A Writing Support System that Scaffolds Language Learners via Autocompletion with Difficulty Prediction

Yo Ehara[1]

[1] *Tokyo Gakugei University, Koganei, Tokyo, 1848501, Japan*

Abstract

To support foreign language learners, we propose a system that automatically completes the learner's writing in a foreign language. The proposed system completes the continuation of the learner's input using a language model. The learner writes a part of a sentence using our system. Our system provides automatic completion by showing the candidate expressions, each of which completes the sentence. Among the candidates, another machine-learning classifier predicts the candidates that the learner is unlikely to know in a learner-adaptive manner. The system scaffolds each learner by showing the candidate expressions that the learner is unlikely to come up with during writing. In experiments, we show qualitative results.

Keywords

Autocompletion, second language learning, reading and writing.

1. Introduction

The complexity and intricacy of language learning necessitate a carefully orchestrated approach, particularly concerning the acquisition and application of appropriate vocabulary in a second-language context. Vocabulary knowledge plays a crucial role in expressing ideas effectively and communicating proficiently. However, due to the diverse abilities and levels of language learners, as well as the range of unfamiliar words they encounter, developing a universally effective method of support poses unique challenges.

This study addresses the aforementioned challenges by proposing an innovative approach that leverages artificial intelligence (AI) to facilitate second- language vocabulary acquisition. While previous studies and traditional teaching methods have offered some support in clarifying grammatical constructs, such as prepositions, there is a substantial gap in providing targeted assistance for vocabulary building. The primary difficulty lies in accurately predicting the specific words that individual learners struggle with when writing compositions, which necessitates extensive data collection and analysis.

In the field of second language acquisition, two key concepts are recognized: receptive vocabulary and productive vocabulary [7]. Receptive vocabulary refers to words that learners can understand and recognize when reading or listening, while productive vocabulary encompasses words that learners can accurately use when speaking or writing. Generally, an individual's receptive vocabulary is larger than their productive vocabulary, as they cannot productively use a word that they do not recognize or understand receptively.

Building on this premise, this study hypothesizes that augmenting a learner's receptive vocabulary would subsequently enhance their productive vocabulary. To achieve this, we propose an AI-based support system that emphasizes targeted vocabulary building through a reading vocabulary test. This AI system employs machine learning methods to evaluate a learner's receptive vocabulary, identify gaps, and predict potential areas of difficulty in their productive vocabulary.

Workshop on Artificial Intelligence in Support of Guided Experiential Learning, Held in conjunction with the International Conference on Artificial Intelligence in Education (AIED) 07 July 2023, Tokyo, Japan
EMAIL: ehara@u-gakugei.ac.jp (A. 1)

The proposed system, which involves a classifier and a masked language model, operates as follows: First, the learner is asked to take a vocabulary test that requires them to determine the meaning of words in a sentence. This test typically takes approximately 30 minutes.

Next, the data from this test are used to train or fine-tune a machine-learning classifier that allows for learner-adaptive prediction. This model determines whether a given expression is difficult for the learner in a learner-adaptive manner. It is important to note that productive vocabulary is nearly always a subset of receptive vocabulary. Therefore, if the classifier determines that an expression is difficult (i.e., not part of the learner's receptive vocabulary), it is highly likely that the expression is also absent from the learner's productive vocabulary.

Our system functions as follows: Using the system, the learner writes part of a sentence, and the system presents candidate expressions, each of which can complete the sentence. This automatic completion is conducted by a typical masked language model, which simply predicts the next word based on the learner's input. This masked language model enables the learner to explore potential candidates for the rest of the sentence. Then, the previously mentioned classifier for determining difficult words for the learner identifies candidates that are particularly difficult but important for the learner. The learner can discover expressions that complete the sentence, especially those with which they are un- familiar. Moreover, the system scaffolds the learner by continuously displaying important candidates predicted to be unfamiliar to them.

