國立臺灣師範大學理學院資訊工程研究所

碩士論文

Department of Computer Science and Information Engineering

College of Science

National Taiwan Normal University

Master's Thesis

探討提升自動英語口語評估準確性之方法 - 以會話測試為例

Exploring Methods to Enhance Accuracy in Automated Speaking Assessment- English Interview as a Case Study

李俊廷

Jiun-Ting Li

指導教授: 陳柏琳 博士

Advisor: Berlin Chen, Ph.D.

中華民國 113 年 5 月

May 2024

Acknowledgements

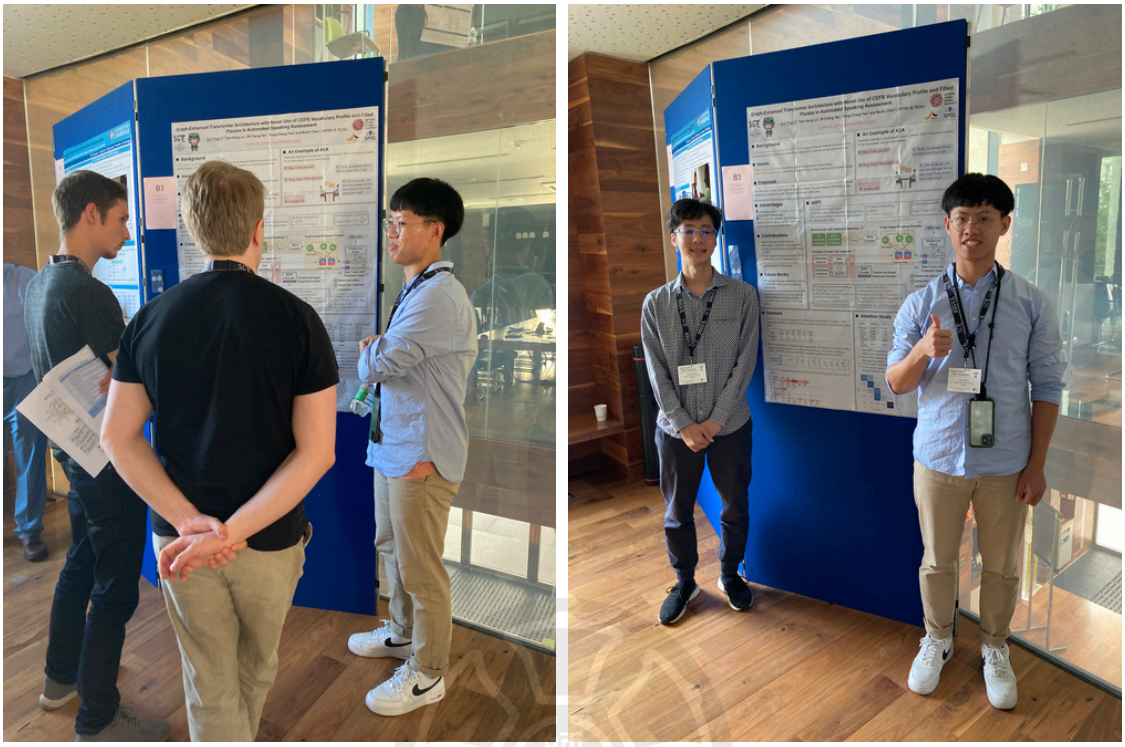
I want to express my sincere appreciation to the following individuals and organizations who have contributed to my journey:

Berlin Chen: My advisor, whose guidance has been invaluable in teaching me the intricacies of research. Berlin Chen always suggests we show our enthusiasm for what we do, encouraging us to submit publications to international conferences. I appreciate his teaching and guidance untiringly, and his support in making me participate in SLaTE 2023 and Interspeech 2023 at Ireland, Dublin.

My advisor always reminds us to keep improving our abilities and techniques. Thank you for your guidance. I feel the same after looking for a job recently. But I will never forget what the teacher mentioned about my strengths. Even if the other person doesn't want to hire me or something like that, I won't hold a grudge against the other person, because as a human being, I don't know which day I will meet him again, or I will have the opportunity to meet him again. Help him, it's just that the time has not come yet.

Table 1: The left is taken at the outside front of Interpseech 2023 Venue, at North Dock B in Ireland, Dublin. The right photo is taken on the first floor of the Department of Computer Science in the National Taiwan Normal University Branch Gongguan.

Table 2: Both the photos are taken at Trinity College, Dublin in the workshop of SLaTE 2023. The two people in the left photos are the other PhD students in the CS department. The one on the right side of the photo is Tien-Hong Lo.



顏必成 (Yan, Bi-Cheng) and 羅天宏 (Lo, Tien-Hong): My Ph.D. elder classmates, whose assistance paved the way for my research endeavors. Bi-Cheng Yan is a straight man. What he said was very straightforward and he did not shy away from anything he said. Sometimes I may find other people being mad on what he said, but these words help me understand whether what I have done in my research is wrong or inappropriate. Tien-Hong Lo is very good at observing others and giving explanations and teachings that are appropriate and easy to understand. Every time I discuss research with him, I receive usable knowledge or information to help me with my research. Furthermore, I thank him for suggesting this book: 底層邏輯：看清這個世界的底牌，劉潤. It is one of the best books ever I read before. I hope that we can get in touch after leaving this laboratory.

劉憶年 (Liu, Yi-Nian)'s mother: For generously providing delicious lunches almost every week, and fueling me through long study sessions. This lunch was originally for my advisor, Berlin Chen. However, my advisor always says that he wants to put more effort into research, so he eats less lunch or he may fall asleep. But I always share other things inside the lunch bags with the members in the laboratory, such as bananas, cookies, oranges, etc.

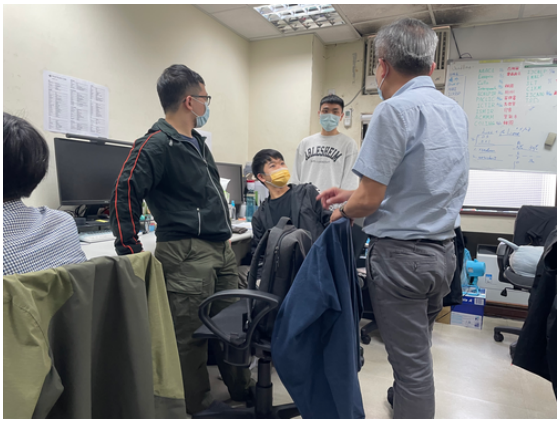
Table 3: The photos from left to right side are the conferences of TAAI 2021, Interspeech 2023, and ASRU 2023, respectively. Notice that the fee for participation in ASRU 2023 is afforded by myself.

My classmates: 馨偉, 又升 (仙草↑～), 沁穎, 宛庭, 詩彥, and the younger generation: 鄭皓天, 蔡孟庭, 王詣承, 何冠勳, 楊子霆, 吳姿儀, 游恩倫, 林孟欣, 彭玟瑄, 楊憶婷, 白立亭, 盧泇良, 王建鈞, 李佩穎. Your camaraderie and support have made this journey memorable. Every day, I feel energized by the joyful atmosphere in the laboratory, which adds meaning to our research efforts.

曹又升 (Tsao, Yu-Sheng), his name translated into Taiwanese sounds like SamTsao, 仙草, a traditional jelly cuisine in Taiwan. He focuses on the ability of hard code implementation and likes to build up systems. His seat is beside me in the laboratory.

There are some **solemn warnings** I want to inform them: 王馨偉 (Wang, Hsin-Wei)

Table 4: Some events in the laboratory: the left is the cleaning up in the laboratory, and the right side is the discussion for our progress.

and 楊子霆 (Yang, Tzu-Ting). I know you both contribute to the laboratory a lot, even more than the PhD students. But there is something you should not do as follows: (1)

Occupying the best GPU cards in this laboratory for many weeks. I do not know what kind of experiments you conduct, it looks like ASR-related experiments, but your behavior influences the efficacy of other members' works. What's worse is that when someone else wants to run an experiment with this card, the two of you will deliberately kill the other person's experiment and then continue to run your experiment. You should stop such a selfish act. (2) Copy other member's experimental results or code without their permission. It is certain that sharing their ideas or experimental results with members of the laboratory promotes the development of techniques and makes it possible for members to bear new ideas in their research. However, you two, especially 馨偉, copy other people's codebase without permission or acknowledge them, and use the results in your report. Some members of the laboratory already complain about this issue, 玟瑄, 必成, 天宏, 詣承. I claim again that you can propose to cooperate with other people or ask them if possible to share their work. But you should not just duplicate their hard work without their permission. It is very BAD.

宥勤 (Previous Name: 憶庭 (Yang, Yi-Ting)): For suggesting interview opportunities that broadened my horizons. She works hard to try every method possible to solve specific problems, such as her research on the keyword spotting field. Even if she is not good at coding, or cannot address the problem she met, but she do not give up. I hope she can keep that in her mind.

Table 5: All the photos are Shih-Hsuan Chiu.

邱世弦 (Chiu, Shih-Hsuan) and the elder master students: A big shoutout for your kind support in my master, 邱↑邱↓邱↓邱↓邱↑, for their guidance and support throughout

my time in the laboratory, alongside heartfelt thanks to 趙福安 (Chao Fu-An), 范姜紹瑋 (Fan-Jiang, Shao-Wei), 林筱芸 (Lin, Hsiao-Yun), 鄭宇森 (Cheng, Yu-Sen), 林韋廷 (Lin Wei-Ting) for leading me into this laboratory.

Shih-Hsuan always makes funny things in the laboratory, every day he comes to the laboratory, it is sure something hilarious happens on that day. Typically, I call his name like 邱-↑-↓. And he definitely asks me to shut up. THAT WAS FUNNY. Or we deliberately misheard what he said, and he would say: fuck you. One time, his beverage was left a little bit, and I drank all the remain. After he come back, he kept sipping the drink, but when he noticed that the drink was gone, he realized that I had drunk it all. Then interesting conversations happened. In conclusion, he is a funny guy and argues to do something well.

Table 6: They love to play basketball after the whole day on doing their research.

世弦, 福安, 范姜, and 韋廷 are brothers who love to play basketball or work out at the gym.

EZAI Corp.: Providing opportunities to represent my ideas, and the delicious lunch boxes to fill my belly. I want to thank Mic, Nick, and other colleagues in EZAI for concentrating on our representation and allowing us to know the possible path to keep our research going

without deviating from realistic situations.

陳浩然 (Howard Hao-Jan Chen) Professor: He is the professor who supports my research with the foundation. I will try my best to improve the pronunciation system.

Dr. Kate Knill and the ALTA Institute, Cambridge University: For their invaluable contribution in providing SST to CEFR converting scores in the NICT JLE corpus. Nice to meet you in person in Ireland!

Netherlands Dwarf Bunnies: MUGI and MIMI: I would like to say research needs creative ideas. Sometimes I feel that I don't have the energy to do research. But the adorable bunnies (2 ぴょん) always heal my heart and give me the energy to keep going on what I want to do in my research. https://www.youtube.com/@bunny_mugi_channel Everyone please subscribe to their YouTube channel!

Beatbox music: I want to show my appreciation to Gene Shinozoki and Chris Celiz (Spiderhorse) for producing such great work in beatboxing. I like Daily Beats #76 | Tell you Something https://www.youtube.com/watch?v=E3JbBA7GZvk when doing my work. I almost listen to it every day! It is nice to meet you guys in person at Grand Beatbox

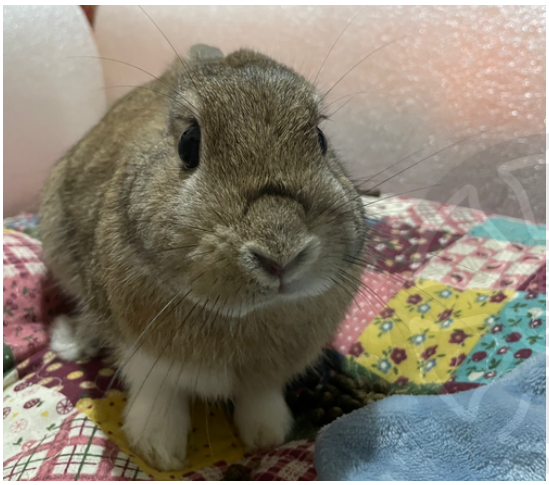
Table 7: The photo with Spiderhorse (left side: Gene Shinozaki, and right side: Chris Celiz). It is taken in the venue of Grand Beatbox Battle 2023 at EX-THEATER ROPPONGI.

Battle 2023 in Tokyo, Japan, and take the photo together.

My mother: MAMIMAMĨ. I love you. Thank you so much for supporting my decision

Table 8: The photo with my other friends.






to enroll in NTNU, CSIE for my master's degree.

Other guys: There are so many other accompany in my academic journey for my master's degree. I can not itemize you all here, but I appreciate the time we can meet each other.

Your support and encouragement have played a pivotal role in my academic journey. I am truly grateful.

摘要

由於全球化與網路的普及，人們需要學習第二語言的需求急劇增加，尤其是英文作為最主要的知識傳遞語言。雖然現今有許多免費或付費的英文教學影片、補習班等資源可供選擇，然而語言教師的增加速度卻跟不上學習者的需求。因此，為了解決此問題，我們需要有效率的方式處理學習者在語言學習過程中獲得的資訊，協助非母語者在沒有足夠語言教師的情況下，仍能順利地學習第二語言。在各種補足人力的方法中，電腦作為人力輔助的角色最為適合，尤其是語音辨識技術已經成熟，並出現許多商業應用案例，如電腦輔助語言學習 (Computer Assisted Language Learning, CALL) 的錯誤發音偵測與診斷 (Mispronunciation Detection and Diagnosis, MDD)、可讀性評量，以及我們本研究的主題：自動口說評量。自動口說評量是英文評量中的一個方面，透過受訪者的口說聲音和內容來進行能力評估，但需要英文專家花費時間進行評分。如果可以藉由電腦完成相同任務，將節省大量的人力、時間和金錢。然而，目前在此領域的研究遇到幾個問題，例如不同等級的語者數量不平衡，尤其是在最高和最低等級的語者數量和其他等級之間呈倍數差距，以及自由口說容需要考慮更細緻的子句關係代名詞關係和面試官的資訊。我們嘗試從資料、訓練技巧和模型架構等方面入手，提升整體效能，同時兼顧可解釋性，使本研究能夠真正在實際應用中被接受。模型的程式碼在 https://github.com/a2d8a4v/HierarchicalContextASA/、資料前處理的程式碼在 https://github.com/a2d8a4v/local_for_nict_jle。

關鍵字： Automated Speaking Assessment、Bidirectional Encoder Representations from Transformers、Graph Neural Network、Spoken Response Coherence

Abstract

Due to globalization and the prevalence of the Internet, there has been a sharp increase in the demand for second language learning, especially in English, which is the primary language of knowledge transfer. While there are many free or paid resources such as English tutorial videos and cram schools available today, the rate of increase in language teachers cannot keep up with the demand of learners. Therefore, to address this problem, we need an efficient way to process the information acquired by learners in the language learning process, to assist non-native speakers in successfully learning a second language without sufficient language teachers. Among various methods to supplement manpower, the computer plays the most suitable role as a human assistant, especially since speech recognition technology has matured and many commercial applications have emerged, such as Mispronunciation Detection and Diagnosis (MDD) in Computer Assisted Language Learning (CALL), readability assessment, and the topic of our research: automatic speaking assessment. Automatic speaking assessment is an aspect of English assessment that evaluates the ability of respondents through their oral speech and content, but requires English experts to spend time grading. If the same task can be completed by a computer, it will save a lot of manpower, time, and money. However, current research in this field has encountered several problems, such as the imbalance of the number of speakers in different levels, especially the multiple differences

in the number of speakers between the highest and lowest levels and other levels, and the need to consider more detailed clause relationships, pronoun relationships, and interviewer information in free speaking content. We attempted to improve the overall performance of our research from the aspects of data, training techniques, and model architecture, while also considering interpretability so that our research can be truly accepted in practical applications. The model's implementation code is available at https://github.com/a2d8a4v/HierarchicalContextASA/, and the code for the data preprocessing stage is at https://github.com/a2d8a4v/local_for_nict_jle.

Keywords: Automated Speaking Assessment, Bidirectional Encoder Representations from Transformers, Graph Neural Network, Spoken Response Coherence

Contents

Chapter 1 Introduction

1.1 Research Motivation

Today, in our globalized society, English dominates as the primary language for communication and information exchange [40]. Such a situation allows instant access to the latest knowledge, making English valuable to learn. Consequently, the population learning English as a second language is explosively increasing. However, within the constraints of limited learning resources, teaching resources for listening, speaking, reading, and writing might not adequately meet everyone's needs. Additionally, the human resources for language teaching cannot fulfill all individuals' requirements for efficient and accurate language learning. Presently, in Asian countries such as South Korea [106], Japan [92], Taiwan [42, 57, 70], and others, there is a significant push for second language education [35, 96], primarily centered on English. Numerous supplementary classes focusing on English examinations or oral practice are prevalent in these countries. It's challenging to tailor instruction according to each student's learning needs, whether within formal schooling or supplementary classes, due to the larger number of students compared to teachers. Consequently, this leads to diminished effectiveness in English language learning and might even erode students' confidence in learning English.

With the advancement of technology, English language learning has been facilitated through devices such as computers, tablets, or smartphones. These tools enable individuals to utilize their fragmented time for language learning, breaking free from the constraints imposed by schools or after-school classes [84]. Moreover, for schools or supplementary classes, such technology aids in real-time assessment of students' English abilities, allowing for tailored instructional content delivery. It even enables computers to provide students with learning recommendations, thereby freeing up more time for teach-

ers in schools or supplementary classes to provide students with the essential assistance they require.

