GazeReader: Detecting Unknown Word Using Webcam for English as a Second Language (ESL) Learners

Jiexin Ding*
jxding17@gmail.com
Global Innovation Exchange
Tsinghua University
Beijing, China

Bowen Zhao*
bowen98@uw.edu
Global Innovation Exchange
University of Washington
Seattle, United States
Tsinghua University
Beijing, China

Yuqi Huang*
huangyuqi1999@gmail.com
Global Innovation Exchange
University of Washington
Seattle, United States
Tsinghua University
Beijing, China

Yuntao Wang†
yuntaowang@tsinghua.edu.cn
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Yuanchun Shi
shiyc@tsinghua.edu.cn
Department of Computer Science and Technology
Tsinghua University
Beijing, China

ABSTRACT

Automatic unknown word detection techniques can enable new applications for assisting English as a Second Language (ESL) learners, thus improving their reading experiences. However, most modern unknown word detection methods require dedicated eye-tracking devices with high precision that are not easily accessible to end-users. In this work, we propose GazeReader, an unknown word detection method only using a webcam. GazeReader tracks the learner's gaze and then applies a transformer-based machine learning model that encodes the text information to locate the unknown word. We applied knowledge enhancement including term frequency, part of speech, and named entity recognition to improve the performance. The user study indicates that the accuracy and F1-score of our method were 98.09% and 75.73%, respectively. Lastly, we explored the design scope for ESL reading and discussed the findings.

CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**.

KEYWORDS

unknown words detection, webcam, eye tracking, natural language processing

ACM Reference Format:
Jiexin Ding, Bowen Zhao, Yuqi Huang, Yuntao Wang, and Yuanchun Shi. 2023. GazeReader: Detecting Unknown Word Using Webcam for English as a Second Language (ESL) Learners. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585790

*The authors contribute equally to this paper.
†denotes as the corresponding author.

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585790

1 INTRODUCTION

Unknown words, which we defined as words that the user spends extra time thinking about, can significantly increase the reading difficulty for English as a Second Language (ESL) learners [19, 26], since vocabulary knowledge is considered an essential feature of reading ability. Automatic detection of unfamiliar words can help users improve reading fluency and reading experience [14]. Previous works used eye movement features such as fixation duration and the number of regressions to detect unknown words, considering the strong correlation between eye movement and reading difficulty [7, 9, 18]. However, all these methods are based on dedicated eye-tracking devices. Even for portable eye trackers, prices of hundreds or thousands of dollars prevent most people from using these technologies in their daily lives.

The ubiquity of webcams makes it possible to track eye movement non-obtrusively. Researchers have used webcam-based eye-tracking methods to analyze users' reading behavior [13], but the low precision [21] makes it infeasible to extract eye movement features proposed by previous works. Therefore, these reading analysis approaches cannot be directly transferred to webcam-based unknown word detection, which requires fine-grained tracking.

In this work, we detect unknown words using the webcam for ESL learners by integrating gaze and text information with the help of transformer-based language models. Our method tracks eye movement using WebGazer [22] and embeds the positional information of gaze and texts with Long Short-term Memory [10] (LSTM) models. In order to improve the model's performance given only noisy gaze data acquired by webcam, we leverage pre-trained language models to encode the context information for assistance. The gaze and context data are then fed to the model's classifier to predict the targeted unknown word. The F1-score of our model

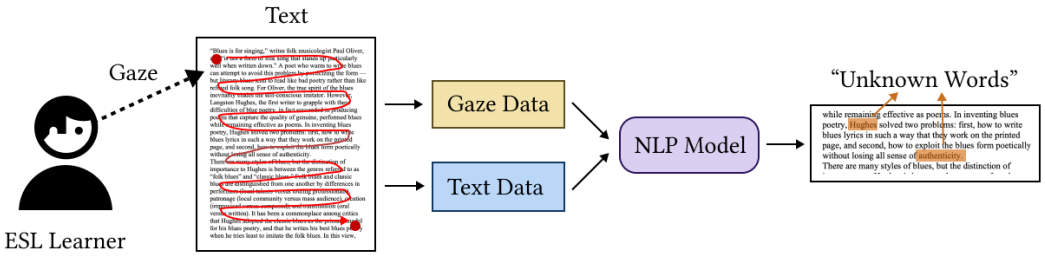

Figure 1: GazeReader collects gaze data and text information and uses a transformer-based machine learning model to detect unknown words for ESL.

is 75.73% which is higher than the text- and gaze-only model. We also conduct a user study to discover the needs of ESL learners and explore the design scope according to our method. The results show that our future design should focus more on proper nouns, multi-meaning words, and long and complex sentences, as well as exploring ways to reduce disturbance to the user's reading process through interactive design. In summary, the main contributions of this work include the following:

1) We propose GazeReader, a novel webcam-based unknown word detecting method that leverages gaze and text information for ESL learners. The accuracy is 98.09%, the F1-score is 75.73%, and the cross-user F1-score is 78.26%.