To evaluate this study, ideally, we require large datasets consisting of learners' partial sentences to be autocompleted. However, to the best of our knowledge, such a learner corpus does not exist, making rigorous evaluation difficult. In this presentation, as a preliminary study, we will showcase the system's construction, provide a demonstration of how it works, and present qualitative results from several examples. Finally, we present a qualitative evaluation of the proposed system.

1.1. Related Work

Several studies have been proposed to support second language writing using neural language models [6, 1]. However, to the best of our knowledge, no previous study proposed a system that combines the personalized prediction of each learner's vocabulary and writing.

An intelligent reading support system with the personalized prediction of each second language learner's vocabulary was previously proposed [5]. However, their paper does not deal with writing support.

2. System Overview

Figure 1 shows an overview of the proposed system. Figure 1 shows that the system is composed of two machine-learning based modules: "personalized classifier" and "LM" (a language model). These two modules are independent. First, in step (a), each user takes a 30-minute vocabulary test [4, 2] and submits their results to the personalized classifier. The submitted results from multiple users are collected and used as training data for training the personalized classifier to make predictions. When provided with a user ID and a word not included in the vocabulary test, the personalized classifier predicts the probability of the user knowing the word.

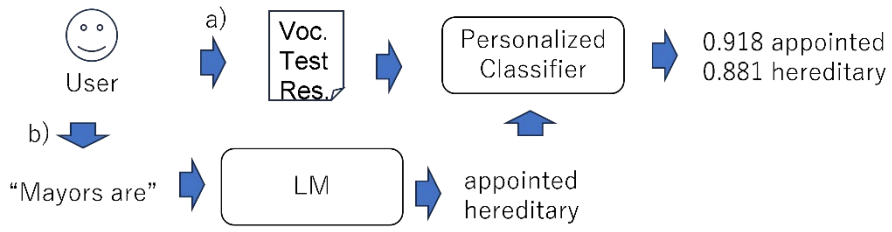

Figure 1. Overview of the proposed system

Next, in step (b), each user submits a half-written text to the system. The system forwards it to the LM module for auto-completion of the input. The LM module generates a ranked list of the top-k candidate words to complete the half-written input. Then, each word in the ranking of the completed text is input to the personalized classifier. The personalized classifier predicts the probability that the user knows the word. By examining the words that complete the input, the system can identify which expressions are likely to be unfamiliar to the user. In the provided example, each of the words "appointed" and "hereditary" is a suitable continuation for the phrase "Mayors are,". However, the learner is predicted to be unfamiliar with "hereditary".

Figure 2 is the actual screenshot of the proposed system shown in Figure 1. Two text boxes are shown. The upper one displays the half-written input, while the bottom one displays the auto-completed results. In the bottom box, each line consists of three components. The first value is a score that is returned by the "topk" function of the transformers library and indicates how well the completed word suits the input. The second value indicates how likely the user is to know the completed word. Finally, the third is the completed input.

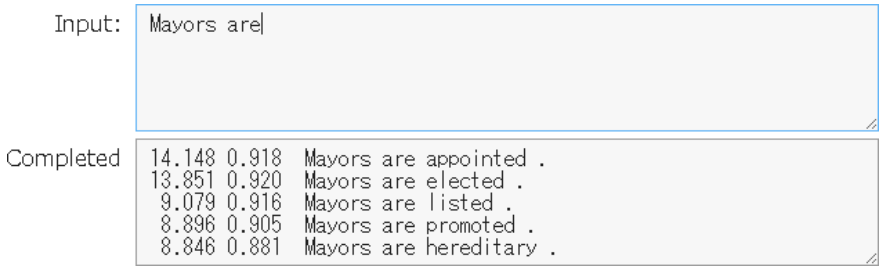

Figure 2. The screenshot of the proposed system

In Figure 2, we can see that the word "appointed" is predicted to be most likely to complete the input. The word "hereditary" is also considered appropriate as an expression that completes the input. However, it is not likely to be known to the learner as shown in Figure 2.