Technological applications in language learning, such as automatic speech recognition (ASR) and automated speech assessment [7] (ASA), have matured significantly. Related products like English conversational learning robots offer students suggestions on grammar, vocabulary, and pronunciation, aligning more closely with native speaker usage. They can provide timely feedback after person-to-person English interviews and aid in reviewing and expanding English assessments. This paper focuses on evaluating human-to-human English tests. I am particularly interested in assessing spoken content, which presents certain challenges distinct from written content, and providing precise overall English assessment results. There's still considerable room for improvement in terms of current performance. Through analyzing the speaker spoken content and voice information, this study aims to develop more ideal modeling methods to provide more accurate feedback.

1.2 Mission Description

1.2.1 Introduction and Problem Statement

The population of English-as-a-second-language (ESL) learners has increased worldwide due to the accumulative significance of the English language in many fields, including computer science [2, 55]. However, it is acknowledged that people retain respective orientated pronunciation practices or grammar conventions in their mother languages, which is called interlanguage, leading to wrong delivery to others.

In language learning, there are four main aspects: listening, speaking, reading, and writing, serving as skills and approaches to learning a language. It's common for Taiwanese individuals to habitually 'watch' television or movies with subtitles, reading textbooks while studying. However, this habit often leads many Taiwanese students to believe that learning vocabulary merely involves looking at subtitles to grasp pronunciation. Additionally, they continuously learn English vocabulary and phrases by copying them, which in the long run results in Taiwanese students being proficient in the aspects of reading

and writing in language learning but struggling to express their own thoughts effectively in English. To rectify the direction of language learning and understand if students can naturally express themselves in English communication, this study primarily focuses on 'speaking'. It involves researching and evaluating conversational English content. To assign scores to the conversation content, ASA studies aim to address this issue by evaluating the speaking abilities of ESL learners using widely recognized scoring frameworks, such as the Common European Framework of Reference (CEFR) [71], which has been implemented in Dutch [97, 98] and French [38], and providing them with understandable feedback.

1.2.2 Language Assessment

The concept of language proficiency, while not definitively established, is guided by certain standards that assess the capabilities of non-native speakers. A notable example is the Common European Framework of Reference for Languages (CEFR) [71], widely used in examinations, teaching, and classification studies. This framework outlines criteria for determining the proficiency levels of language test-takers. However, the interpretation of language proficiency varies among different entities, leading to diverse assessments.

Language assessment serves examination purposes and provides learners with valuable feedback, enabling them to adjust their learning approach. By integrating various perspectives, a more objective assessment can be achieved. Also, language assessment becomes a vital part of teaching techniques or means of language acquisition [44].

In recent years, ASA systems have gained prominence, particularly speech recognition-based ones. These systems employ frameworks like the one described by [65] for grammar assessment and the Vocabulary Profile for vocabulary estimation [1]. Despite these advancements, challenges remain, such as the contextual interpretation of spoken content and the creative aspect of language use, which is often overlooked in automated assessments [9, 84]. Higher-level language learners typically exhibit a precise expression of content, complex vocabulary, and sophisticated grammar structures. In contrast, novice learners often use simpler structures and vocabulary, and may demonstrate errors influenced by their native language. This dichotomy highlights the nuanced nature of language

[1] https://www.englishprofile.org/

proficiency assessment.

In conclusion, this paper explores the multifaceted nature of language proficiency assessment, focusing on the advancements and challenges in automated scoring systems.

1.2.3 Technologies in Automated Speaking Assessment

The field of language assessment has evolved over decades, with significant contributions from research institutes and tech companies. Automated scoring (AS) has emerged as a solution to the inefficiencies of traditional language learning assessments [100]. AS includes automatic essay scoring (AES) [33, 82, 83, 89, 109, 111, 115, 117, 121] and ASA [7]. AES focuses on grading written content by evaluating structure, vocabulary, grammar, and sentence construction, benefiting from advancements in language models. ASA, however, aims to assess spoken language proficiency, including content, pronunciation, grammatical errors, and cohesion. ASA faces unique challenges, such as data collection and annotation, the inherent disfluency in spontaneous speech, and often leverages transformation from written to spoken tasks [47, 48].

The exploration of features in Automated Scoring (AS) plays a critical role in shaping the design of the system. Current methods in automated spoken assessment (ASA) predominantly utilize linguistic and acoustic features sourced from automatic speech recognition (ASR) technologies. These features, while not exhaustive, include acoustic elements, recognized phonemes and words, and detailed time-aligned data. They are closely associated with various proficiency areas such as fluency, pronunciation, prosody, and text complexity [17, 31, 119, 120]. These features are then integrated into specialized analytic modules for grading different language aspects. Nevertheless, manually crafted features might miss some critical information pertinent to ASA. To address this, a bi-directional long short-term memory (BLSTM) based prompt-aware encoder with an attention mechanism has been introduced. This encoder focuses on capturing both an individual's response and its context, relying on text prompts to assess speaking proficiency against a standardized scoring system [79, 80]. While these methods are adept at identifying long-range correlations, they may face challenges with lengthier text [50]. Alternative approaches involve using Transformer models [107], which employ multi-head self-attention for more flexible word associations. Research in [22] indicates that Transformer-based models sur-

pass LSTM-based models in evaluating monologic responses. Furthermore, a study in [85] applies a Transformer-based language model to assess the content quality of speaking proficiency in online interactions with non-native speakers.

1.2.4 Methodology Overview

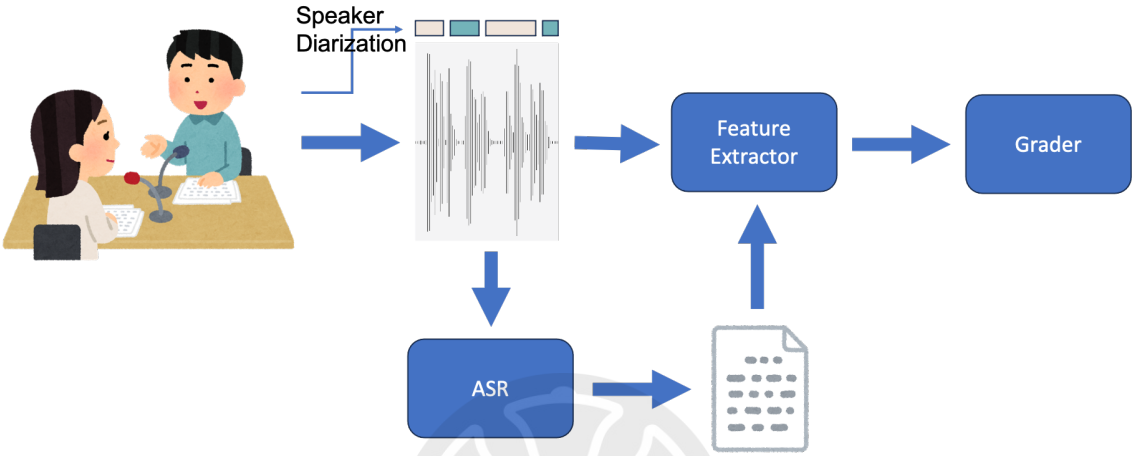


Figure 1.1: The overall scheme of doing automated speech assessment in a traditional way. It typically obtains, for example but not limit to, acoustic time-aligned features and n-gram features for the final grading.

Before providing students with detailed error feedback, the system should initially offer an overall assessment of the student's proficiency level in that particular skill. This paper defines research on conversational speaking assessments involving an interviewer and an interviewee. They engage in a pre-designed interview based on evaluations and language education theories, encompassing various scenarios of daily life. It is aimed to conduct automated speaking assessments and provide feedback based on the textual and audio information from the interviewee's spoken content. As illustrated in Figure 1.1 Typically, a single-channel microphone records all the spoken content of the assessment. Subsequently, I employ speaker diarization, a technology for automatic speaker identification, to segment the audio files. The entire conversation's audio file is then segmented into portions for each speaker, labeled either the candidate or the interviewer. These segments undergo speech recognition to obtain their textual content. I assess the interviewee's English speaking ability using this text and audio information obtained from speech recognition.

The focus of this paper lies in the textual information assessment. I consider this examination as a complete conversation, not just the monologue of one party. In real-

ity, the interviewer and the interviewee engage in a role-playing scenario for the spoken assessment, which includes interaction from both sides. This constitutes a conversation. After extracting text and audio information separately, it predicts the interviewee's speaking ability based on these two major features. However, I have the constraint which only provides text transcriptions in their publicly available corpus, restricting the direction of my research. To conquer this issue, this paper adopts another framework, as illustrated in Figure 1.2, to process our research and experiments.

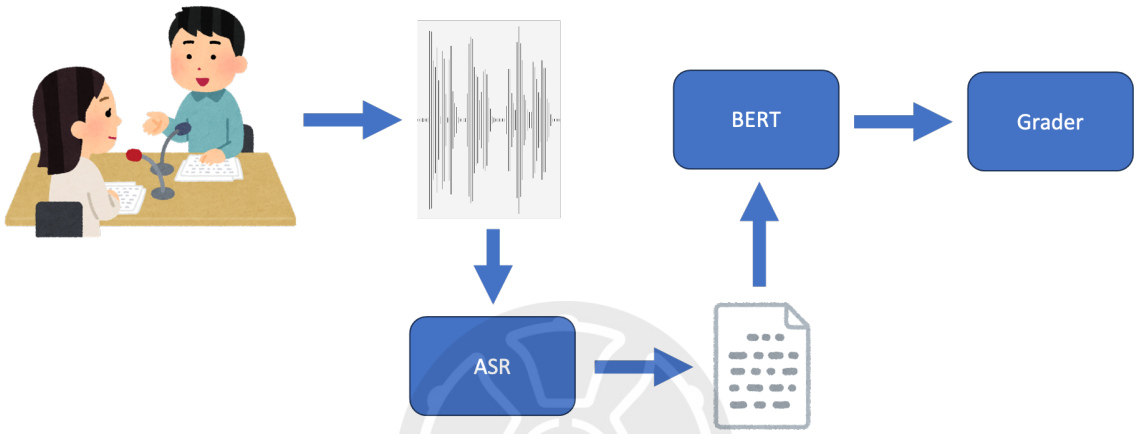


Figure 1.2: The scheme of grader under a text-only constraint. The whole conversation is first processed via an automatic speech recognizer to get the decoded text. Then the decoded text is conveyed to the BERT-based language foundation model to obtain the hidden states for the down-streaming grader model.

Furthermore, the proposed method meets the demand for education providers who only access text results converted from users' audio recordings due to privacy issues. The content-scoring engine used sparse features, such as unigrams and bi-grams. It was trained on a large set of responses, which is not inferior in performance comparable to that based on fluency and pronunciation [61]. In [22], text transcripts generated by an off-the-shelf ASR system are used to predict overall spoken language proficiency. Similar methods in conjunction with natural language processing techniques have been applied with success in analogous fields, such as automated essay grading (AEG) [86, 105].

Despite the success of these neural methods in capturing syntactic structures, they are faced with challenges in the lack of consideration to cooperate with the existing structured dialogue information, such as the transmission in the course of conversation, and the speaker's intent inside each response. To fill this gap, I propose to effectively integrate hierarchical context inside the original conversation grader, where I adopt heterogeneous

graph neural networks (HGNNs) that provide a promising solution by leveraging relay nodes to transmit information between super nodes, enabling capturing both local and global hierarchical context [19]. HGNNs also can extract fine-grained details and specific meanings from content and words. Notably, [33] is the first work that employed HGNNs to evaluate the English written ability of test-takers. As a remedy, I suggest alleviating the issue by constructing the sentence nodes as a proxy interacting with word nodes in heterogeneous graphs, stepwise refining the relationship in interview content.

Additionally, while a multi-modal approach to grading proficiency shows promise [85], the challenge of obtaining necessary audio data remains, particularly due to privacy concerns that restrict access to such data outside academic settings. To mitigate this, focusing on text-based features—whether human-annotated or generated by ASR becomes crucial for evaluation purposes.

To our best knowledge, research has yet to address the ASA problem by utilizing HGNNs. In view of this, I propose a novel approach: graph-enhanced response encoder based on transformer architecture (GEPT) for ASA, so as to improve spoken language proficiency assessment based on a graph attention network (GAT) [101] to explicitly model the dynamic interaction between responses in a meeting with conditional response information, building upon the existing Transformer-based model.

1.2.4.1 Coherence Modeling

In addressing the challenge of assessing automated spoken assessment (ASA) within conversational contexts, it becomes apparent that existing methodologies often fall short in capturing the coherence between utterances. Previous research primarily relies on single sentences as input for foundational models, disregarding the interconnectedness and transitions inherent in conversation [30, 116, 121]. However, this oversight underscores the necessity of considering the flow of spoken content and extracting paragraph-level embeddings from both intra-sentence and inter-sentence tokens. Recent efforts, such as employing enhanced graph-based encoders [56], aim to model the relationship between interlocutors and candidate responses, facilitating a more comprehensive evaluation of proficiency.

The challenges of extending coherence modeling across conversational turns reflect the complexity inherent in communication dynamics. Coherence plays a crucial role in facilitating logical transmission during conversations, enabling language learners to negotiate meaning with interlocutors or interlocutors to solicit optimal responses from candidates. This concept operates across two levels: at the macro level, intentional structures, characterized by Dialogue Acts [3, 108], serve as transition markers within conversations, while at the micro level, finer-grained knowledge units within responses illustrate actual actions. Both levels are identified as vital elements in conversations [14], revealing a gap in their exploration within interactive assessments. Consequently, there is a need to delve deeper into understanding and integrating these elements within assessment frameworks to ensure a more nuanced evaluation of conversational proficiency.

1.2.4.2 Word CEFR Ranking Integration and Disfluencies

In this paper, the word CEFR ranking integration is introduced, which utilizes the vocabulary profile to tag the word CEFR ranking, to the best of our knowledge, this is the first time used as a feature in the ASA task. However, the word of the CEFR ranking information should be a consistent representation under an assumption of normal distribution instead of dynamically changing during training time. As a remedy, I designed a new CEFR node connected to the word nodes to make word nodes encapsulate CEFR information from them. Then the embedding integrates with word CEFR ranking information is leveraged with the foundation model.

IR	please	very	short	one				
IE	can	i	use	she	or	i		
VP	A1	A1	A1	A1	A1	A1		
IR	she	uh	she	well				
IE	oh	cough						
VP	X	X						
IR	whatever	you	like					
IE	so	um	now	it	was	a	dinner	time ...
VP	A2	X	A1	A1	A1	A1	A1	A1

Figure 1.3: Word position focusing in attention layer compared to the CEFR ranking of words

Spoken language assessments are typically conducted on spontaneous speech, in which disfluencies, such as hesitation 'um' shown in Figure 1.3, are common in spoken

language [58], which does not exist in AEG. In addition, previous studies have proved that incorporating incremental disfluencies can reduce learner errors' impact on model precision [93]. Separately, LMs prioritize lexical words over disfluencies and ignore the word difficulties tagged in Figure 1.3, which is a crucial proxy for evaluating spoken language proficiency and incorporating it as a supplement to assist in grading content [33]. In our proposed structure, I also incorporate information on the filled pauses (FP) and vocabulary scale relationships to improve the render of the relationships between CEFR ratings, hesitations, and the content of the verbal response.

1.3 Contributions

The contributions in this paper are like the following items:

Advancement in Automated Speaking Assessment. For the previous advancements, few works were investigated in the conversation test in ASA. This thesis contributes to the evolution of ASA methodologies in conversation tests by integrating deep learning techniques, specifically Transformer-based models and GNNs. GNNs serve as structural knowledge models, storing rich factual information in a way that enhances the foundation models. In contrast, the foundation models excel in generating unseen facts. The exploration and implementation of these advanced frameworks showcase their potential to capture nuanced linguistic cues and contextual information in spoken language evaluations.

Addressing Coherence in Conversation Assessments. An essential contribution lies in addressing the challenge of coherence modeling within conversation assessments. The incorporation of hierarchical attention mechanisms and discourse relations within Graph Neural Networks (GNNs) aims to enhance the evaluation of spoken proficiency by capturing holistic contextual information.

Integration of Common European Framework of Reference (CEFR) Rankings. This research pioneers the incorporation of CEFR ranking information as a consistent feature in ASA methodologies. Ensuring the uniform representation of proficiency levels throughout training enables a more standardized and accurate evaluation of speaking proficiency.

1.4 Structure of the Thesis

This thesis is organized into four main chapters, following the introductory Chapter 1. The initial chapter sets the stage, outlining the research motivations and providing an overview of related evaluation studies. The following of this paper is divided into four sections. Related works in Chapter 2 conduct a literature review, discussing past approaches in ASA, foundation model, heterogeneous graph model learning, and structural dialogue. In Methodology at Chapter 3, I give a clear explanation of the proposed approach to modeling the hierarchical context and coherence. Following, Experimental Results in Chapter 4 and Dataset section gives an introduction of corpus and the analysis of dataset. The latest, Conclusions and Future Works conclude the proposed methods and the efficacy, also extend the possibilities of investigating coherence. And the related topics also need attentions to process on.

Chapter 2 Related Work

2.1 Automated Speaking Assessment (ASA)

From the theoretical aspect, [54] is the pioneer who established the importance of the field of L2 assessment, who thought that the challenges in learning languages can be anticipated by comparing the learners' native language and their target language. In which a set of distinct elements, such as sounds, grammar, and vocabulary, are analyzed from his perspective of contrastive analysis and structuralist perspective of language. Later, [21] built upon these findings, establishing the groundwork for error analysis and approaching the idea of error from a developmental angle. The later research by [77] covers a language user's grasp of grammar, syntax, morphology, phonology, and similar elements. However, it redefines this knowledge as a practical, socially-oriented understanding of when and how to employ speech effectively. [13] subsequently formalized the conveyance of meaning, comprising grammatical understanding, sociolinguistic adeptness, and strategic competence. It then developed and established as [71].

