misclassifications. The confusion matrix further confirmed this: while 786 CN samples were correctly classified, a considerable number of samples were misclassified as JP or KR. For the JP and KR labels, the recall rates were extremely low, with only a small number of samples correctly classified, and most were misclassified as CN. This indicates that the model struggled significantly in distinguishing between these three labels.

In contrast, the RoBERTa-8to10 model employed a phased classification strategy. First, the CN, JP, and KR labels, which were prone to confusion, were merged into a new CJK label. This CJK label, along with the remaining seven labels, was used for initial model training. In the second phase, samples predicted as CJK were further subdivided. Metric evaluations showed an improvement in precision for the CN label to 33.06, with a corresponding increase in the $F_{0.5}$ score to 36.28, despite a slight drop in recall (0.5950). This indicates that the model effectively reduced misclassifications while maintaining a good ability to identify CN-labeled samples. For the JP and KR labels, both precision and $F_{0.5}$ scores saw significant improvements, particularly for the JP label, where precision increased from 35.29 to 42.37. The confusion matrix revealed that the number of correctly classified JP samples doubled to 25, while the number of correctly classified KR samples surged to 168. This highlights the effectiveness of the phased classification strategy in enhancing the model's ability to differentiate between these confusing labels [29].

In summary, the RoBERTa-8to10 model demonstrated substantial improvements in classification performance for the CN, JP, and KR labels. By merging the easily confused labels in the initial phase, the phased classification strategy reduced the complexity of the initial classification task. In the subsequent fine-tuning phase, the model was able to better learn the subtle differences between these labels, leading to enhanced precision and reliability. These results validate the advantages of a phased classification approach in multi-label classification tasks and provide an effective solution for addressing similar classification challenges.

B. Discussion and Future Work

In this work, we have demonstrated the effectiveness of the proposed method, showing that the phased classification strategy has significant advantages in handling multi-label classification tasks with easily confused labels. Due to limitations in data volume and the available types of data, our current exploration is restricted to English texts written by native speakers of Chinese, Japanese, and Korean. The linguistic and cultural similarities among these three languages make the classification task more challenging, but also provide an ideal testing ground for our model.

Looking ahead, we aim to address the issue of data scarcity. By collecting and constructing larger, more diverse datasets, we hope to validate our approach across a broader range of languages and cultural contexts. This would not only improve the generalization capabilities of the model but also allow us to explore the application of the phased classification strategy to more complex groupings and finer-grained labels.

Additionally, while this study focuses on sentence-level text classification, our approach is theoretically applicable to longer texts as well. For example, texts at the paragraph or full-article level contain richer contextual and semantic information [30], which could enable the model to more accurately capture linguistic features and distinctions. Thus, we plan to extend our method to longer texts in future research, with the expectation of achieving even better classification performance.

VII. Conclusion

In this study, we focus on NLI for ultra-short EAP texts and propose an innovative methodological strategy. This strategy primarily involves the phase-wise adjustment of L1 label granularity to fine-tune deep learning models, thereby more accurately capturing the subtle linguistic features of L2 English texts from similar cultural backgrounds. The core of this method integrates in-depth grammatical analysis with advanced computational techniques, combining traditional linguistic knowledge with modern machine learning technologies to enhance model performance and adaptability in complex linguistic environments. To validate our approach, we designed a series of scientific experiments that clearly demonstrate the effectiveness of the proposed method by comparing model performance before and after the experiments. The results show that the granularity-adjusted models are more sensitive to and can better identify linguistic variations caused by subtle cultural differences, thereby significantly improving the accuracy of NLI. Moreover, we conducted multiple rounds of verification to ensure the reliability and repeatability of the results. Through exhaustive data analysis, we not only proved the effectiveness of our method but also explored its potential for future research. The analysis indicates that this strategy is applicable not only to ultra-short texts but can also be effectively applied to texts of varying lengths, offering new perspectives and tools for managing diverse linguistic inputs. Our research further reveals that further optimization of the label granularity adjustment strategy could allow for more precise capture of linguistic features, providing new directions and possibilities for future research.

Acknowledgment

This work was supported by the Information, Production and Systems Research Center, Waseda University; the Future Robotics Organization, Waseda University; the Humanoid Robotics Institute, Waseda University, under the Humanoid Project; and the Waseda University Grant for Special Research Projects (grant number 2024C-518).

References

[1] S. Malmasi and M. Dras, "Multilingual native language identification," Natural Language Engineering, vol. 23, no. 2, pp. 163-215, 2017.

[2] B. Berti, A. Esuli, and F. Sebastiani, "Unravelling interlanguage facts via explainable machine learning," Digital Scholarship in the Humanities, vol. 38, no. 3, pp. 953-977, 2023.

[3] B. Berti, A. Esuli, and F. Sebastiani, "Unravelling interlanguage facts via explainable machine learning," Digital Scholarship in the Humanities, vol. 38, no. 3, pp. 953-977, 2023.