In fact, although the majority of the data in [4] are from Japanese learners of English, according to the Weblio (Can be seen from https://ejje.weblio.jp/), "appoint" is the word that may be under- stood by an English learner achieving a TOEIC (Test of English for International Communication, https://www.iibc-global.org/english.html) score of 350, whereas "hereditary" is the word understood by a learner of a TOEIC score of 860. This indicates that "hereditary" is significantly more difficult than "ap- point".

3. Implementation Details

We built the personalized classifier in Figure 1 using the vocabulary-test result dataset [4]. This dataset contains the vocabulary test results of 100 words and 100 learners. In this paper, we assumed that one of the 100 learners of this test used the system. This learner achieved a good score in the dataset of [4] and is assumed to be able to read intermediate texts. We also followed [4] for building personalized classifier using logistic regression based on word frequency features from British National Corpus (BNC) and Corpus of Contemporary American English (COCA) corpora (https://www.wordfrequency.info/).

For the LM module, most language models can be used. We used the BERT (Bidirectional Encoder Representation From Transformer) model [3] for building this module. For the pre-trained model, we simply used "bert-base-uncased". We predicted the word that completes the half-input text by adding "[MASK] ." token to the end of the inputted text. "[MASK]" is the token that used to denote the masked tokens in the BERT models. "." after "[MASK]" was necessary because otherwise the language model predicts tokens that finish sentences like "?", "!", and ".".

While, in this paper, we used BERT, which is a masked language model, we can also use other types of language models, such as large language models (LLMs), including ChatGPT (https://chat.openai.com/). Most language models, including masked language models and causal language models, can be used as "LM" in Figure 1 because the model simply has to predict the next

word given a context. Hence, also, no fine-tuning is required in our system. Hence, a simple call of ChatGPT can also be used as “LM” in Figure 1.

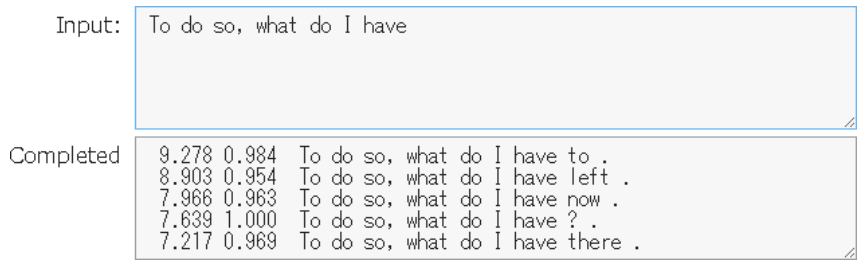

Figure 3. The system understands an idiom

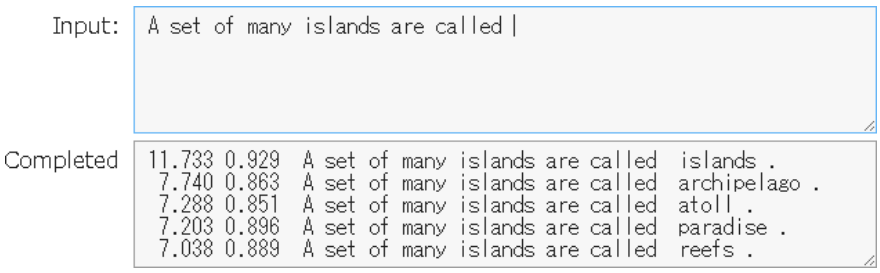

Figure 4. The system can be used to query a difficult word

4. Examples and qualitative results

Here, we show some examples to qualitatively analyze the proposed system. Figure 3 shows an example to show that the proposed system can understand basic English grammar and can handle the idiom “have to”. Various nouns could be used as objects for the verb “have”, such as in “do I have time to do that?”. However, in the example of Figure 3, the noun for the object is already filled in due to the presence of “what”, so nouns such as “time” cannot be the object of the verb “have”. Therefore, in Figure 3, a preposition or adverb correctly comes after the word “have”. Especially, due to the phrase “To do so”, the most likely word to complete the input is correctly predicted as “have to”. This result implies that the language model can scaffold the learner by correctly adding the word “to” when the learner does not remember that the preposition “to” is necessary to complete the idiom.