In practical methods to implement the ASA system, the iconic approaches to ASA of language proficiency typically employ linguistic and acoustic features derived from automatic speech recognition (ASR) systems. These features include, but are not limited to, acoustic features, recognized phones and words, and multi-granular time-aligned information, which is intricately linked to distinct proficiency domains, such as fluency, pronunciation, prosody, and text complexity [17, 31, 119, 120].

In turn, the obtained linguistic and acoustic cues are channeled into the corresponding analytic graders of different aspects. However, these hand-designed features may inevitably leave out some salient information relevant to ASA. As a remedy for such an

issue, deep-learning (DL)-methods recently have become a mainstream mode for ASA of language proficiency to address the problem that hand-crafted features inevitably leave out some salient information relevant to ASA [79] and are typically tuned for a specific task or domain [18, 80]. For instance, a prompt-aware encoder is proposed, which is based on bi-directional long short-term memory (BLSTM) and an attention mechanism to capture response and its contextualized information, thereby conditioning on text prompts (stimulus questions) to evaluate speaking proficiency according to a predefined scoring rubric [79, 80]. Those methods excel at capturing long-range dependencies but may struggle with longer content [50]. Transformer-based [107] language model (LM) is also adopted in other recent work, which is a foundation model trained on broad data good at capturing contextualized information. Other similar attempts also used Transformers [107] with multi-head self-attention to connect words more flexibly. In [22], the Transformer-based models have been shown to outperform LSTM-based models in evaluating monologic responses. In [85], a Transformer-based LM is applied to assess the content aspect of speaking proficiency in online dialogues with non-native speakers.

In summary, as shown in Figure 2.1, there are four main categories of systems commonly used for grading proficiency [7, 62]. The first relies on traditional hand-crafted features like n-grams, time-aligned information, and various aspects like fluency, pronunciation, prosody, and text complexity [17, 31, 39, 119, 120]. The latter system utilizes BERT-based grading, using transcribed spoken text and Natural Language Processing tags like part-of-speech, disfluencies, and grammar errors identified post-transcription [5, 22, 23, 93]. A third method emerged, leveraging self-supervised learning representations like Wav2vec2.0 or HuBERT. These methods tackle issues like limited audio data, lack of crucial speech-related proficiency information, and error propagation due to Automatic Speech Recognition (ASR) [1, 4, 6, 15, 41, 51]. Recently, due to the success of developing a large language model (LLM), the latest method of utilizing LLM to access spoken content is considered but still undergoing research and development [27, 62].

2.2 Foundation Model

Generative foundation models [10] are a type of large-scale model designed to serve as a fundamental basis or foundation for various tasks. These models are typically pre-trained

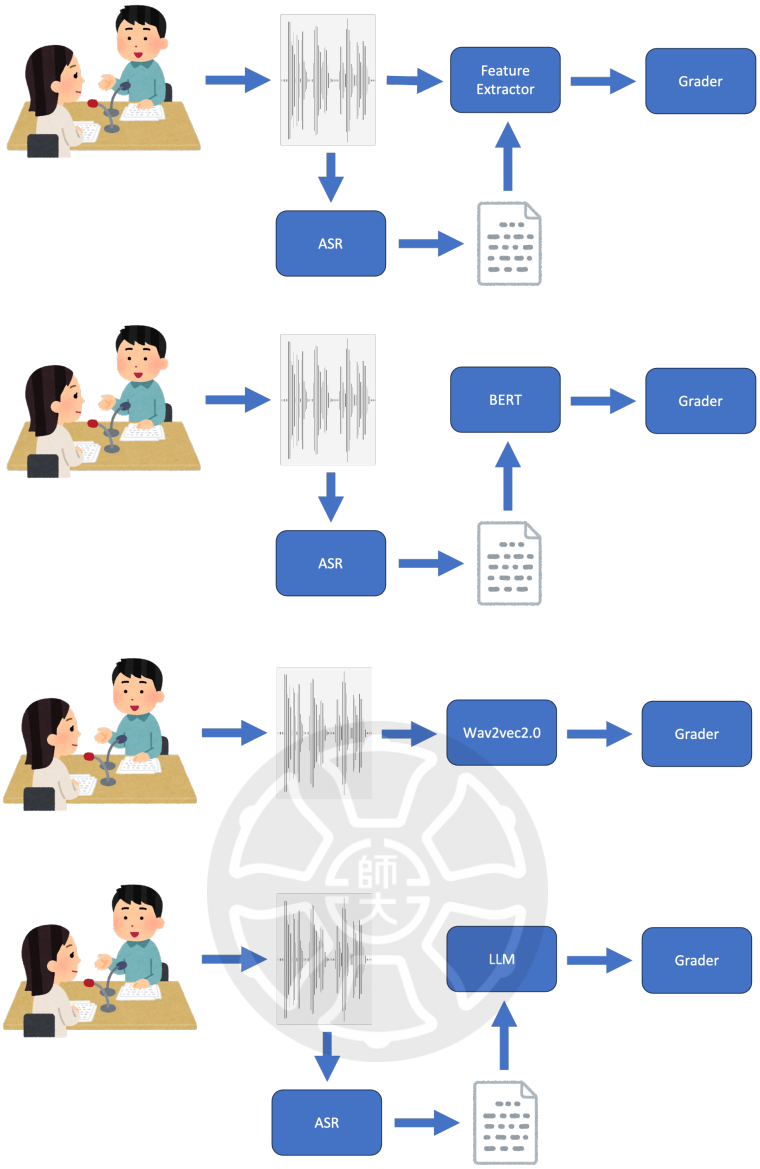


Figure 2.1: Four types of scheme for the grader. The first scheme is a traditional measure that utilizes pre-defined proficiency salient features such as acoustic and linguistic features. The second one, which only utilized the contextual features, ignore most of the acoustic features such as prosody or pronunciation, while the latter one only use acoustic features with self-supervised learning representation. The latest one, utilizes the high performance and capability of LLM to give direct feedback to the users.

on vast amounts of data, taking large language models as examples, often involving diverse sources from the internet, books, articles, and more. The goal is to impart a broad understanding of language and its nuances, enabling these models to perform well across multiple NLP tasks without extensive task-specific fine-tuning.

Some well-known examples of foundation models include the generative pre-trained Transformer (GPT)series by OpenAI [81], such as GPT-3, and Google's BERT (bidirec-

tional encoder representations from transformers) [23]. Foundation models have some characteristics that reason for exploiting them, especially since I am concerned about Transfer Learning and Fine-Tuning Adaptability which is related to the ASA topic. Due to their pre-training on diverse data, foundation models can transfer knowledge from the pre-training phase to downstream tasks. This knowledge transfer helps solve specific tasks more effectively, even with limited task-specific data. While pre-trained, these models are often fine-tuned on smaller, task-specific datasets to adapt their knowledge to the intricacies of a particular problem. This fine-tuning process allows them to perform exceptionally well on various tasks. These models have significantly advanced the field of NLP by providing a versatile starting point for a multitude of applications, enabling researchers and developers to achieve impressive results across diverse language-related tasks. Many works utilize the foundation model to AS, such as AES[72], or automated speaking assessment [67].

In this work, I focus on the text aspects to exploit the foundation model to obtain the contextualized information from the spoken content. Otherwise, I designed several experiments to improve the foundation model on grading as the baseline line method compared to the other hierarchical modeling methods.

2.3 Heterogeneous Graph-based Learning

Recently, a promising technique focusing on heterogeneous graph network representation learning has achieved momentous success [25, 114]. As one of the pioneer works, graph neural network (GNN) [87] was introduced to learn node representations by encapsulating neighbors' information via recurrent neural architecture. Following the graph spectral theory, there is a surge of generalizing graph convolutional networks (GCNs) has emerged and demonstrated superior learning performance by designing different graph convolutional layers. Notably, spectral CNN [11] pioneers the convolution operation in the spectral domain for network representation learning. Subsequent research has built upon this foundation, graph convolutional network (GCN) [53], for example, designs with a localized first-order approximation of spectral graph convolutions. GraphSAGE [37] was then introduced, which is practical in that samples features from a fixed size of the node' s local neighborhood instead of aggregating the whole graph in GCN. Similarly, graph

attention network (GAT) [101] incorporates trainable attention weights to learn the fine-grained importance of neighbors when aggregating neighborhood information of a node. Other recent advances, such as relational graph convolutional networks (R-GCN) [88], graph attention networks (HAN) [104], heterogeneous graph attention networks (HAN) [104], meta-path aggregated graph neural network (MAGNN) [32], and relational graph attention networks (R-GAT) [12], also follow a similar view.

Generally, HGNNs follow the idea of message passing across different graph layers by comprising information propagation and aggregation, capturing the rich neighborhood contextual signals for heterogeneous graph learning. In this work, I propose an enhanced HGNN framework that treats structural dialogic relations and the responses in dialogues as nodes due to the unknown influence between both of them. After constructing the heterogeneous graphs, I then utilize the GAT algorithm in each graph neural network layer to operate message passing. Moreover, I apply a bottom-up encoding process in our framework, where the sentence representations from the pre-trained foundation model pass through the respective sentence-level discourse graph to introduce the discourse relations, and the integrated representations go through the sentence-level structure graph to enhance the dialogical structural information.

In practical scenarios, diverse nodes and connections are commonly found in real-world applications, forming what is known as heterogeneous graphs. Heterogeneous network representation learning [25, 114] is a type of representation learning that aims to leverage relay nodes to transmit rich information between super nodes, enabling preserving relation heterogeneity in global and local hierarchical context. HGNNs offer promising techniques to attain this objective, delivering cutting-edge representation results, and have achieved unprecedented success in multiple natural language processing tasks, such as automatic essay scoring [33, 112], and document summarization [76, 102, 122]. Moreover, due to the inherent sparsity nature of structure leading to the flexibility and capability to fit in any long-length inputs and capture the hierarchical information, and the low memory usage compared to the quadratic self-attention mechanism.

I prepare an enhanced graphic scheme of our framework, in which a content encoding module contains a contextualized encoder that captures the contextualized content and an enhanced graph-based hierarchical encoder that portrays the hierarchical content among

responses [56], and a graph-based sentence-level discourse module that model dialogic relations in conversation tests. The discourse is retrieved from a state-of-the-art (SoTA) structured dialogic parser [18], where one relation connects from one response to another one with a relation type. Subsequently, a regressor computes each random pair from the above-computed embeddings and fuses them to predict the final proficiency score.

2.4 Structural Dialogue

Structural dialogue refers to dialogues' systematic analysis and parsing, considering their unique format and dynamics. Unlike traditional text, dialogues often involve multiple participants and can contain overlapping speech, interruptions, and non-linear discourse. Structural dialogue analysis aims to understand and represent these complex interactions. This is particularly important in situations like multi-party meetings, customer service interactions, and automated dialogue systems, where understanding the structure and flow of conversation is crucial for accurate interpretation, response generation, or summarization. The goal is to capture the essence of conversations, including the roles of different speakers, their interactions, and the progression of topics [64].

Initial research in discourse parsing primarily used rhetorical structure theory [63] for written texts. This method was not fully suitable for multi-party dialogues due to its limitation in handling non-adjacent discourse units. Then, the entity-grid [49] approach is proposed to represent the structure of a document and then applied in dialogue [14].

Recently, deep learning-based methods have been applied for higher performance,[91] introduced a novel approach for parsing multi-party dialogues, predicting dependency relations, and constructing discourse structures incrementally, thus handling the complex nature of multi-party dialogues more effectively. [18] further builds upon these methods. It presents a deep sequential model that not only predicts dependency relations but also constructs a discourse structure in a more integrated and alternating fashion. This model also incorporates global information encoding and speaker highlighting mechanisms to enhance understanding of dialogue dynamics. I then utilize this SoTA model to retrieve the discourse relation.

Dialogue Acts (DA) play a pivotal role in structural dialogue analysis. They represent

the functional aspect of language in communication, essentially conveying the speaker's intentions through speech [3]. Each dialogue act serves as a unique linguistic function, reflecting the speaker's purpose, whether it's asking a question, making a statement, or responding. In the context of language assessment and conversational modeling, DAs are critical. They not only reveal the proficiency level within individual utterances but also contribute to understanding the coherence and flow of conversation. [14, 16, 30] have shown that DAs are integral in modeling interactions and make it possible to make succinct summarization in conversation [16], as they mark key transitions and intentions within a dialogue. This aspect of DA, especially in the realm of automated assessments, requires further exploration to enhance the accuracy and effectiveness of proficiency evaluations.

Chapter 3 Methodology

3.1 Task Formulation

Given (X, Z) is an input pair of candidate's and interlocutor's responses bonded to holistic proficiency Y, where $X = \{x_0, x_1, ..., x_{|X|}\}$, $Z = \{z_0, z_1, ..., z_{|Z|}\}$. The input is then preprocessed to make $|X| = |Z|$, where $\{x_i, z_i\}$ are a conversation turn. A candidate's response x is represented as $x = [\mu_0, \mu_1, ..., \mu_{|x|}]$, consisting of $|x|$ paragraphs, and an interlocutor's response z is represented as $z = [\nu_0, \nu_1, ..., \nu_{|z|}]$, consisting of $|z|$ paragraphs. The μ_i can be represented as a sequence of sentences $\mu_i = [se_{i,0}, se_{i,1}, ..., se_{i,|\mu_i|}]$, where $se_{i,j}$ denotes the j-th sentence in the i-th candidate's response. The ν_k can be represented as a sequence of sentences $\nu_k = [sr_{k,0}, sr_{k,1}, ..., sr_{k,|\nu_k|}]$, where $sr_{k,l}$ denotes the l-th sentence in the k-th interlocutor's response. Each sentence can be further split into tokens, that is, $se_{i,j} = [w_{i,j,1}, w_{i,j,2}, ..., w_{i,j,|se_{i,j}|}]$, where $w_{i,j,m}$ denotes the m-th word of the j-th sentence in the i-th paragraph, and $sr_{k,l} = [w_{k,l,1}, w_{k,l,2}, ..., w_{k,l,|sr_{k,l}|}]$, where $w_{k,l,n}$ denotes the n-th word of the l-th sentence in the k-th paragraph.

3.2 Overall Framework

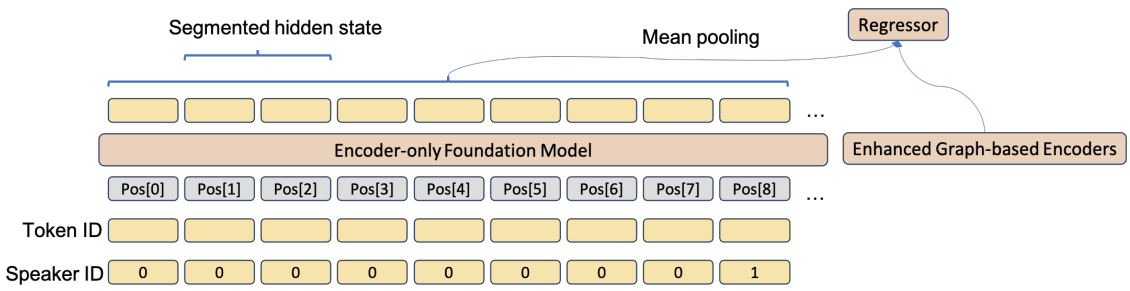


Figure 3.1: The illustration of the overall framework.

As illustrated in Figure 3.1, our model consists of (1) a transformer-based contextualized encoder that captures the contextualized information from the conversation (2) hierarchical sentence-aware content graphs to enhance the high-importance words existing in responses (3) discourse sentence-aware relation graphs that portray the speaker's interaction with responses and discourse relations in sentence granularity (4) sentence-level action graphs that represents the fine-grained act knowledge of the inter-response in token-level, and (5) a final predictor that utilizes contextualized information from (1) and the refined hierarchical embedding from (2)(3)(4) to predict the final holistic proficiency.

3.3 Encoders

3.3.1 Contextualized Encoder

The objective of the contextualized encoder centers around acquiring context-rich information, for which I utilize a foundation model. In line with [67], I employ a widely trained language model like RoBERTa [59], renowned for its impressive performance across various downstream natural language tasks. Despite its strengths, RoBERTa is constrained by a token limit of 512 in its input sequences. To accommodate our requirements for dialogues with input tokens extending up to 1,600, I have adjusted the model to accept the maximum input length by using a sliding window approach [8]. The contextualized encoder outputs a last hidden state **LHS** representing the conversation for downstream modules.

3.3.2 Enhanced Hierarchical Graph Encoders

Building upon the contextualized encoder, graph encoders aim to learn expressive representations of nodes in the constructed graphs. It consists of three main steps: heterogeneous graph construction, node embedding initialization, and graph attention operation.

