Language Assessment Using Word Family-Based Automated Item Generation: Evaluating Item Quality Using Teacher Ratings

S. Susan Marandi, Department of English, Faculty of Literature, Alzahra University, Iran
Shaghayegh Hosseini, Department of English, Faculty of Literature, Alzahra University, Iran

WorldCALL 2023 – CALL in Critical Times
Conference Proceedings

Abstract

The integration of Artificial Intelligence (AI) technologies has initiated a new era in language assessment practices, revolutionizing the field with its innovative approaches. This study introduces an advanced Automated Item Generation (AIG) system that utilizes word families as a foundation to automatically generate test items. The primary objective of this research is to investigate the effectiveness of the AIG system in producing high-quality questions through a comprehensive evaluation that combines both quantitative and qualitative measures. The AIG system is developed using cutting-edge machine learning and deep learning techniques, enabling it to enhance and facilitate the language assessment process by generating a substantial number of items. To assess the quality of the generated questions, a group of 30 experienced English teachers participated in the evaluation process. The participants assessed the quality of multiple-choice and fill-in-the-blank questions generated by the AIG system using a 4-point scale. To supplement the quantitative analysis, interviews were conducted to capture the perspectives of the teachers concerning the integration of AIG in language assessment. The findings demonstrate highly promising outcomes in terms of question quality, validating the efficacy of employing word families as a linguistic basis for generating test items. By shedding light on the advantages and effectiveness of utilizing word families as a fundamental lexical unit for AIG, this study contributes to the field of automated item generation in language assessment.

Keywords: Automated Item Generation (AIG), Intelligent Computer-Assisted Language Learning (ICALL), Language Assessment, Word Families

worldCALL

WorldCALL Conference 2023 in Chiang Mai, Thailand

Introduction

With the recent advancements in Artificial Intelligence (AI), its applications have significantly impacted various domains, and the field of language learning is no exception. In language education, AI has evolved as a powerful tool, introducing innovative methods for learners to acquire proficiency. This evolution extends further into language assessment, where AI applications facilitate and enhance evaluation processes, ensuring more accurate and efficient assessments. The creation of test items is a crucial step in the assessment process which can be facilitated by AI. Typically, teachers invest a significant amount of time and effort in creating test items, resulting in a process that is both time-consuming and costly (Fulcher, 2013). This highlights the need to develop a reliable, efficient, and cost-effective method for generating high-quality test items that align with assessment objectives and accurately reflect the proficiency level of test takers.

Automated Item Generation (AIG) involves the utilization of computer technology to automatically generate test questions based on a predetermined model, combining the expertise of test developers with modern computational techniques (Gierl and Haladyna, 2013). This process employs pre-determined models and algorithms, resulting in the creation of an AIG system which is capable of rapidly generating an extensive number of test items within seconds. Aligning with the demands of the digital age, this capability significantly accelerates the item generation process, offering a wide range of high-quality test items and alleviating the workload of human item developers (Alves et al., 2010).

In the context of this study, the AIG system was specifically designed to employ word families as its primary lexical unit. In linguistic terms, a word family is a group of words which comprises a headword and its derived and inflected forms (Bauer and Nation, 1993). For instance, the headword *able* expands to encompass various derived and inflected forms, such as *ability*, *abilities*, *inability*, and *unable*. By incorporating the diverse forms originating from a single headword, the AIG system ensured a comprehensive coverage of language elements within the generated test items.

1. Methods

The current study included two main phases. Initially, an Automated Item Generation (AIG) system was created, which served as the fundamental element of this research. This was followed by the gathering and analyzing of quantitative and qualitative data, employing a mixed-method approach. The quantitative analysis was conducted through an item evaluation list with the aim of evaluating item quality, while the subsequent qualitative phase involved semi-structured interviews for exploring the participants' perspectives on the potential benefits and challenges posed by AIG.

The created AIG system was designed to generate word-family-based multiple-choice and fill-in-the-blank items for all three grades of Iranian high schools, which include grades 10, 11, and 12. The system was built according to a three-step process proposed by Gierl et al. (2021), with modifications in the second step. The initial step involved identifying the content for item generation, which, in this case were Iranian high school books. The second step included applying this content to an item model, with some adjustments to the original approach to avoid fixed item models. The final step involved the generation of items by the system.