[4] J. Geluso, "Grammatical and functional characteristics of preposition-based phrase frames in English argumentative essays by L1 English and Spanish speakers," J. Engl. Acad. Purp., vol. 55, p. 101072, 2022.

[5] S. Steinbakken and B. Gambäck, "Native-language identification with attention," in Proc. 17th Int. Conf. Natural Language Processing (ICON), Dec. 2020, pp. 261-271.

[6] L.J. Zhang and X. Cheng, "Examining the effects of comprehensive written corrective feedback on L2 EAP students' linguistic performance: A mixed-methods study,"J. English for Academic Purposes, vol. 54, p. 101043, 2021.

[7] S. Malmasi et al., "Native language identification: explorations and applications," Ph.D. dissertation, Dept. of Sci. and Eng., Macquarie Univ., Sydney, Australia, 2016.

[8] J. A. Linck, J. F. Kroll, and G. Sunderman, ”Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second-language learning,” Psychological Science, vol. 20, no. 12, pp. 1507-1515, 2009.

[9] J. Tetreault, D. Blanchard, and A. Cahill, ”A report on the first native language identification shared task,” in Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications, Jun. 2013, pp. 48-57.

[10] S. Latifi, S. Syed, and M. Gierl, ”Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing,” Language Testing, vol. 38, no. 1, pp. 62-85, 2021.

[11] B. Gyawali, G. Ramirez, and T. Solorio, ”Native language identification: a simple n-gram based approach,” in Proc. Eighth Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, GA, USA, Jun. 2013, pp. 224-231.

[12] J. F. Werker and R. C. Tees, ”Cross-language speech perception: Evidence for perceptual reorganization during the first year of life,” Infant Behavior and Development, vol. 7, no. 1, pp. 49-63, 1984.

[13] S. Crossley, et al., ”Frequency effects and second language lexical acquisition: Word types, word tokens, and word production,” Int. J. Corpus Linguistics, vol. 19, no. 3, pp. 301-332, 2014.

[14] S. A. Kotz, ”A critical review of ERP and fMRI evidence on L2 syntactic processing,” Brain and Language, vol. 109, no. 2-3, pp. 68-74, 2009.

[15] G. Kasper, ”Four perspectives on L2 pragmatic development,” Applied Linguistics, vol. 22, no. 4, pp. 502-530, 2001.

[16] Y. K. Muthusamy, E. Barnard, and R. A. Cole, ”Automatic language identification: A review/tutorial,” IEEE Signal Processing Magazine, vol. 11, no. 4, pp. 33-41, Oct. 1994.

[17] S. A. Crossley and D. S. McNamara, ”Computational assessment of lexical differences in L1 and L2 writing,” Journal of Second Language Writing, vol. 18, no. 2, pp. 119-135, 2009.

[18] L. Ortega, ”Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing,” Applied Linguistics, vol. 24, no. 4, pp. 492-518, 2003.

[19] UYSAL, Hacer Hande. Argumentation across L1 and L2 Writing: Exploring Cultural Influences and Transfer Issues. Vigo International Journal of Applied Linguistics, 2012, 9: 133-159.

[20] P. Durrant, ”Discipline and level specificity in university students’ written vocabulary,” Applied Linguistics, vol. 35, no. 3, pp. 328-356, 2014.

[21] Z. Du and K. Hashimoto, ”Decoding Academic Language: The Symbiotic Relationship Between Boosters, Hedges, and Voice in EAP,” in 2024 12th International Conference on Information and Education Technology (ICIET), 2024, pp. 46-52.

[22] Z. Du and K. Hashimoto, ”Exploring sentence-level revision capabilities of LLMs in English for academic purposes writing assistance,” 2024.

[23] M. Lui and P. Cook, ”Classifying English documents by national dialect,” in Proc. Australasian Lang. Technol. Assoc. Workshop 2013 (ALTA 2013), 2013, pp. 5-15.

[24] R. Zheng, J. Li, H. Chen, and Z. Huang, ”A framework for authorship identification of online messages: Writing-style features and classification techniques,” J. Am. Soc. Inf. Sci. Technol., vol. 57, no. 3, pp. 378-393, 2006.

[25] J. Howard and S. Ruder, ”Universal language model fine-tuning for text classification,” arXiv preprint arXiv:1801.06146, 2018.

[26] Z. Du and K. Hashimoto, ”Data augmentation for sentrev using back-translation of lexical bundles,” in Proc. 37th Pacific Asia Conf. Lang., Inf. Comput., Dec. 2023, pp. 70-79.

[27] Z. Du and K. Hashimoto, “Tcnaec: Advancing sentence-level revision evaluation through diverse non-native academic english insights,” IEEE Access, vol. 2023.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ”RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[29] A. Tharwat, ”Classification assessment methods,” Applied Computing and Informatics, vol. 17, no. 1, pp. 168-192, 2021.

[30] K. Stock and J. Yousaf, ”Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data,” International Journal of Geographical Information Science, vol. 32, no. 6, pp. 1087-1116, 2018.