Spoken content in conversation is converted into hierarchical layers, transitioning from word or phrase level to sentence level and culminating at the highest discourse level. These assemble the semantic content with two decoupled granularities: the actions encapsulated within responses and the underlying semantic context. Sentence nodes connect to

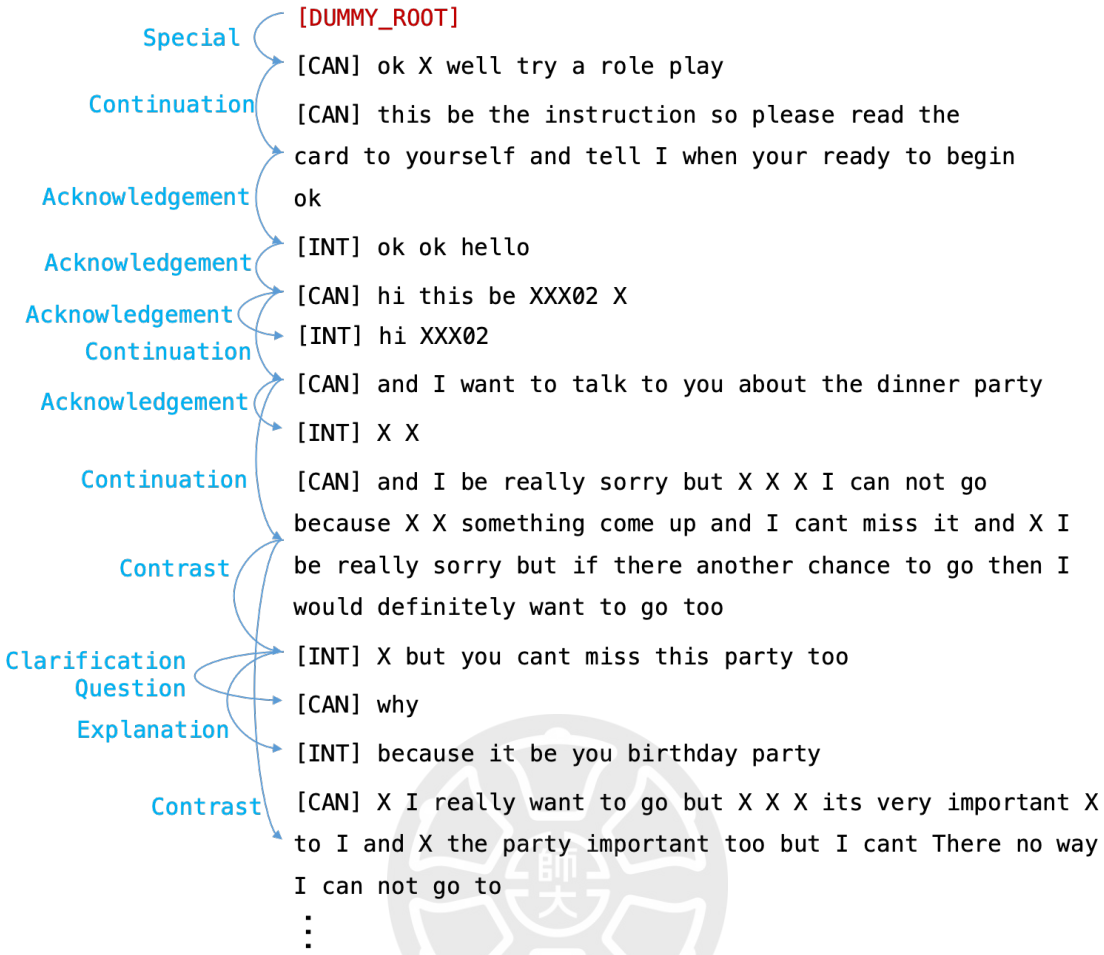


Figure 3.2: An illustration of the dialogue from the test set in the NICT JLE [45] corpus. All the nodes are responses in the dialogue. Starting from a dummy root node, the relation is constructed in a tree graph, meaning no circle exists in the graph, and the path goes from top to bottom. The discourse relation portrays the relation from one forward to another response. [CAN] is the candidate and [INT] is the interlocuter.

build two separate graphs. Similarly, discourse nodes link to sentence nodes, constructing a sentence-aware discourse relation graph. Inspired by [26], each subgraph employs a global node to unidirectionally summarize the exchanged information in the other nodes. In summary, it can be divided into three subgraphs: a hierarchical sentence-aware content graph, a discourse sentence-aware relation graph, and a sentence-level action graph.

For specific instructions for use, the hierarchical sentence-aware content graphs as the function of hierarchical context, consisting of words linked to their response based on the idea of [56, 102], which emphasizes high `tf-idf` words in spoken content. Then its encoder, denoted as Enc_{G_c}, collects the hierarchical information of spoken content. As illustrated in Figure 3.2, responses are jointly connected from top to bottom based on dis-

course relations. Accurately modeling these response relations in conversations can aid the grader model in recognizing key content for concise proficiency assessment. Discourse sentence-aware relation graphs are constructed based on structure dialogue discourse relations [16, 18, 28]. Its encoder, denoted as Enc_{G_d}, captures the structure of dialogue knowledge in the conversation.

Action graphs utilize the pattern `Subject-Predicate-Object` (SPO) triplets, the knowledge units, in responses, which express the fine-grained speaker's intent at the token level and strengthen discourse sentence-aware relation from the profound view of inter-response. Its encoder is denoted as Enc_{G_a}. Upon those three graphs, I can refine the speaker intent embedding at a fine-grained word level. A piece of detailed information about the graphs is below:

Encoder of Hierarchical Sentence-Aware Content Graph. In spoken content, words serve as the fundamental unit, whereas sentences, embodying more complex meanings, exist at a higher level. Although the hierarchical context structure has been successfully modeled in various tasks [56, 76, 102], it still falls short in effectively representing proficiency assessment. To address this, [56] suggests infusing words with prior knowledge of parsed CEFR ranks to bolster proficiency assessment features at the word level. Subsequently, I incorporate CEFR nodes into the graph, establishing a lower-tier connection with the word nodes.

Encoder of Discourse Sentence-Aware Relation Graph. In the absence of explicit structural dialogic discourse ground truth, I initially leverage a SoTA pre-trained model [18] to discern structural discourse relations among successive responses in dialogues. The resulting parsed tree delineates 18 varieties of structural dialogic discourse relations [1], such as continuation, contrast, and acknowledgment. These relations are subsequently transformed into nodes, and sentence-to-relation connections become directed edges in the graph, following methodologies outlined in [26, 28].

Encoder of Sentence-Level Action Graph. To capture and integrate knowledge units from responses into entities, I employ the AllenNLP's open information extraction (OIE) toolkit [94]. This toolkit is used to extract SPO triplets as knowledge units from each conversational response. I then construct a graph where a specially designed global entity

[1] The detail of parsed discourse relation is shown in Table 4.1

node is bi-directionally linked to SPO nodes, collectively representing the SPO triplets for each action graph.

I also make the hierarchical layers between G_c and G_d via sentence nodes. Thus, the two encoders, Enc_{G_c} and Enc_{G_d}, are connected, where the propagation of information from word to sentence and then to structural discourse is a bottom-up proceeding of hierarchical information passing with a connection of two graph encoders by sentence nodes embedding to complete the encoding process. Continuing with the message passing operation in the above graph encoders, the goal of the graph encoder is to learn a mapping connection that projects the node embeddings $\mathbf{H}^0 \in \mathbb{R}^{|V| \times d_s}$ of the t-time iterative message passing to a new representation $\mathbf{H}^t$ through an attention mechanism that encapsulates the neighbor node features. The Enc_{G_s} first operates message passing to obtain the sentence embedding and then delivers it to the Enc_{G_d} to operate another message passing for retrieving the $\mathbf{H}_{sd}$.

3.3.3 Structured Graph Construction

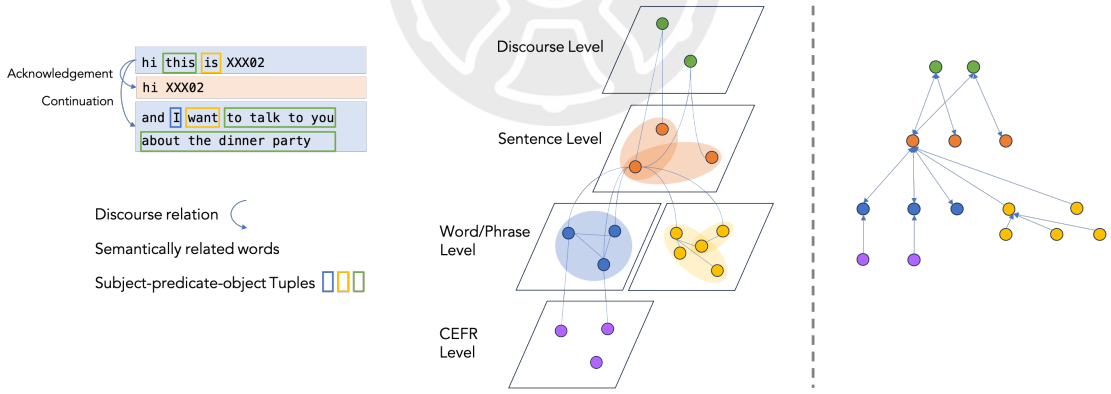


Figure 3.3: From left to right, specify (1) hierarchical contexts in inter- and intra-responses, (2) hierarchical levels, and (3) the propagation path moving from bottom to top. To implement the concept of coherence, hierarchical context in conversation data is proposed in this work. In the case of response level (intra-response), it aggregates semantic information from the semantically related words and the intents aggregated from the corresponding SPO tuple. After that, the response information propagates to the discourse level for interaction in inter-responses.

As mentioned in building the graph encoders G_c, G_d, and G_a, three main steps are necessary before retrieving the refined embedding: heterogeneous graph construction, node embedding initializing, and graph attention operation. The goal of the graph encoder is to learn a mapping connection that projects the node embeddings $\mathbf{H}^0 \in \mathbb{R}^{|V| \times d_s}$ of the t-time

iterative message passing to a new representation $\mathbf{H}^t$ through an attention mechanism that encapsulates the neighbor node features.

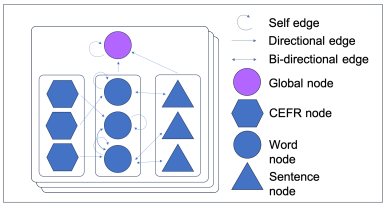


Figure 3.4: The illustration of G_c.

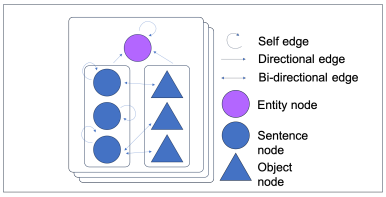


Figure 3.5: The illustration of G_d.

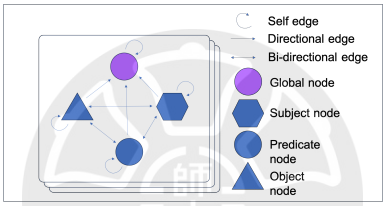


Figure 3.6: The illustration of G_a.

Heterogeneous Graph Construction The visualization of graph constructing depicts in Figure 3.4, 3.5 and 3.6. The heterogeneous graph is an information network consisting of relative primary semantic units as relay nodes and other discourse units as supernodes. Let $G = \{V, E\}$ represent an arbitrary graph, where V and E denote the node and edge sets, respectively. I define the graph G_c for Enc_{G_c} as $V = V_c \bigcup V_w \bigcup V_s \bigcup V_{cg}$ and $E = \{E_{c2w}, E_{w2w}, E_{w2s}, E_{s2w}\}$, where V_c, V_w, V_s and V_{sg} denotes CEFR, word, sentence and global nodes, respectively. And E_{c2w}, E_{w2w}, E_{w2s} , E_{s2w} and E_{s2g} stand for CEFR-to-word, word-to-word (pairwise mutual information), word-to-sentence and sentence-to-word, respectively. Another graph G_d is defined as $V = V_s \bigcup V_d \bigcup V_{sg}$ and $E = \{E_{s2d}, E_{d2s}, E_{sd2g}\}$, where V_d and V_{dg} stands for discourse relation nodes and global nodes, and E_{s2r}, E_{r2s} and E_{sr2g} represents for the edges of sentence-to-relation nodes, relation-to-sentence nodes and sentence-relation-to-global nodes, respectively. In the case of G_a, it is defined as $V = V_{subject} \bigcup V_{predicate} \bigcup V_{object} \bigcup V_{entity}$ and $E = \{E_{subject2predicate}, E_{predicate2subject}, E_{predicate2object}, E_{object2predicate}, E_{subject2object}, E_{object2subject}, E_{spo2e}\}$.

Node Embedding Initializing. Let the embedding of CEFR, word, sentence, and global

node in G_c be denoted as $\mathbf{H}^c_{G_c} \in \mathbb{R}^{|V_c| \times d_c}$, $\mathbf{H}^w_{G_c} \in \mathbb{R}^{|V_w| \times d_w}$, $\mathbf{H}^s_{G_c} \in \mathbb{R}^{|V_s| \times d_s}$, and $\mathbf{H}^{cg}_{G_c} \in \mathbb{R}^{|V_{cg}| \times d_{cg}}$, respectively. In the case of the CEFR node, I prepare its representation with Xavier initializer [36] and then build the CEFR-to-word relationship E_{c2w} with CEFR ranking of words in spoken content.

The whole scheme of the embedding computation route is illustrated in Figure 3.3. In the case of the word node, I initialize its representation by using GloVe [75]. In the case of the sentence node, the sentence node embedding is obtained by:

$$\mathbf{H}'_s = \text{MLP}([W * \mathbf{LHS}[p_{start} : p_{end}] + b); \mathbf{H}^{ngram}_s]) + \mathbf{H}^{ngram}_s, \tag{3.1}$$

$$\mathbf{H}^0_s = \text{MLP}([\mathbf{H}'_s : E^g_{G_a}]) + \mathbf{H}'_s, \tag{3.2}$$

where p_{start} start and p_{end} end segment index of tokens in response, and $[\cdot : \cdot]$ represents the segmentation operation in the axis of the input length, while $[\cdot ; \cdot]$ represents the concatenation operation at the axis of hidden state dimension. Then those responses also obtain their n-gram embedding $\mathbf{H}^{ngram}_s$ following [76, 102]. For more details, it is computed as $\text{CNN}(x_{1:|x_n|}) \odot \text{BLSTM}(g_{1:|x_n|})$, which is the local and global information from n-gram in spoken content are captured by a CNN encoder and a bidirectional long short-term memory (BLSTM), respectively.

I then fuse both via a multilayer perceptron (MLP) layer and residual mechanism to obtain the final sentence node embedding. The global node embedding is initialized after the message operation of other nodes in the graph: it uses a self-attentive mechanism to compute the global node embedding before aggregating, symbolized as $\mathbf{H}^g_{G_c}$.

For G_d, its initialized sentence node embedding is inherent from the aggregated sentence node embedding in G_c, while the relation node embeddings are initialized with a uniformized embedding table.

For G_a, I follow the method in [108] to initialize the embedding of subject, predicate, object nodes, and entity nodes. For subject, predicate, and object nodes, I first utilize AllenNLP's OIE toolkit to obtain the character of tokens in a response. Notably, a response may have multiple SPO tuples. After that, I utilize Myers' algorithm [68] [2] to operate alignment, mapping them together to obtain the parallel embedding of BERT final hidden

[2] Code available at https://github.com/tamuhey/tokenizations

state with it token's positional indexes. If the obtained embeddings are more than one token, I use an attention-pooling mechanism to limit them to a single size. The entity node is treated as a global node like in G_c and G_d; I obtain its initial embedding after the message passing operation of subject, predicate, and object nodes. I add them up as the initial embedding for each entity node.