The system was coded using Python programming language, integrating state-of-the-art NLP, machine learning, and deep learning techniques. A deep learning model was trained using SentenceTransformers, enabling the system to assign difficulty levels to the sentences of a corpus, ensuring that the generated questions differed from exact sentences in the books. The sentences utilized in the generated questions were sourced from the Race dataset (Lai et al., 2017). Over 2,000 generated questions were examined, leading to system improvements and enhanced question quality. Provided below are examples of multiple-choice (MC) and fill-in-the-blank (FB) items generated by the system:

MC: Many children use the Internet to get _____ knowledge and information, and to relax in their free time.

a) use b) useful c) used d) usefulness

FB: This online encyclopedia is _____ (write) by thousands of people around the world.

The participants of the study were 30 Iranian English teachers, including 22 females and 8 males. They were recruited through purposive sampling and were required to have at least 5 years of teaching experience in English at Iranian high schools, as well as experience in test design, specifically item generation. Table 1 presents a summary of demographic information on the participants, highlighting an average teaching experience of 9.17 years, ranging from 5 to 20 years and a mean age of 28.60, with participants' ages varying from 23 to 40 years old.

Demographic variable	Minimum	Maximum	Mean	SD
Teaching experience	5	20	9.17	4.18
Age	23	40	28.60	4.87

Table 1: Demographic information of participants

In the quantitative phase of the study, the teachers assessed the quality of 18 automatically generated items using a 4-point scale (Gierl et al., 2021). This scale graded items based on their overall quality, where a score of 1 indicated that the item was considered unacceptable; 2 indicated that major revisions were necessary, 3 denoted the need for only minor revisions, and a score of 4 meant that the item was acceptable with no further revisions required. To ensure a balanced representation across questions formats, the 18 items were evenly divided across 10^{th}, 11^{th}, and 12^{th} grade, with three multiple-choice, and 3 fill-in-the-blank questions per grade. A digital form created with Google Forms was utilized for administering the evaluation list, enabling the teachers to access and submit their assessments digitally.

Semi-structured interviews were conducted to collect qualitative data on participants' perceptions of AIG, which allowed for an in-depth exploration of the participants' viewpoints. The interviews were conducted until data saturation was achieved, indicating that further interviews did not provide significantly new insights beyond what was already mentioned by the participants. A total of 8 interviews were conducted during this phase, mainly focusing on three key aspects: participants' familiarity with AIG, their perceived benefits of using AIG in language assessment, and their perspectives on the challenges associated with the implementation of AIG in language assessment. These interviews were conducted and recorded through online meetings, using *Google Meet*.

2. Data Analysis

Analysis of the quantitative data was performed through *SPSS*, version 27. Within this phase, the study primarily focused on evaluating the quality of the generated test items by teachers. They rated the items on a scale of 1 to 4, as explained above. During this phase, one of the participants was identified as an outlier and was subsequently excluded from the analysis to prevent skewing the results. The mean rating from the remaining 29 participating teachers was 3.58 out of 4, suggesting an overall positive perception of item quality. Additionally, the standard deviation of 0.29 indicated that the ratings were closely clustered around the mean. This implies general agreement among the participants regarding the quality of the items.

The skewness and kurtosis indices presented values of -0.54 and -0.50 respectively. The negative skewness of -0.54 implied that the distribution of ratings was somewhat skewed towards the higher end of the scale (Larson-Hall, 2015), suggesting that more participants rated the items as ‘acceptable’ or ‘needing minor revisions’ rather than the items being ‘unacceptable.’ Similarly, the kurtosis index of -0.50 indicated a uniform and balanced spread of teacher ratings, reflecting a consistent perception among the teachers regarding the quality of test items.

This data collectively indicated a positive inclination towards the quality of the generated items. Teachers predominantly assessed the questions as either acceptable or needing only minor revisions, with a stronger inclination towards the items being acceptable with no need for revision at all. In Table 2, the summary statistics of the quantitative analysis are presented, showing the mean, standard deviation, skewness, and kurtosis values obtained from the teachers' ratings of the generated test items. Additionally, Figure 1 illustrates the distribution curve, depicting the spread of these ratings among the participating teachers.