Figure 4 shows an example in which the system is used to find a difficult expression. Here, the definition of the word is specified in the phrase “A set of many islands are called”. Then, the language model returns a list of the word that is suitable for this definition. Here, the word “archipelago” comes the second of the list. According to the Weblio dictionary, “archipelago” is very difficult word. This result shows that the system can scaffold the learner by correctly showing the word suitable for the definition even if the learner does not know the word that the learner wants to express and used the other plain expression to indicate the word.

5. Conclusion

In this paper, we showed that a writing support system that scaffolds language learners. By a simple combination of two machine-learning models, namely a language model and a personalized classifier, the proposed system can auto- complete the second language learner’s input. The system can also output how likely the expression is known to the learner. Through the qualitative analysis, we showed that our system can effectively scaffold the learner by showing the word that the learner may not be able to think of.

Future work includes more thorough user studies of the system. Another important future work includes the retraining/re-fine-tuning of the model for learning difficult expressions based on the learner’s choice among the candidates.

6. Acknowledgements

This work was supported by JST ACT-X, Grant Number JPMJAX2006, Japan.

7. References

The references should be formatted according to the following gudelines: A paginated journal article [2], an enumerated journal article [3], a reference to an entire issue [4], a monograph (whole book) [5], a monograph/whole book in a series (see 2a in spec. document) [6], a divisible-book such as an anthology or compilation [7] followed by the same example, however we only output the series if the volume number is given [8] (so series should not be present since it has no vol. no.), a chapter in a divisible book [9], a chapter in a divisible book in a series [10], a multi-volume work as book [11], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [12], a proceedings article with all possible elements [13], an example of an enumerated proceedings article [14], an informally published work [15], a doctoral dissertation [16], a master's thesis: [17], an online document / world wide web resource [18, 19, 20], a video game (Case 1) [21] and (Case 2) [22] and [23] and (Case 3) a patent [24], work accepted for publication [25], prolific author [26] and [27]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [28]. Multi-volume works as books [29] and [30]. A couple of citations with DOIs: [31, 28]. Online citations: [32, 18, 33, 34].

[1] Arau´jo, S., Aguiar, M., Monteiro, J.: A bert-powered writing assistant for academic purposes in european portuguese. In: Perspectives and Trends in Education and Technology: Selected Papers from ICITED 2022, pp. 513–520. Springer (2023)
[2] Beglar, D., Nation, P.: A vocabulary size test. The Language Teacher 31(7), 9–13, (2007)
[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidi- rectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/ N19-1423
[4] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English- as-a-Second-Language Learners Using Crowdsourcing. In: Proc. of LREC (May 2018).
[5] Ehara, Y., Shimizu, N., Ninomiya, T., Nakagawa, H.: Personalized Reading Support for Second-language Web Documents. ACM Trans. Intell. Syst. Technol. 4(2), 31:1– 31:19 (Apr 2013). https://doi.org/10.1145/2438653.2438666, http://doi.acm.org/ 10.1145/2438653.2438666.
[6] Narimatsu, H., Koyama, K., Dohsaka, K., Higashinaka, R., Minami, Y., Taira, H.: Task definition and integration for scientific-document writing support. In: Proceed- ings of the Second Workshop on Scholarly Document Processing. pp. 18–26 (2021).
[7] Nation, I.: How Large a Vocabulary is Needed For Reading and Listening? Canadian
[8] Modern Language Review 63(1), 59–82 (Oct 2006).