Graph Attention Operation. Briefly, the graph attention operation encapsulates the information from neighboring nodes under the scheme of the multi-head attention head and a residual connection. The residual connection is to sidestep the gradient vanishing problem. After each graph attention layer, I utilize a position-wise feed-forward layer composed of two linear convolutional transformations.

$$z_{ij} = \text{LeakyReLU}(\mathbf{W}_a[\mathbf{W}_q\mathbf{h}_i; \mathbf{W}_k\mathbf{h}_j]), \tag{3.3}$$

$$\alpha_{ij} = \frac{\exp(z_{ij})}{\sum_{l\in N_i}\exp(z_{il})}, \tag{3.4}$$

$$\mathbf{h}'_i = \|_{k=1}^{K}\sigma(\sum_{j\in N_i}\alpha_{ij}\mathbf{W}_v\mathbf{h}_j), \tag{3.5}$$

$$\mathbf{h}''_i = \mathbf{h}_i + \mathbf{h}'_i, \tag{3.6}$$

where $\mathbf{W}_a$, $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are trainable matrixes. α_{ij} is the attention weights computed from the query and key (the neighbor node $\mathbf{h}_j$ and its node embedding $\mathbf{h}_i$). Then it aggregate the neighbor node embeddings and integrate its original node embedding to update towards as the next new node embedding. It also adds a residual connection to avoid gradient vanishing after several iterations. Inside G_c graph, I further modify its GAT layers to infuse the scalar edge weights e_{ij}, which are mapped to the multi-dimensional embedding space $\mathbf{e}_{ij} \in \mathbb{R}^{mn\times d_e}$. Thus, Equal 3.3 is modified as follows:

$$z_{ij} = \text{LeakyReLU}(\mathbf{W}_a[\mathbf{W}_q\mathbf{h}_i; \mathbf{W}_k\mathbf{h}_j; \mathbf{e}_{ij}]). \tag{3.7}$$

Sequentially, the update order is from fundamental nodes to advanced nodes. I describe the process in the groups of graphs. For the hierarchical context in Enc_{G_a}, it update

sentence nodes with their neighbor word nodes via the above GAT and FFN layer:

$$\mathbf{H}^1_{s\to w} = \text{GAT}(\mathbf{H}^0_w, \mathbf{H}^0_s, \mathbf{H}^0_s), \tag{3.8}$$

$$\mathbf{H}'^1_{s\to w} = \text{FFN}(\mathbf{H}^1_{s\to w}, \mathbf{H}^0_s). \tag{3.9}$$

It then keeps updating word node embedding from the CEFR nodes and the other surrounding word node embeddings.

$$\mathbf{H}^1_{w\to w} = \text{GAT}(\mathbf{H}^0_w, \mathbf{H}^0_w, \mathbf{H}^0_w), \tag{3.10}$$

$$\mathbf{H}'^1_{w\to w} = \text{FFN}(\mathbf{H}^1_{w\to w} + \mathbf{H}^0_w), \tag{3.11}$$

$$\mathbf{H}^1_{CEFR\to w} = \text{GAT}(\mathbf{H}^0_w, \mathbf{H}^0_c, \mathbf{H}^0_c), \tag{3.12}$$

$$\mathbf{H}'^1_{CEFR\to w} = \text{FFN}(\mathbf{H}^1_{CEFR\to w} + \mathbf{H}^0_{CEFR}). \tag{3.13}$$

Then, it obtains the aggregated word node embeddings from its origin, CEFR, and sentence node embedding of summarization.

$$\mathbf{H}^1_w = \mathbf{H}'^1_{s\to w} + \mathbf{H}'^1_{w\to w} + \mathbf{H}'^1_{CEFR\to w}. \tag{3.14}$$

For sentence node embedding, it obtains the embedding via the following computation:

$$\mathbf{H}^1_{w\to s} = \text{GAT}(\mathbf{H}^0_s, \mathbf{H}^0_w, \mathbf{H}^0_w), \tag{3.15}$$

$$\mathbf{H}^1_s = \text{FFN}(\mathbf{H}^1_{w\to s} + \mathbf{H}^0_s), \tag{3.16}$$

it only uses the word node embedding as a source to aggregate the sentence node embedding. The latest global node in G_c obtains its embedding via its belonging word node and sentence node embeddings. I first package word and sentence node embedding aligning with its global node in per sample graph to obtain H^0_{ws}, then:

$$\mathbf{H}^1_{G^g_c} = \text{GAT}(\mathbf{H}^0_{G^g_c}, \mathbf{H}^0_{ws}, \mathbf{H}^0_{ws}), \tag{3.17}$$

$$\mathbf{H}'^1_{G^g_c} = \text{FFN}(\mathbf{H}^1_{G^g_c} + \mathbf{H}^0_{G^g_c}). \tag{3.18}$$

For the next graph G_d, I obtain the sentence node embedding via:

$$\mathbf{H}^1_{w\to s} = \text{GAT}(\mathbf{H}^0_w, \mathbf{H}^0_s, \mathbf{H}^0_s), \tag{3.19}$$

$$\mathbf{H}'^1_{w\to s} = \text{FFN}(\mathbf{H}^1_{w\to s}, \mathbf{H}^0_s). \tag{3.20}$$

Finally, for the graph G_a, I follow the below formula to compute the subject, object, and predicate node embeddings, which are computed with the following formulas:

$$\mathbf{H}^1_{subject\to predicate} = \text{GAT}(\mathbf{H}^0_{predicate}, \mathbf{H}^0_{subject}, \mathbf{H}^0_{subject}), \tag{3.21}$$

$$\mathbf{H}'^1_{subject\to predicate} = \text{FFN}(\mathbf{H}^1_{subject\to predicate} + \mathbf{H}^0_{subject}), \tag{3.22}$$

$$\mathbf{H}^1_{object\to predicate} = \text{GAT}(\mathbf{H}^0_{predicate}, \mathbf{H}^0_{subject}, \mathbf{H}^0_{subject}), \tag{3.23}$$

$$\mathbf{H}'^1_{object\to predicate} = \text{FFN}(\mathbf{H}^1_{object\to predicate} + \mathbf{H}^0_{predicate}), \tag{3.24}$$

$$\mathbf{H}^1_{predicate} = \mathbf{H}'^1_{subject\to predicate} + \mathbf{H}'^1_{object\to predicate}, \tag{3.25}$$

where $\mathbf{H}^1_{subject\to predicate}$ and $\mathbf{H}^1_{object\to predicate}$ are the aggregated predicate node embedding from subject node embeddings and object node embeddings, respectively.

$$\mathbf{H}^1_{predicate\to subject} = \text{GAT}(\mathbf{H}^0_{subject}, \mathbf{H}^0_{predicate}, \mathbf{H}^0_{predicate}), \tag{3.26}$$

$$\mathbf{H}'^1_{predicate\to subject} = \text{FFN}(\mathbf{H}^1_{predicate\to subject} + \mathbf{H}^0_{subject}), \tag{3.27}$$

$$\mathbf{H}^1_{object\to subject} = \text{GAT}(\mathbf{H}^0_{subject}, \mathbf{H}^0_{object}, \mathbf{H}^0_{object}), \tag{3.28}$$

$$\mathbf{H}'^1_{object\to subject} = \text{FFN}(\mathbf{H}^1_{object\to subject} + \mathbf{H}^0_{subject}), \tag{3.29}$$

$$\mathbf{H}^1_{subject} = \mathbf{H}'^1_{predicate\to subject} + \mathbf{H}'^1_{object\to subject}, \tag{3.30}$$

where $\mathbf{H}^1_{predicate\to subject}$ and $\mathbf{H}^1_{object\to subject}$ are the aggregated subject node embedding from predicate node embeddings and object node embeddings, respectively.

$$\mathbf{H}^1_{subject\to object} = \text{GAT}(\mathbf{H}^0_{object}, \mathbf{H}^0_{subject}, \mathbf{H}^0_{subject}), \tag{3.31}$$

$$\mathbf{H}'^1_{subject\to object} = \text{FFN}(\mathbf{H}^1_{subject\to object} + \mathbf{H}^0_{object}), \tag{3.32}$$

$$\mathbf{H}^1_{predicate\to object} = \text{GAT}(\mathbf{H}^0_{object}, \mathbf{H}^0_{predicate}, \mathbf{H}^0_{predicate}), \tag{3.33}$$

$$\mathbf{H}'^1_{predicate\to object} = \text{FFN}(\mathbf{H}^1_{predicate\to object} + \mathbf{H}^0_{object}), \tag{3.34}$$

$$\mathbf{H}^1_{object} = \mathbf{H}'^1_{subject\to object} + \mathbf{H}'^1_{predicate\to object}. \tag{3.35}$$

where $\mathbf{H}^1_{subject\to object}$ and $\mathbf{H}^1_{predicate\to object}$ are the aggregated object node embedding from

subject node embeddings and predicate node embeddings, respectively. All add a residual connection to avoid gradient vanishing. And then I process the entity node embedding so as to concatenate them for sentence node embedding in G_c.

$$\mathbf{H}^1_{G^a_e} = \text{GAT}(\mathbf{H}^0_{G^a_e}, \mathbf{H}^0_{spo}, \mathbf{H}^0_{spo}), \tag{3.36}$$

$$\mathbf{H}'^1_{G^a_e} = \text{FFN}(\mathbf{H}^1_{G^a_e} + \mathbf{H}^0_{G^a_{ce}}), \tag{3.37}$$

those entity node embeddings are then used as an MLP layer with a residual mechanism to add back to its belonging sentence node embedding. And I will get the embeddings after t-times recursive aggreagation: $\mathbf{H}^t_{G_c}$, $\mathbf{H}^t_{G_d}$, and $\mathbf{H}^t_{G_a}$, which can be denoted as $\mathbf{H}^g_{G_c}$, $\mathbf{H}^g_{G_d}$, and $\mathbf{H}^g_{G_a}$, respectively.

3.4 Regressor

Figure 3.7: The illustration of operating in the regressor.

The regressor, as shown in Figure 3.7, uses several sources of embedding: mean pooled **LHS**, $\mathbf{H}^g_{G_c}$, $\mathbf{H}^g_{G_d}$, and $\mathbf{H}^g_{G_a}$, to predict the holistic proficiency score $\hat{Y}$. The regressor is a pairwise attentive structural network in which non-duplicated embedding pairs are selected to operate a self-attentive mechanism and then concatenate them all to operate the MLP layer for the final score.

For a clear contour of the regressor, I illustrate its computation in the below:

$$hid_{\text{bert}} = \text{ProjectLinear}_{BERT}(\mathbf{LHS}) \in \mathbb{R}^{D_{\text{hidden}}}, \tag{3.38}$$

$$hid_{\text{graph}} = \text{ProjectLinear}_{graph}(\mathbf{H}^g_{G_c} or \mathbf{H}^g_{G_d}) \in \mathbb{R}^{D_{\text{hidden}}}, \tag{3.39}$$

where ProjectLinear represents the projection layer into the same feature space, hid is the hidden states after projection. Then, I produce the maximum number of pairs from input

embeddings, which is heads $\in \mathbb{R}^{C_2^{en}}$, where en is the total number of input embeddings.

$$\mathbf{H}^{\text{combo}}_{(m,k)} = \mathbf{H}_m \odot \mathbf{H}_k \in \mathbb{R}^{2D_{\text{hidden}}}, \tag{3.40}$$

$$\mathbf{H}^{\text{head}}_{(m,k,l)} = \text{ReLU}(A_k(\mathbf{H}^{\text{combo}}_{(m,k)})) \in \mathbb{R}^{D_{\text{hidden}}}, \tag{3.41}$$

$$\mathbf{H}_{\text{combo}} = \odot_{\substack{m,k\in\{1,\ldots,N_{\text{combo}}\}\\ l\in\{1,\ldots,N_{\text{heads}}\}}} \mathbf{H}^{\text{head}}_{(m,k,l)} \in \mathbb{R}^{D_{\text{hidden}} N_{\text{heads}} N_{\text{combo}}}, \tag{3.42}$$

$$\hat{Y} = \text{Linear}(\mathbf{H}_{\text{combo}}), \tag{3.43}$$

where $\mathbf{H}^{\text{combo}}_{(m,k)}$ represents the combined embedding of pair (m,k), $m \neq k$, and $N_{\text{combo}} = \frac{1}{2} N^{\text{Reg.}}_{\text{inputs}} (N^{\text{Reg.}}_{\text{inputs}} - 1)$, $N^{\text{Reg.}}_{\text{inputs}}$ is the number of inputs for the regressor. $\mathbf{H}^{\text{head}}_{(m,k,l)}$ denotes the output of the l-th attention head, $\mathbf{H}_{\text{combo}}$ is the concatenated output of all attention heads, and $\hat{Y}$ represents the predicted proficiency score obtained from the linear function applied to $\mathbf{H}_{\text{combo}}$.

3.5 Optimization

Reweighted Loss Function. During the training process, we seek to minimize the weighted mean squared error (MSE) loss between the predicted and target holistic score. The strategy is adopted from [56]. The final layer of the prediction head is bound to the scoring scale (0-6), which is an ordinal ranking number. We regard this problem as a regression problem. Therefore, we use the MSE loss function for the prediction task. Furthermore, to reduce the impact of imbalanced data, we adopt reweighting techniques in the loss function:

$$loss_{reweighted} = loss_{MSE} * (1 - \frac{N}{N_c})^{\beta}, \tag{3.44}$$

where β is a controllable parameter, N represents the total number of speakers in the training set, while N_c refers to the number of speakers in each CEFR level within it.

Posttraining Strategy on Multiple Auxiliary Tasks. This strategy aims to optimize the initial parameters of the contextualized encoder. Inspired by [22], we facilitate the LM's ability to learn multi-aspects of assessment knowledge and promote prediction accuracy on the main scoring task. It follows the training stages in [5]. In the first three epochs,

the model undergoes posttraining on the EFCAMDAT dataset [34], which is rich in full CEFR labels (A1-C2). Subsequently, the model is trained on the NICT JLE corpus in the next three epochs. This strategy is applied to optimize the initial parameters of the contextualized encoder.

Chapter 4 Experimental Settings and Results

4.1 Overview

Our research corpus is labeled with corresponding CEFR (Common European Framework of Reference for Languages) tags, based on the widespread adoption of CEFR. CEFR is a systematic classification of language proficiency levels, comprising A1/A2 (Basic Users), B1/B2 (Independent Users), and C1/C2 (Proficient Users), totaling six levels. Since I focus on studying English language corpora, the CEFR proficiency labels here pertain specifically to the English language. Additionally, as our research corpus primarily targets general English learning, rather than requiring advanced skills or specific domain knowledge (e.g., business English), data availability tends to be scarcer in levels C1 and C2.

Here, I introduce the NICT-JLE corpora, which include speaking evaluation items, while the others are related to written evaluations. Regarding the corpora, I will analyze them from several perspectives: data distribution, SST score to CEFR score, the distribution of structural discourse relation in different CEFR levels, the action distribution in each response, and a sample in the corpora to explain our analysis.

4.2 The NICT JLE corpus

NICT JLE is a collection of English speaking proficiency test data from 1,301 speakers spanning from 1999 to 2002. The entire corpus is designed as interview evaluations for English assessments. Each interview evaluation represents a unique participant and in-

cludes their past English learning experiences or other English proficiency test scores. Finally, each test is rated on a Japanese Standard Speaking Test (SST) scale, ranging from 1 to 9, indicating the overall speaking proficiency level.

Each test is structured into five stages: an ice-breaking question, describing a single image, simulating a conversation in a specific scenario, narrating a story based on multiple assigned images, and concluding with closing questions. The topics of conversation revolve around issues common in public life. The range of topics covered is diverse, with some topics further categorized by difficulty levels, making it suitable for distinguishing English proficiency levels among speakers. Due to its design for interview evaluation, it is well-suited for research and system design aligned with real-world application scenarios.

Detailed annotations are provided for spoken content, covering aspects like grammar errors (only 167 instances are marked), emotions, pauses, non-fluency in speech, and instances where the interviewee's speech overlaps with the interviewer's [46, 99] [1]. These annotations are beneficial for conducting detailed studies among individuals with varying English proficiency levels. However, this corpus only offers manually annotated transcriptions of speech and lacks audio information, limiting analysis from a sound perspective.

Moreover, the distribution of predicted discourse relation is listed in Table 4.1. Under the observation, four out of them over 10% are Continuation, Question-answer Pair, Comment, Clarification Question, and Acknowledgement. From the perspective of CEFR rankings, the Continuation remains subtle, while the Question-answer Pair and the Clarification Question gradually decline from the beginner to advanced learner. On the other hand, the Comment and the Acknowledgement show an upward trend. For the definition and examples of the above relations:

Continuation. It indicates the extension of the same line of thought, topic, or narrative without a significant shift in focus or argumentative direction. Given an example: "I went to the market yesterday. I also visited the library." Here, the second sentence continues the narrative of the speaker's activities without changing the topic.

[1] https://alaginrc.nict.go.jp/nict_jle/src/taglist.pdf describes the function of each tag. https://alaginrc.nict.go.jp/nict_jle/src/readme_transcription.pdf accounts for the meaning of each grammar error tag with a succinct example.

Table 4.1: The distribution of predicted discourse relation types on the NICT-JLE Corpus.

Discourse Types	A1	A2	B1	B2	C
Continuation	11.750	12.343	12.638	12.700	11.649
Question-answer pair	19.843	17.444	14.972	13.838	14.018
Contrast	1.577	2.621	4.526	5.436	5.513
Q-Elab	2.359	1.935	1.066	0.908	1.160
Explanation	0.975	1.348	1.937	2.179	1.871
Comment	13.752	13.250	13.637	11.150	15.989
Background	0.000	0.000	0.002	0.000	0.000
Result	2.936	4.668	4.838	5.218	5.014
Correction	1.088	0.950	0.884	0.690	0.935
Parallel	0.043	0.055	0.052	0.073	0.062
Alternation	0.235	0.187	0.298	0.460	0.249
Conditional	0.003	0.009	0.023	0.048	0.012
Clarification question	14.876	12.198	10.911	10.775	10.102
Acknowledgement	22.693	25.685	27.616	30.448	26.802
Elaboration	4.632	4.274	3.862	3.656	4.502
Narration	0.000	0.001	0.000	0.000	0.000
Special	3.237	3.034	2.736	2.421	2.120

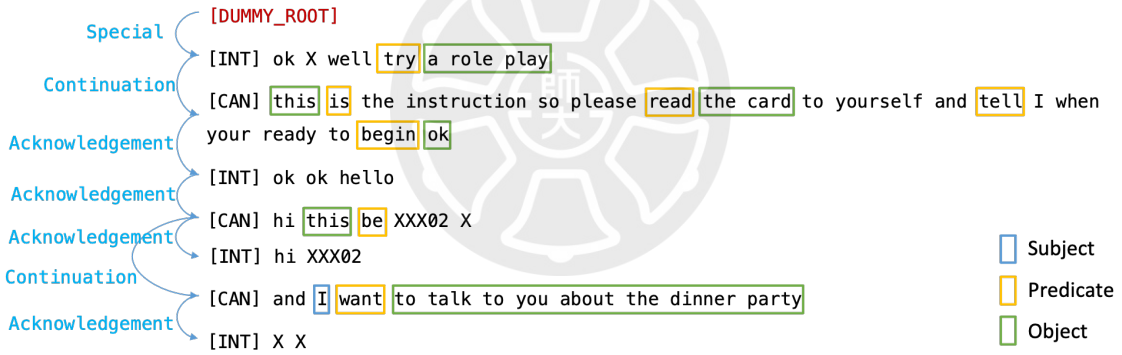


Figure 4.1: The multiple action tuples in response and the discourse relation in the response pairs.

Question-Answer Pair. This is a relation where one segment poses a question and the subsequent segment provides the answer. Example: "What time is the meeting? The meeting is at 3 PM." The first segment asks a question, and the second segment directly answers it.

Comment. This relation occurs when one segment adds information, opinion, or an evaluative remark about the preceding segment. For instance, "It's raining heavily. That's probably why the traffic is so bad." The second sentence provides a comment or explanation regarding the situation described in the first sentence.