	N	Mean	Standard Deviation	Skewness	Kurtosis
WF	29	3.58	0.29	-0.54	-0.50

Table 2: Summary of quantitative data analysis results

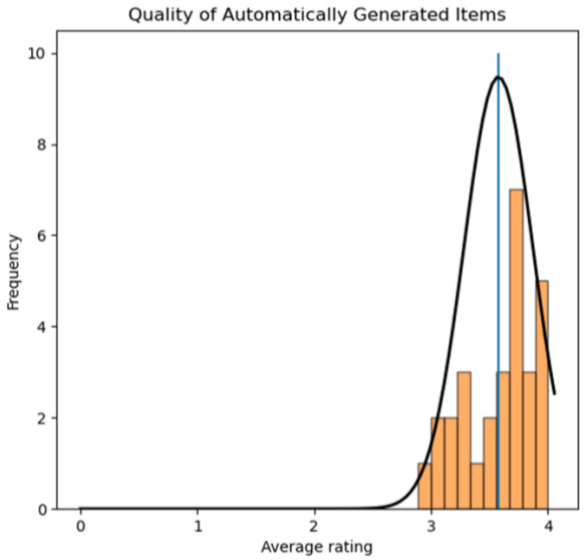

Figure 1: Distribution curve illustrating quantitative results

On the other hand, the analysis of the qualitative data, carried out using MAXQDA 2022 software, revealed distinct themes surrounding the benefits and challenges associated with AIG. Among the benefits highlighted by the teachers, the first was ‘efficiency and time optimization.’ Teachers emphasized that AIG facilitates the otherwise time-consuming process of question generation for both educators and test developers. The next benefit was ‘increased assessment frequency and focus,’ as the participants emphasized that AIG’s ability to rapidly generate questions allows for more frequent assessments throughout the school year. Another significant identified benefit was ‘fairness and objectivity in assessments.’ Participants highlighted AIG’s role in creating standardized assessments regardless of teacher biases or preferences. All in all, these advantages highlight how AIG contributes to elevating both the quantity and quality of assessments.

Regarding the use of AIG in language assessment, the primary challenge mentioned was ‘adapting to student levels and personalization,’ which stemmed from the fact that AIG systems typically follow a standardized approach. Consequently, the generated questions may not align with the individualized needs of students. ‘Diversifying question types’ was also considered a challenge, due to limitations of automated systems in generating a diverse range of question types. Finally, there were concerns around ‘cheating and security’ in case of computer-based delivery of the test. This concern arose from the possibility that students could gain access to question-generation algorithms, potentially risking the security of assessments.

3. Conclusions

In our exploration of the data, quantitative analysis of the data revealed positive ratings (mean=3.58 out of 4) regarding the quality of word family-based automatically generated items. Moreover, interview findings highlighted AIG's efficiency in facilitating the question generation and assessment processes. The combination of this quantitative and qualitative evidence underlines the benefits of AIG in facilitating and accelerating the process of question generation.

Teachers’ positive perception of the AIG system not only confirms its potential but also positions it as a useful tool in AI-driven language assessment and item development. The capacity of AI technology to analyze extensive text corpora and create items suited to specific lexical units and difficulty levels can significantly enhance the quality of assessments. Educators can benefit from these advancements to enhance their utilization of AI-aided item generation effectively.

Moving forward, we aim to diversify question formats beyond conventional multiple-choice and fill-in-the-blank formats. Personalization is another area of focus, aiming to adapt AIG to accommodate various student proficiency levels and diverse learning needs. In addition, exploring alternative lexical units beyond word families will help us expand the scope for more versatile item generation techniques. Finally, our aim is to improve and refine the distractors, to enhance the overall quality of assessments.

References

Alves, C. B., Gierl, M. J., & Lai, H. (2010). Using automated item generation to promote principled test design and development. *American Educational Research Association, Denver, CO, USA.*

Bauer, L., & Nation, P. (1993). Word Families. *International Journal of Lexicography*, 6(4), 253-279. https://doi.org/10.1093/ijl/6.4.253

Fulcher, G. (2013). *Practical language testing*. Routledge.

Gierl, M. J., & Haladyna, T. M. (2013). Automatic item generation: An introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 13-22). Routledge.

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation.* Routledge.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R.* Routledge.

Contact emails: susanmarandi@alzahra.ac.ir
sh.hosseini01@gmail.com