Clarification Question. A clarification question is a relation where a segment in the

discourse seeks to clarify or request elaboration on a point made in the previous segment. ”We need to finish the project by next week. Could you explain how we’re going to do that?” The second sentence asks for clarification on the plan or method mentioned in the first sentence.

Acknowledgement. It is used when one segment acknowledges or shows receipt of the information or statement made in the previous segment. Example: ”I’ve finished the report. Great, thank you for your hard work.” The second sentence acknowledges the completion of the task mentioned in the first sentence.

Let’s take a look at the sample, as illustrated in Figure 4.1, each response may have more than one SPO tuple. These tuples can be seen as a fine-grained action in responses. Different from `tf-idf` words that focus on word frequency in inter- or intra-documents that can eliminate the influence of some stopwords, the SPO tuple provides a clear three-element structure to make modeling focusing on those real speaker intents. The discourse relation then helps connect the two responses at a higher level of sentence.

4.3 The EFCAMDAT Corpus

EFCAMDAT (EF-Cambridge Open Language Database) is a large, publicly available corpus designed for writing assessments. It is intended to aid research in English language learning and assist in English language teaching. The database encompasses participants from 198 countries worldwide, with 174,743 English learners. The corpus comprises proficiency scores ranging from levels 1 to 16, convertible to CEFR. Its test design features multi-unit questions to address and respond to various contextual topics. As illustrated in Table 4.2, there are a total of 1,180,310 scripts (7,126,752 sentences, 83,543,480 words), providing a substantial quantity of data, alleviating concerns regarding data scarcity. In addition to annotating sentence grammar information, some responses are marked with grammar error information, along with the English proficiency level of each user’s content, which can be converted into CEFR levels, as shown in Table 4.3. It aims to utilize grammar error information to pre-filter some grammatical errors in spoken content before conducting oral content assessments. This not only provides information for predicting English proficiency levels but also offers direct feedback to English learners on how to

improve their spoken English content.

Table 4.2: Title themes designed in EFCAMDAT.

ID	Essay topic	ID	Essay topic
1:1	Introducing yourself by email	7:1	Giving instructions to play a game
1:3	Writing an online profile	8:2	Reviewing a song for a website
2:1	Describing your favourite day	9:7	Writing an apology email
2:6	Telling someone what you' re doing	11:1	Writing a movie review
2:8	Describing your family' s eating habits	12:1	Turning down an invitation
3:1	Replying to a new penpal	13:4	Giving advice about budgeting
4:1	Writing about what you do	15:1	Covering a news story
6:4	Writing a resume	16:8	Researching a legendary creature

Table 4.3: Score alignment of mapping CEFR to score in the EFCAMDAT corpus.

CEFR	EFCAMDAT
A1	1,2,3
A2	4,5,6
B1	7,8,9
B2	10,11,12
C1	13,14,15
C2	16

I use EFCAMDAT for pertaining the initial parameters of the contextualized encoder, that is, the extended RoBERTa. It is mentioned in Section Section 3.5. Furthermore, the EFCAMDAT has some error tags in its annotation, which should be revised or influence the effectiveness of the grader model. I follow [60, 90] to proceed with data preprocessing, also I replace many meaningless tokens and fix many misspelling words which are in a long-term distribution that may degrade the model to give a proper grade. After doing the above revision, I convert the score to CEFR labels following Tabel 4.3. The code for preprocessing the EFCAMDAT corpus is available at https://github.com/a2d8a4v/EFCAMDAT_local.

4.4 Implementational Details

Our implementation is based on the code of [102]. I utilize deep graph library [103] to construct our graphs and operate message passing. The size of hidden states is adjusted to 256, and the dimension of SPO embedding is set to 64.

Last but not least, in constructing a sentence-aware action graph, it is necessary to

align two token sequences from the same input but via different tokenizers. In such a case, a response text input via a tokenizer in RoBERTa's WordPiece and AllenNLP's OIE tokenizer, I utilize Myers' algorithm [68] [2] to operate alignment, mapping them together to obtain the parallel embedding of BERT final hidden state with it token's positional indexes.

4.5 Experimental Setup

The model is trained with an Adam optimizer [52], with hyperparameters of the batch size as 64 and the gradient accumulation steps as 2. To alleviate the model uncertainty [110], I train each kind of experiment 5 times with an exponentially decaying learning rate initialized at different numbers and a decay factor of 0.85 per epoch. The model will stop training if validation loss does not descend four times. I utilize the last epoch as the best checkpoint.

The model performance evaluation adopts root-mean-square error (RMSE), while further comparisons include Pearson's correlation coefficient (PCC). Additionally, to determine the accuracy of classification, I designed margin accuracy within 0.5 (ACC05) and 1.0 (ACC10), under-estimate rate (UR), and over-estimate rate (OR). However, I also wanted to estimate the situation at individual levels. Therefore, RMSE, PCC, ACC05, ACC10, UR, and OR are individually divided into two types: micro and macro.

The target Y can be further extended to $y_1, y_2, y_3, ..., y_n$, while the $\hat{Y}$ is the target denotes $\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_n$, n is the number of samples in total. For more details of the definition of each metric, RMSE is a measure used to assess the differences between values predicted by a model and the values actually observed. It's defined as the square root of the average of squared differences between prediction and actual observation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{4.1}$$

[2] Code available at https://github.com/tamuhey/tokenizations

where n is the number of observations, $\hat{y}_i$ represents the predicted value, and y_i represents the observed value.

RMSE is commonly used in regression analysis, forecasting, and predictive modeling. It provides a way to quantify the difference between values predicted by a model and the values observed. A lower RMSE value indicates a better fit of the model to the data.

In terms of PCC, measures the linear correlation between two variables, $\hat{y}_i$ and y_i. It quantifies the degree to which a relationship between these two variables can be described using a line.

$$PCC = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4.2}$$

where $\hat{y}_i$ and y_i are the values of the variables, $\bar{\hat{y}}$ and $\bar{y}$ are the mean values of the variables y and $\hat{y}$, respectively. n is the number of data points. PCC is a fundamental statistical tool used to assess the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. It is widely used in various fields like finance, medicine, and social sciences.

Following ACC05, it is a statistical measure used to evaluate the performance of a classification or prediction model. It is defined as the proportion of predictions that are within a margin of 0.5 units from the true value. It is especially useful in scenarios where predictions that are close enough to the actual value are considered acceptable. This metric is a relaxed form of accuracy measurement, allowing a small margin of error in the predictions. It is often applied in fields like forecasting and classification, where exact precision is not always necessary.

$$\text{ACC05} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(|\hat{y}_i - y_i| \leq 0.5), \tag{4.3}$$

in which $\hat{y}_i$ is the predicted value for the i^{th} observation. y_i is the actual value for the i^{th}

observation. n is the total number of observations. $\mathbb{I}$ is the indicator function, which is 1 if $|\hat{y}_i - y_i| \leq 0.5$ and 0 otherwise.

ACC10 has a similar definition as ACC05 but increases its margin to 1. It is worth noticing that a margin of 1 is across to the upper or lower rank, having a higher capability to contain more data points.

For the latest metrics, UR and OR, UR is a metric that quantifies the frequency at which a predictive model's estimations are lower than the actual values. It is defined as the ratio of the number of underestimations to the total number of predictions.

$$\text{UR} = \frac{\text{Number of underestimations}}{\text{Total number of predictions}} \tag{4.4}$$

An underestimation occurs when the predicted value ($\hat{y}_i$) is less than the actual value (y_i).

while OR measures how often a model's predictions exceed the actual values. It is calculated as the ratio of the number of overestimations to the total number of predictions.

$$\text{OR} = \frac{\text{Number of overestimations}}{\text{Total number of predictions}} \tag{4.5}$$

An overestimation occurs when the predicted value ($\hat{y}_i$) is greater than the actual value (y_i).

UR and OR are important in evaluating the bias in predictive models. They are especially useful in scenarios where the consequences of underestimating differ from those of overestimating. For instance, in financial forecasting, different strategies might be required to address revenue underestimations as opposed to cost overestimations.

Table 4.4: Conversion table from SST to CEFR according to [29]. Label 0 represents failure.

SST	Original CEFR	Modified CEFR	Labels
-	-	-	0
1	preA1	-	-
2	A1	A1	1
3	A1+	A1	1
4	A2	A2	2
5	A2+	A2	2
6	B1	B1	3
7	B1+	B1	3
8	B1+	B1	3
9	B2	B2	4
9	B2+	B2	4
9	C	C	5

Table 4.5: Distribution of training, validation, and test sets. The number of data is equal to the number of speakers.

	A1	A2	B1	B2	C	Total
Train	220	501	228	32	16	997
Validation	14	127	9	4	2	156
Test	23	90	26	4	2	145

4.6 Data Preprocessing

The NICT JLE [45] corpus comprises oral interviews conducted in English with 1,281 Japanese individuals and 20 American native individuals. It contains 2 million human-annotated words with rich annotations such as disfluencies (e.g., filled pauses), but only text-annotated transcriptions can be accessed. The interviews in the corpus involve a test-taker being interviewed by an English tutor or examiner and consist of multiple stages, including warm-up questions, single picture description, simulated conversation (role-play), story description using designated pictures, and wind-down questions. The interviews resemble conversations on specific topics, which were held with non-fixed prompt questions, and the interviewees are expected to respond promptly. Each interview has a standard speaking test score (SST), which is ranked on a scale and can be converted to corresponding CEFR levels [29]. Among them, C is a native-like fluency. For our experiments, I further assigned labels for the converted CEFR levels as shown in Table 4.4. I used up 1,298 interviews and split them into training, validation, and test sets, as summarized in Table 4.5 except for three preA1 speakers.

In the case of data preprocessing, I utilized the NICT JLE corpus in the version of 4.1 [45] [3] in which is an XML-like format that can not parsed directly via XML parsers. Also, some special occasions exist in this corpus, such as Japanese Kanatana words, Arabic numbers, span words with privacy concerns, United Kindom English and United States English of the same word, overlapped responses, and mislabelled tags. It leads to careful preprocessing before tagging the human-annotated spoken content. To avoid wrong labeling, I first fix the above issues in the corpus, then process it sentence-wise to retrieve several kinds of tags described in [118] parallelly aligned with spoken content. In the detail of conversation segmentation, the candidate's and interlocutor's responses are partitioned based on the closing delimiter marking at the end of each stage into paragraph boundary inputs; I then divide all the paragraphs into responses by sentence boundary closing delimiter marking at the end of each response.

The latter, I follow [22] to tag other kinds of text-based information, such as part-of-speech and morphic-syntactic, via UDPipe [95] toolkit, in which the LM is trained on mass web available media text content. The words of parts-of-speech (A1 to C2)[4][5] are then labeled by utilizing the CEFR-J Wordlist [33, 69] Vocabulary Profile 1.6 via Stanza [78] toolkit. It is worth noticing that a word can have multiple CEFR levels depending on its phrasing. All details on data preprocessing are released at https://github.com/a2d8a4v/local_for_nict_jle.

Table 4.6: Distribution of stages in each CEFR level.

	A1	A2	B1	B2	C	All
Max	5	5	5	5	10	10
Min	4	4	5	5	8	4

For the detail of this corpus after analysis, I first demonstrate the stages' distribution in each CEFR level in Table 4.6. Due to the situations that occurred in collecting the dataset, few conversation tests have lower than 5 stages in a conversation test; otherwise, most of them are at an average of 5 stages. For C-level speakers, candidate has more than one topic in a stage of a conversation test, I divide different topics as a new stage to sidestep the problem of overlapping in topic-specific dialogue.

From the perspective of credibility when using this corpus, The statistic of the token

[3] Version 4.1 was released in 2012.
[4] Data available at http://www.cefr-j.org/data/CEFRJ_wordlist_ver1.6.zip
[5] Data available at https://github.com/openlanguageprofiles/olp-en-cefrj

Table 4.7: Distribution of number of responses in each CEFR level.

	A1	A2	B1	B2	C	All
Max	61	88	77	82	57	88
Min	1	1	1	2	5	1

count in response from several aspects and granularities is validated. The average sentence length of the speaker in each level (A1-C) is 76.833, 81.987, 91.139, 103.225, and 209.300 tokens. The maximum and minimum sentences of each stage in each level are illustrated in Table 4.7. The '1' can be a continuous response such as hesitation or a simple answer to the counterpart in a conversation, while 88 is the number that presents the speaker's proficiency in language. If considering both perspectives (token and sentence count), it is sure that C can handle their learning language well: speak more numbers of responses and use more words in their responses.

Table 4.8: Distribution of the number of responses in each CEFR level.

	A1	A2	B1	B2	C	All
Max	122	176	154	164	114	176
Min	2	2	2	4	10	2

Nevertheless, if the interlocutor counts, The maximum and minimum sentences of each stage in each level will become in Table 4.8, which do not change the summarization of the analysis above. Then I go to a higher level, stage level, to observe the variation in analysis. Each level's average number of tokens is 674.195, 1057.148, 1401.863, 1660.475, and 4979.550, respectively. Apparently, C especially has more tokens than others.

Table 4.9: Distribution of number of maximum and minimum tokens of each sentence in each level.

	A1	A2	B1	B2	C	All
Max	273	472	451	404	479	479
Min	0	0	0	0	0	0

If observing from the perspective of token level, the maximum and minimum tokens of each sentence in each level are illustrated in Table 4.9, the maximum and minimum tokens of each stage in each level are illustrated in Table 4.10, And the average number of tokens per sentence in each level in Table 4.11. C jumps to a higher number of tokens even in the minimum. Furthermore, the speaker may keep speaking without the interruption of the interlocutor, so the tokens can be up to 400, even if it is not a normal amount

Table 4.10: Distribution of number of maximum and minimum tokens of each stage in each level.

	A1	A2	B1	B2	C	All
Max	425	661	1089	849	1209	1209
Min	2	3	9	23	106	106

Table 4.11: Distribution of average numbers of tokens per sentence in each level.

	A1	A2	B1	B2	C	All
Max	9.507	661	1089	849	1209	1209
Min	2	3	9	23	106	106

of free speaking, about 250 words in experience. After averaging, the numbers become 9.507, 14.331, 17.491, 17.581, and 24.574. It gradually increases the number. All the observations confirm that fluency-like people speak fast, well, and more.

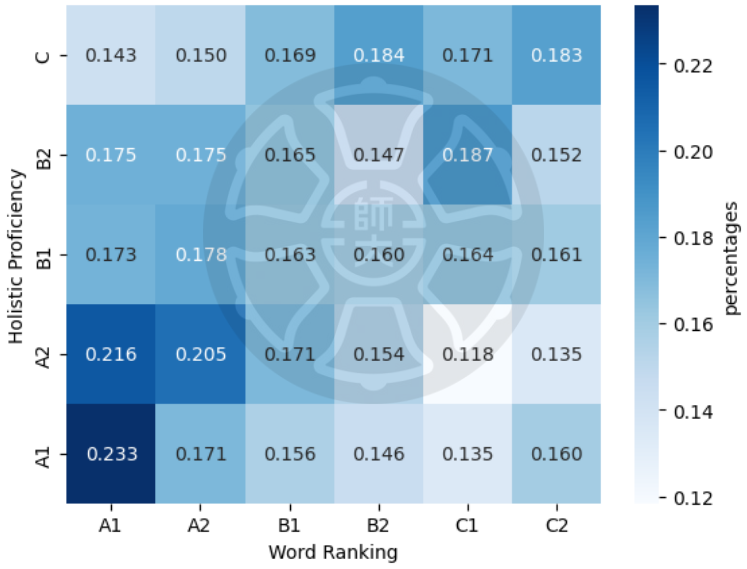


Figure 4.2: The confusion matrix of comparing the changes of frequency of word CEFR ranking among holistic proficiency.

I further analyze the percentage matrix comparing word CEFR rankings at the x-axis with the holistic proficiency in conversation tests at the y-axis. I depict Figure 4.2 via normalizing each holistic CEFR group to observe the percentage of a specific word ranking in holistic CEFR groups. It shows that A1 speakers use A1 ranking words a lot, and A2 speakers use A2 ranking words a lot. Then B2, C1, and C2 speakers, use every ranking word in a fair distribution but high in difficult words.

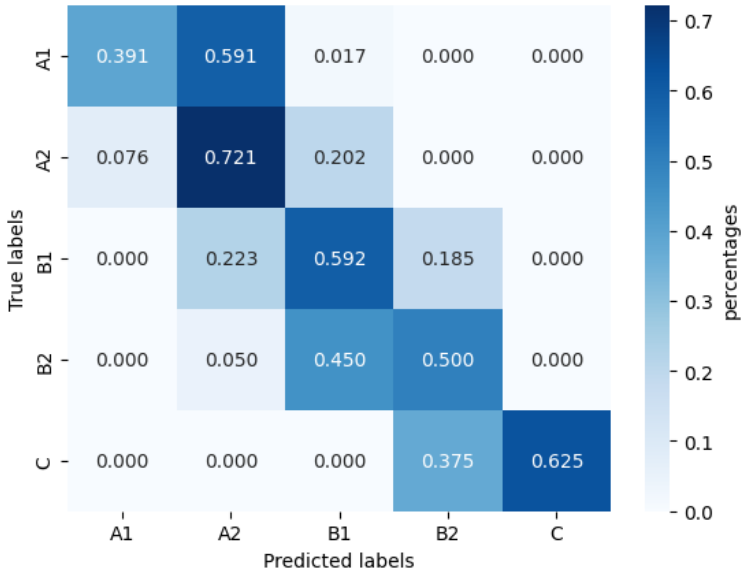


Figure 4.3: The confusion matrix of evaluation of our proposed method (BERT+CDA) on testing set.

4.7 Main Results

In this section, I present the key findings from our experiments, showcasing the performance of our proposed methods. Table 4.12 offers a comprehensive comparison of our approaches against two reference models doing on conversation test: BLSTMATT, a BLSTM-attention mechanism of prompt encoder [79, 80] focuses on encompassing context, language usage, and delivery aspects. I omitted the delivery aspect for our experiments due to constraints, relying solely on annotated transcriptions. Another is BERT, which primarily focuses on leveraging contextual information to assess linguistic proficiency holistically [62]. I use the contextual encoder mentioned in Chapter 3 to fit the input size so a little bit different to the original method in [62].

Our proposed method demonstrates a notable overall improvement across all metrics. This outcome suggests that the graph-based encoder, incorporated into our approach, aids the grading model in effectively emphasizing the hierarchical context and speaker intents, outperforming the BERT and BLSTMATT models with stable scores in other metrics. On the other hand, BLSTMATT meets the obstacle when the input is too long, and its performance is unstable with the input length. BERT encounters the problem of fail-to-training when only small data, which graph-enhanced methods can help alleviate based on observations in the experimental results. I further visualize these improvements in

Figure 4.3, highlighting that most scores align with the correct positions, albeit with a slight tendency to under-score.

The **hierarchical sentence-aware content graph** Enc_{G_c}, **discourse sentence-aware relation graph** Enc_{G_d}, and **sentence-level action graph** Enc_{G_a}, represents the utilization of G_c, G_d, and G_a, respectively. In Table 4.12, they then are symbolized as C, D and A, respectively. Posttraining is the MLM task to pre-train the contextualized encoder on the NICT-JLE corpus.

4.8 Ablation Studies

In our quest to delve deeper into the factors impacting the performance of the grader model, I conducted a series of comprehensive ablation studies. These experiments were designed to dissect the influence of hierarchical context and speaker intents on our model's effectiveness.

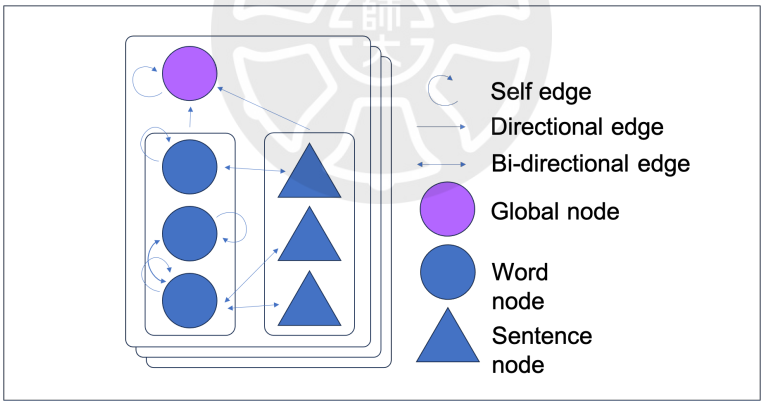


Figure 4.4: The illustration of G_c but removing the CEFR nodes.

The Effectiveness of CEFR Nodes. One critical component of our approach is the introduction of CEFR nodes within the hierarchical context, which inform the CEFR ranking of inclination words. To evaluate the significance of these CEFR nodes, I conducted an experiment wherein I removed them from the hierarchical sentence-aware content graph, like in Figure 4.4. The results shown in Table 4.13 conclusively indicate that the inclusion of CEFR nodes significantly contributes to the model's performance.

Another noteworthy factor is TF-IDF, which assesses the importance of keywords within the context. It can help our grader model identify attention-worthy parts that the

Table 4.12: The results of the grading model: BLSTMSTT and BERT are the baseline models. Upon the foundation model, the proposed measures are C, D, and A, representing the hierarchical context, discourse relation, and action in responses, respectively.

Model	**RMSE↓**	**Macro-RMSE↓**	**PCC↑**	**Accuracy↑**		**Macro-Accuracy↑**		**Over-estimate rate ↓**		**Under-estimate rate ↓**	
				≤0.5	≤1.0	≤0.5	≤1.0	micro	macro	micro	macro
BSTLMATT	0.802 ± 0.049	1.502 ± 0.092	0.042 ± 0.023	60.515 ± 1.384	78.517 ± 1.158	19.979 ± 0.098	40.122 ± 0.294	15.529 ± 0.544	19.391 ± 0.679	23.955 ± 1.922	60.629 ± 0.620
BSTLMATT+lu	0.761 ± 0.068	1.374 ± 0.192	0.340 ± 0.251	58.217 ± 3.137	83.343 ± 5.082	24.873 ± 4.626	46.803 ± 7.087	11.226 ± 3.522	13.674 ± 4.732	30.557 ± 6.473	61.453 ± 2.152
BERT	0.656 ± 0.034	1.029 ± 0.068	0.273 ± 0.075	61.812 ± 2.945	85.890 ± 1.498	29.687 ± 2.813	55.340 ± 3.401	20.129 ± 2.485	21.592 ± 2.690	18.058 ± 3.409	48.720 ± 1.583
BERT+D	0.520 ± 0.005	0.682 ± 0.004	0.729 ± 0.001	65.961 ± 0.518	94.763 ± 0.189	42.037 ± 0.357	84.671 ± 0.160	20.334 ± 1.027	16.028 ± 0.714	13.705 ± 0.697	41.935 ± 0.455
BERT+C	0.539 ± 0.022	0.714 ± 0.101	0.711 ± 0.036	64.234 ± 0.223	93.677 ± 0.820	47.024 ± 4.811	81.489 ± 8.288	20.696 ± 2.014	16.687 ± 2.063	15.070 ± 2.157	36.289 ± 6.866
BERT+CD	0.502 ± 0.001	0.606 ± 0.002	0.751 ± 0.000	67.437 ± 0.378	96.323 ± 0.167	57.731 ± 0.405	92.936 ± 0.141	19.304 ± 0.258	16.673 ± 0.250	13.259 ± 0.417	25.596 ± 0.525
BERT+CDA	0.507 ± 0.001	0.590 ± 0.001	0.755 ± 0.000	66.323 ± 0.346	96.407 ± 0.056	55.654 ± 0.901	91.968 ± 0.066	21.365 ± 0.380	17.271 ± 0.264	12.312 ± 0.111	27.075 ± 1.014
C	0.525 ± 0.002	0.683 ± 0.002	0.724 ± 0.001	65.376 ± 0.287	93.928 ± 0.189	49.310 ± 0.218	82.170 ± 0.257	18.524 ± 0.374	15.318 ± 0.239	16.100 ± 0.111	35.372 ± 0.081
D	0.567 ± 0.068	0.822 ± 0.203	0.643 ± 0.142	63.760 ± 3.434	92.618 ± 3.835	38.780 ± 6.691	75.889 ± 14.771	20.334 ± 3.664	15.877 ± 3.282	15.905 ± 1.620	45.342 ± 3.930
C+D	0.567 ± 0.068	0.822 ± 0.203	0.643 ± 0.142	63.760 ± 3.434	92.618 ± 3.835	38.780 ± 6.691	75.889 ± 14.771	20.334 ± 3.664	15.877 ± 3.282	15.905 ± 1.620	45.342 ± 3.930
C+D+A	0.501 ± 0.001	0.646 ± 0.003	0.745 ± 0.001	67.584 ± 0.752	95.926 ± 0.060	47.452 ± 0.967	89.493 ± 0.122	16.852 ± 1.535	13.839 ± 1.327	15.564 ± 0.790	38.709 ± 0.382

Table 4.13: The results of BERT+C and BERT+C without the CEFR nodes.

Model	RMSE↓	Macro-RMSE↓	PCC↑	Accuracy		Macro-Accuracy		Over-estimate rate↓		Under-estimate rate↓	
				≤0.5 (↑)	≤1.0 (↑)	≤0.5 (↑)	≤1.0 (↑)	micro	macro	micro	macro
BERT+C	0.539 ± 0.022	0.714 ± 0.101	0.711 ± 0.036	64.234 ± 0.223	93.677 ± 0.820	47.024 ± 4.811	81.489 ± 8.288	20.696 ± 2.014	16.687 ± 2.063	15.070 ± 2.157	36.289 ± 6.866
BERT+C(-CEFRn)	0.556 ± 0.000	0.725 ± 0.002	0.699 ± 0.000	60.501 ± 0.167	93.928 ± 0.142	39.969 ± 0.069	80.770 ± 0.084	26.685 ± 0.142	20.203 ± 0.136	12.813 ± 0.233	39.828 ± 0.138

contextualized encoder may otherwise overlook. By examining a conversation segment, I can discern the differences between TF-IDF and the contextualized encoder, highlighting the need for additional supporting information in this task. This job should be done under both contextual embedding and graph-based model learning due to their complementary [73]. By observing Table 4.14, the improvements of performance can directly prove that thing.

Table 4.14: The results of BERT and BERT+C.

Model	RMSE↓	Macro-RMSE↓	PCC↑	Accuracy		Macro-Accuracy		Over-estimate rate↓		Under-estimate rate↓	
				≤0.5 (↑)	≤1.0 (↑)	≤0.5 (↑)	≤1.0 (↑)	micro	macro	micro	macro
BERT	0.656 ± 0.034	1.029 ± 0.068	0.273 ± 0.075	61.812 ± 2.945	85.890 ± 1.498	29.687 ± 2.813	55.340 ± 3.401	20.129 ± 2.485	21.592 ± 2.690	18.058 ± 3.409	48.720 ± 1.583
BERT+C	0.539 ± 0.022	0.714 ± 0.101	0.711 ± 0.036	64.234 ± 0.223	93.677 ± 0.820	47.024 ± 4.811	81.489 ± 8.288	20.696 ± 2.014	16.687 ± 2.063	15.070 ± 2.157	36.289 ± 6.866

Speaker Intents. Our analysis examines the number of tagged discourse relations within each proficiency level. As indicated in Table 4.15, I observed that the RMSE and PCC improved a little bit but a subtle degrading in accuracy and OR and UR. It indicates that

Table 4.15: The results of BERT+CDA and BERT+CD.

Model	RMSE↓	Macro-RMSE↓	PCC↑	Accuracy		Macro-Accuracy		Over-estimate rate↓		Under-estimate rate↓	
				≤0.5 (↑)	≤1.0 (↑)	≤0.5 (↑)	≤1.0 (↑)	micro	macro	micro	macro
BERT+CDA	0.507 ± 0.001	0.590 ± 0.001	0.755 ± 0.000	66.323 ± 0.346	96.407 ± 0.056	55.654 ± 0.901	91.968 ± 0.066	21.365 ± 0.380	17.271 ± 0.264	12.312 ± 0.111	27.075 ± 1.014
BERT+CD	0.502 ± 0.001	0.606 ± 0.002	0.751 ± 0.000	67.437 ± 0.378	96.323 ± 0.167	57.731 ± 0.405	92.936 ± 0.141	19.304 ± 0.258	16.673 ± 0.250	13.259 ± 0.417	25.596 ± 0.525

One interesting thing is that Figure 4.5 shows that the action in speaker responses has strong correlations relevant to the high level of holistic proficiency. But if the speaker is below B1, the action becomes meaningless because the percentage is spared fairly in each holistic proficiency.

Candidate and Interlocutor. I engage in a thoughtful discussion on a pivotal issue concerning dialogue evaluation. While I primarily assess the candidate, it is imperative to consider that the candidate's spoken content can vary widely. However, when the interlocutor's dialogue process remains smooth, with rare breakdowns, incorporating the interlocutor's spoken content maintains overall dialogue fluency and provides a more comprehensive perspective on completeness.

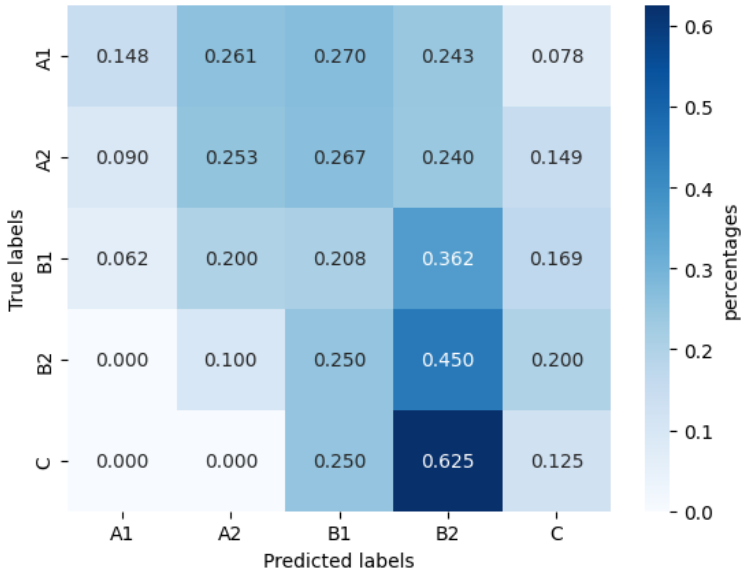


Figure 4.5: The illustration of the pure hierarchical context of action (A).

The Needs of LM. Despite hierarchical context outperforming baselines, hierarchical with LMs outperform hierarchical context without LMs. Figure 4.6 depicts the offset toward the left downside by comparing BERT+CD to pure C+D.

Furthermore, the graph modeling limits to having structural knowledge which is predefined, while LM can learn unfactual knowledge. This indicates both are indispensable components in modeling conversation data.

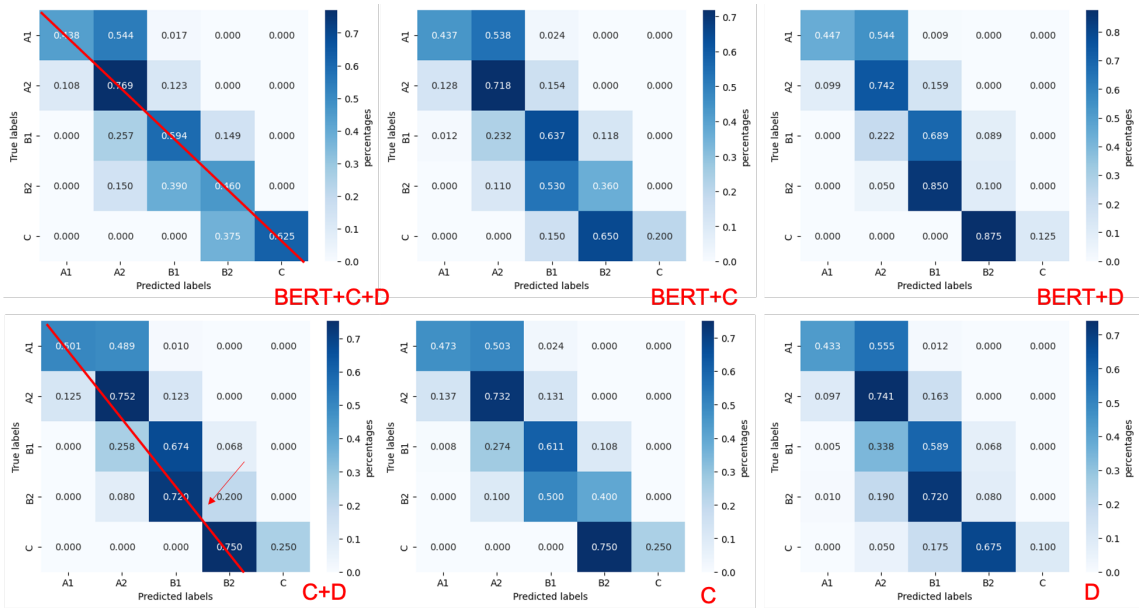


Figure 4.6: The illustration of pure hierarchical context or with BERT.

Proficiency Level Performance Analysis. To assess the performance of the proposed

Table 4.16: The results of BERT+CDA, analytic showing the ACC0.5, 1.0, OR, UR in proficiency levels. Score is the column name that maps CEFR A1-C to labels 1-5.

Score	RMSE↓	Accuracy↑		Over-estimate rate↓	Under-estimate rate↓
		≤0.5	≤1.0		
1	0.628 ± 0.004	44.870 ± 0.696	92.174 ± 0.000	55.130 ± 0.696	0.000 ± 0.000
2	0.431 ± 0.001	74.472 ± 0.305	99.281 ± 0.090	15.685 ± 0.360	9.843 ± 0.168
3	0.552 ± 0.001	61.385 ± 0.576	93.385 ± 0.377	15.538 ± 0.576	23.077 ± 0.000
5	0.477 ± 0.005	52.500 ± 5.000	100.000 ± 0.000	0.000 ± 0.000	47.500 ± 5.000
4	0.861 ± 0.004	45.000 ± 0.000	75.000 ± 0.000	0.000 ± 0.000	55.000 ± 0.000

methods against baselines, I treated our five integer scores (from 1 to 5) as classes, and evaluated PCC, RMSE, ACC0.5 and ACC1.0, UR, OR. As depicted in Figure 4.6, despite the results influenced by the data imbalance, our proposed method can mitigate the data imbalance problem, provide useful information made from spoken response coherence to make the grader model differ CEFR groups. In details among CEFR levels, Table 4.16 again indicates the same thing that my proposed method can really help avoid the data imbalance problem.

Other Issues. Actually, there are some problems I left for future discussions. (1) The error indicating of which character (interlocutor or candidate). Speaker diarization aims to distinguish two or the above different speakers in a period of speech. However, the NICT-JLE corpus has tags 'A', and 'B' to specify each response to the interlocutor or candidate, respectively. (2) The errors in ASR decoded transcriptions. ASR aims to recognize speech in text. Despite the superior performance in nowadays ASR systems, error propagation still has an effect on the downstream tasks. For instance, transcription with wrong-decoded words from ASR may impact the results of accessing a speaker's holistic proficiency. However, the NICT-JLE is a human-annotated corpus, in which the ASR errors is absent. I leave this issue to future discussions by trying another corpus: ICNALE - spoken dialogue section [43]. (3) The comparison with other traditional methods in modeling local coherence and global coherence.

Chapter 5 Conclusions and Future Works

5.1 Conclusions

In this paper, I proposed using the hierarchical graph modeling for conversation tests, which amalgamates hierarchical context: semantically related words, actions, and discourse relations. The experiments have revealed the efficacy of our proposed methods and the chances of investigating in coherence. The ablation study shows the efficacy of each hierarchical context in holistic proficiency groups. In conclusion, the knowledge graph helps LMs improve their performance on their downstream tasks, such as ASA.

5.2 Future Works

Further challenges related to ASA remain in this field For instance. For example, uncertainty in ASA is a vital topic that contributes to the impartiality of feedback when it meets the real world. The application of the Monte-Carlo theorem that ensembles multiple foundation models trained with randomized hyperparameters [110]. Another topic like the cold-start problem, which decreases performance when encountering unknown questions, has also been explored [74]. Also, the investigation of feature usage in ASA occupies a lot, but it is still worth studying, like self-supervised learning representations whose effectiveness has been proved [6, 15], in the future. Moreover, due to the ripe of ASA, many frameworks or systems have been released, and the upcoming popular issue in system framework also become a topic recently.

Another related issue is about different traits of feedback. Grammar errors and cor-

rection play a vital role when encountering users' needs. Proper suggestions can indeed help language learners improve their grammar complexity. Also, the generation of spoken content is an art, that is, generating any kind of modified spoken content depending on several circumstances would be an important thing in the course of communication or interaction with people.

Foreign accent conversion [24] would also become another popular issue cause it imitates the L1 accent but with L2's accent and spoken content, with corrected pronunciation. Correcting the speaker's mispronunciation in synthesized sound is also a vital issue in foreign accent conversion as well. Articulation classification [20, 113] and the conversion to phoneme [66] also directly influence the core paths to fulfill meaningful pronunciation feedback to language learners. Or learn multiple English native speakers' accents to ease living in countries. Yet no such system can do all the things in one, which indicates that such a market still has great potential in the future.

References

[1] R. Al-Ghezi, Y. Getman, E. Voskoboinik, M. Singh, and M. Kurimo. Automatic rating of spontaneous speech for low-resource languages. In *Proceedings of IEEE the Spoken Language Technology Workshop (SLT)*, pages 339–345, 2023.

[2] S. M. Armenti. *Computer Science Education with English Learners*. University of Rhode Island, 2018.

[3] J. L. Austin. *How to Do Things with Words*. Oxford: University Press, 1962.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems (NIPS)*, 33:12449–12460, 2020.

[5] S. Bannò and M. Matassoni. Cross-corpora experiments of automatic proficiency assessment and error detection for spoken english. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 82–91, 2022.

[6] S. Bannò and M. Matassoni. Proficiency assessment of l2 spoken english using wav2vec 2.0. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1088–1095. IEEE, 2023.

[7] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. Gales. Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning. In *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, pages 126–130, 2023.

[8] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document Transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[9] J. Bérešová. The impact of the CEFR on teaching and testing English in the local context. *Theory and practice in language studies*, 7(11):959–964, 2017.

[10] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[11] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[12] D. Busbridge, D. Sherburn, P. Cavallo, and N. Y. Hammerla. Relational graph attention networks. 2019.

[13] M. Canale and M. Swain. Theoretical bases of com-municative approaches to second language teaching and testing. *Applied linguistics*, 1(1):1–47, 1980.

[14] A. Cervone, E. Stepanov, and G. Riccardi. Coherence models for dialogue. In *Proceedings Interspeech*, pages 1011–1015, 2018.

[15] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen. 3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to English pronunciation assessment. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 575–582. IEEE, 2022.

[16] J. Chen and D. Yang. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2021.

[17] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma, et al. Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine. *ETS Research Report Series*, 2018(1):1–31, 2018.

[18] T.-C. Chi and A. Rudnicky. Structured Dialogue Discourse Parsing. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDD)*, pages 325–335, 2022.

[19] S.-H. Chiu, T.-H. Lo, F.-A. Chao, and B. Chen. Cross-utterance reranking models with BERT and graph convolutional networks for conversational speech recognition. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1104–1110, 2021.

[20] C. J. Cho, P. Wu, A. Mohamed, and G. K. Anumanchipalli. Evidence of vocal tract articulation in self-supervised learning of speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[21] S. P. Corder. The significance of learner's errors. 1967.

[22] H. Craighead, A. Caines, P. Buttery, and H. Yannakoudakis. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2258–2269, 2020.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[24] S. Ding, G. Zhao, and R. Gutierrez-Osuna. Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language*, 72:101302, 2022.

[25] Y. Dong, Z. Hu, K. Wang, Y. Sun, and J. Tang. Heterogeneous network representation learning. In C. Bessiere, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4861–4867. International Joint Conferences on Artificial Intelligence Organization, 2020.

[26] H. Du, Y. Feng, C. Li, Y. Li, Y. Lan, and D. Zhao. Structure-discourse hierarchical graph for conditional question answering on long documents. In *Proceedings of the Association for Computational Linguistics (ACL, Findings)*, pages 6282–6293, 2023.

[27] A. Farajidizaji, V. Raina, and M. Gales. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models, 2023.

[28] X. Feng, X. Feng, B. Qin, and X. Geng. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In Z.-H. Zhou, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3808–3814, 2021.

[29] B. Flanagan, S. Hirokawa, E. Kaneko, E. Izumi, and H. Ogata. A multi-model SVR approach to estimating the CEFR proficiency level of grammar item features. In *Proceedings of International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 521–526, 2017.

[30] C. Fu, Z. Chen, J. Shi, B. Wu, C. Liu, C. T. Ishi, and H. Ishiguro. HAG: Hierarchical attention with graph network for dialogue act classification in conversation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[31] J. Fu, Y. Chiba, T. Nose, and A. Ito. Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, 116:86–97, 2020.

[32] X. Fu, J. Zhang, Z. Meng, and I. King. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*, pages 2331–2341, 2020.

[33] Y. Fujinuma and M. Hagiwara. Semi-supervised joint estimation of word and document readability. In *Proceedings of the Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs)*, pages 150–155, 2021.

[34] J. Geertzen, T. Alexopoulou, A. Korhonen, et al. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *Proc. of the Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254. Citeseer, 2013.

[35] L. Gilanyi, X. A. Gao, and S. Wang. Emi and clil in asian schools: A scoping review of empirical research between 2015 and 2022. *Heliyon*, 9(6):e16365, 2023.

[36] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[37] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[38] N. Hernandez, N. Oulbaz, and T. Faine. Open corpora and toolkit for assessing text readability in French. In *Proceedings of the Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 54–61. European Language Resources Association, 2022.

[39] D. Higgins, X. Xi, K. Zechner, and D. Williamson. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech Language*, 25(2):282–306, 2011.

[40] P. Howson. *The English effect*. British Council, London, 2013.

[41] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[42] Y.-P. Huang. 以英語授課：一個探討臺灣的大學教師教學情況之質性個案研究. 外國語文研究, (20):27–62, 06 2014.

[43] S. Ishikawa. Design of the ICNALE-spoken: A new database for multi-modal contrastive interlanguage analysis. *Learner corpus studies in Asia and the world*, 2:63–76, 2014.

[44] E. Islam, T. Hain, and P. Nomo Sudro. Simulation of teacher-learner interaction in English language pronunciation learning. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2023.

[45] E. Izumi, K. Uchimoto, and H. Isahara. The NICT JLE corpus. 12:7, 2004.

[46] E. Izumi, K. Uchimoto, and H. Isahara. Error annotation for corpus of japanese learner english. In *Proceedings of the International Workshop on Linguistically Interpreted Corpora (LINC)*, 2005.

[47] P. Jamshid Lou and M. Johnson. Improving disfluency detection by self-training a self-attentive model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3754–3763, 2020.

[48] P. Jamshid Lou, Y. Wang, and M. Johnson. Neural constituency parsing of speech transcripts. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2756–2765, 2019.

[49] A. K. Joshi and S. Kuhn. Centered logic: The role of entity centered sentence representation in natural language inferencing. In *Proceedings of the international joint conference on Artificial intelligence*, pages 435–439, 1979.

[50] U. Khandelwal, H. He, P. Qi, and D. Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 284–294, 2018.

[51] E. Kim, J.-J. Jeon, H. Seo, and H. Kim. Automatic pronunciation assessment using self-supervised speech representation learning. *arXiv preprint arXiv:2204.03863*, 2022.

[52] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[53] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[54] R. Lado. Language testing: The construction and use of foreign language tests. a teacher's book. 1961.

[55] Y. Lei and M. Allen. English language learners in computer science education: A scoping review. In *Proceedings of the ACM Technical Symposium on Computer Science Education*, pages 57–63, 2022.

[56] J.-T. Li, T.-H. Lo, B.-C. Yan, Y.-C. Hsu, and B. Chen. Graph-enhanced Transformer architecture with novel use of CEFR vocabulary profile and filled pauses in automated speaking assessment. In *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, pages 109–113, 2023.

[57] N. Li and J. Wu. Exploring assessment for learning practices in the emi classroom in the context of Taiwanese higher education. *Language Education Assessment*, 1:28–44, 2018.

[58] R. J. Lickley. *Detecting disfluency in spontaneous speech*. PhD thesis, University of Edinburgh, 1994.

[59] 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 preprint arXiv:1907.11692*, 2019.

[60] Y.-C. Lo, J.-J. Chen, C. Yang, and J. Chang. Cool English: a grammatical error correction system based on large learner corpora. In *Proceedings of the International Conference on Computational Linguistics (ICCL)*, pages 82–85. Association for Computational Linguistics, 2018.

[61] A. Loukina and A. Cahill. Automated scoring across different modalities. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 130–135, 2016.

[62] R. Ma, M. Qian, M. Gales, and K. M. Knill. Adapting an ASR Foundation Model for Spoken Language Assessment. In *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, pages 104–108, 2023.

[63] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[64] D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 11 2000.

[65] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[66] C. G. McGhee, K. M. Knill, and M. Gales. Towards Acoustic-to-Articulatory Inversion for Pronunciation Training. In *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, pages 66–70, 2023.

[67] S. W. McKnight, A. Civelekoglu, M. Gales, S. Bannò, A. Liusie, and K. M. Knill. Automatic assessment of conversational speaking tests. In *Proceedings the Workshop on Speech and Language Technology in Education (SLaTE)*, pages 99–103, 2023.

[68] E. W. Myers. An o (nd) difference algorithm and its variations. *Algorithmica*, 1(1-4):251–266, 1986.

[69] M. Negishi, T. Takada, and Y. Tono. A progress report on the development of the cefr-j. In *In Exploring language frameworks: Proceedings of the ALTE Kraków Conference*, 2013.

[70] S. M. Ngangbam. Taiwan’s bilingual nation policy 2030: Concerned issues and suggestions / 2030 年國家雙語政策之重要問題與建議. *European Journal of Literature, Language and Linguistics Studies*, 6(2), 2022.

[71] C. of Europe. *Common European Framework of Reference for Languages: Learning Teaching, Assessment*. Cambridge University Press, Cambridge, UK, 2001.

[72] C. M. Ormerod, A. Malhotra, and A. Jafari. Automated essay scoring using efficient Transformer-based language models. *arXiv preprint arXiv:2102.13136*, 2021.

[73] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[74] J. Park and S. Choi. Addressing cold start problem for end-to-end automatic speech scoring. In *Proceedings of Interspeech*, pages 994–998, 2023.

[75] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[76] T.-A. Phan, N.-D. N. Nguyen, and K.-H. N. Bui. HeterGraphLongSum: Heterogeneous graph neural network with passage aggregation for extractive long document summarization. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 6248–6258, 2022.

[77] J. B. Pride and J. Holmes. Sociolinguistics: selected readings. 1972.

[78] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, 2020.

[79] Y. Qian, P. Lange, K. Evanini, R. Pugh, R. Ubale, M. Mulholland, and X. Wang. Neural approaches to automated speech scoring of monologue and dialogue responses. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8112–8116, 2019.

[80] Y. Qian, R. Ubale, M. Mulholland, K. Evanini, and X. Wang. A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, pages 979–986, 2018.

[81] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[82] D. Ramesh and S. K. Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527, 2022.

[83] R. Ridley, L. He, X.-y. Dai, S. Huang, and J. Chen. Automated cross-prompt scoring of essay traits. In *Proceedings of the Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, volume 35, pages 13745–13753, 2021.

[84] P. M. Rogerson-Revell. Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC Journal*, 52(1):189–205, 2021.

[85] M. Saeki, Y. Matsuyama, S. Kobashikawa, T. Ogawa, and T. Kobayashi. Analysis of multimodal features for speaking proficiency scoring in an interview dialogue. In *Proceedings of the Workshop Spoken Language Technology Workshop (SLT)*, pages 629–635, 2021.

[86] K. Sakaguchi, M. Heilman, and N. Madnani. Effective feature integration for automated short answer scoring. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1049–1054, 2015.

[87] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[88] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. In *Proceedings of The International Conference of The Semantic Web (ESWC)*, pages 593–607. Springer, 2018.

[89] V. J. Schmalz and A. Brutti. Automatic assessment of english cefr levels using bert embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics*, 2021.

[90] I. Shatz. Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236, 2020.

[91] Z. Shi and M. Huang. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[92] S. Shimauchi. English-medium instruction in the internationalization of higher education in japan: Rationales and issues. *Educational Studies in Japan*, 12:77–90, 10 2018.

[93] L. Skidmore and R. Moore. Incremental disfluency detection for spoken learner English. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 272–278, 2022.

[94] G. Stanovsky, J. Michael, L. Zettlemoyer, and I. Dagan. Supervised open information extraction. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[95] M. Straka and J. Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL) Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, 2017.

[96] W. Sun and X. L. Rong. English education reform in Asian countries, 05 2021.

[97] E. Szügyi, S. Etler, A. Beaton, and M. Stede. Automated assessment of language proficiency on German data. In *KONVENS*, 2019.

[98] A. Tack, T. François, P. Desmet, and C. Fairon. NT2Lex: A CEFR-graded lexical resource for Dutch as a foreign language linked to open Dutch WordNet. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 137–146, 2018.

[99] M. Tanimura, K. Takeuchi, and H. Isahara. From learners' corpora to expert knowledge description: Analyzing prepositions in the NICT JLE (Japanese learner English) corpus. In *Proceedings of the IWLeL: an interactive workshop on language e-learning*, pages 139–147. Waseda University, 2005.

[100] A. Van Moere and R. Downey. Technology and artificial intelligence in language assessment. *Handbook of second language assessment*, pages 342–357, 2016.

[101] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[102] D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6209–6219, 2020.

[103] M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

[104] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu. Heterogeneous graph attention network. In *Proceedings of The World Wide Web conference (WWW)*, pages 2022–2032, 2019.

[105] Y. Wang, C. Wang, R. Li, and H. Lin. On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3416–3425, 2022.

[106] D. G. Williams. South Korean higher education English-medium instruction (emi) policy: From 'resentment' to 'remedy'. *English Today*, page 1–6, 2023.

[107] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: state-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, 2020.

[108] T. Wu, X. Bai, W. Guo, W. Liu, S. Li, and Y. Yang. Modeling fine-grained information via knowledge-aware hierarchical graph for zero-shot entity retrieval. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, page 1021–1029, 2023.

[109] T.-I. Wu, T.-H. Lo, F.-A. Chao, Y.-T. Sung, and B. Chen. Effective neural modeling leveraging readability features for automated essay scoring. In *Proceedings of The Workshop on Speech and Language Technology in Education (SLaTE)*, pages 81–85, 2023.

[110] X. Wu, K. M. Knill, M. J. Gales, and A. Malinin. Ensemble approaches for uncertainty in spoken language assessment. In *Proceedings Interspeech 2020*, pages 3860–3864, 2020.

[111] J. Xie, K. Cai, L. Kong, J. Zhou, and W. Qu. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the International Conference on Computational Linguistics (ICCL)*, pages 2724–2733, 2022.

[112] R. Xu, W. Pan, C. Chen, X. Chen, S. Lin, and X. Li. Graph-based model using text simplification for readability assessment. In *Proceedings of The International Conference on Asian Language Processing (IALP)*, pages 401–406, 2022.

[113] B.-C. Yan, H.-W. Wang, Y.-C. Wang, and B. Chen. Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[114] C. Yang, Y. Xiao, Y. Zhang, Y. Sun, and J. Han. Heterogeneous network representation learning: A unified framework with survey and benchmark. *TKDE*, 2020.

[115] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, 2020.

[116] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP, Findings)*, Nov. 2020.

[117] H. Yannakoudakis, Ø. E. Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267, 2018.

[118] K. Yasuda, K. Kitamura, S. Yamamoto, and M. Yanagida. Development and applications of an English learner corpus with multiple information tags. *Journal of Natural Language Processing*, 16(4):447–463, 2009.

[119] K. Zechner, D. Higgins, and X. Xi. SpeechraterTM: A construct-driven approach to scoring spontaneous non-native speech. In *Proceedings of The Speech and Language Technology in Education (SLaTE)*, pages 128–131, 2007.

[120] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895, 2009.

[121] J. Zeng, Y. Xie, X. Yu, J. S. Lee, and D.-X. Zhou. Enhancing automatic readability assessment with pre-training and soft labels for ordinal regression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP, Findings)*, pages 4557–4568, 2022.

[122] H. Zhang, X. Liu, and J. Zhang. Contrastive hierarchical discourse graph for scientific document summarization. In *Proceedings of the Workshop on Computational Approaches to Discourse (CODI)*, pages 37–47, 2023.