2) We explore the design scope for ESL reading based on our method, guiding our future design direction towards proper nouns, multi-meaning words, and long and complex sentences.

2 BACKGROUND AND RELATED WORK

It is time-consuming for second language learners to find the meaning of unknown words by looking up dictionaries or using translation tools. Automatically detecting unknown words can enable a variety of ways to facilitate reading. Researchers have implemented automatic unknown word detection by tracking users' clicks on a web document [6] or combining text features with motion data on a smartphone [8]. In addition, other works used eye movements to detect unknown words. The physiological and psychological reason behind it is that eye movements largely reflect people's tendency to pay attention when reading [23]. Word length, word frequency, and familiarity with the word largely explain the length of time users spend looking at it, in other words, longer fixation time [4, 15]. Mostly related to our work, [7] and [9] leverage features from the gaze and document context to detect difficult words in reading. However, the above methods are all based on eye movement features, such as the number of fixation and regression, so dedicated eye tracking equipment is required to obtain high-precision eye movement data.

The webcam is a ubiquitous device that can track users' eye movements. Researchers apply calibration-free webcam-based eye tracking to predict reading comprehension and mind wandering in reading and achieve comparable accuracy to infrared-based eye tracking on these tasks [13]. Previous work also utilizes fixation points computed from webcam data to analyze search behaviors, and the mean error of the real-time eye tracking is 128.9 pixels [21]. Although webcam-based eye tracking shows the potential of modeling user behavior, the tracking precision is not enough for detecting unknown words, which requires precise locations of the gaze.

Based on the strong connection between eye movement and attention in reading, many works introduce gaze to assist model training in Natural Language Processing (NLP) area [1, 11, 17]. Since pre-trained language models [5, 16] contain rich text information for downstream tasks, we leverage them to compensate for the inaccuracy of the webcam-based eye-tracking method. By combining gaze and text information, we can accurately detect unknown words for users based on a webcam, which can enable various interactions assisting ESL reading.

3 GAZEREADER

GazeReader automatically detects unknown words by embedding the gaze data and text information using LSTM and a transformer-based model. We also introduced knowledge-based features including term frequency, part of speech, and named entity recognition to improve the performance. In data pre-processing, we apply a moving average filter and re-sampling to de-noise gaze data and map the gaze data to the text data.

3.1 Data Preparation

3.1.1 Data Collection. We recruited 12 graduate students ranging in age from 23 to 28 years (M = 23.64, SD = 1.57), including two males and ten females. They were all second-language learners of English and were able to use English in academic settings, and their years of formal English study ranged from 7 to 24 years (M = 16, SD = 4.84). Nine of them wore glasses, and the other three did not.

We chose 36 articles from TOEFL reading materials with an average length of 386 words per article and 13902 words in total. Participants read 12 articles per day for three days and spent four hours reading in total. The experiment was conducted in a quiet meeting room without disturbing using a Thinkpad X1 carbon laptop (CPU: i7-10710U, 6 cores, 1.1 GHz, RAM: 16GB, Storage: 512GB). There was a break every 30 minutes.

We built a web PDF reader to display the article and used WebGazer.js [22] to track users' gaze. Before each section started, the participants calibrated the WebGazer according to its instruction [22] to map gaze points to screen coordinates. Participants first

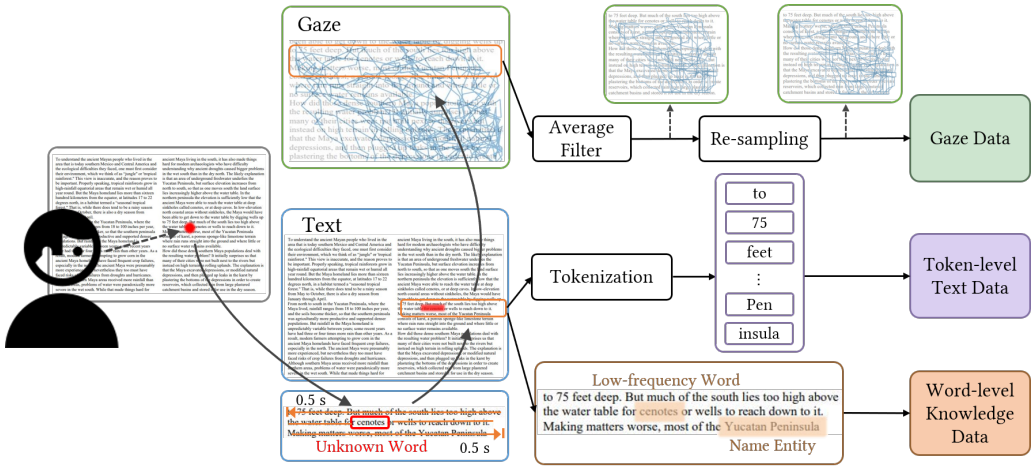

Figure 2: Data pre-processing for GazeReader model.

read the article, recorded the gaze data, and then clicked to mark the unknown words. Our system saved the gaze data, the unknown words, and the text coordinate for each article. We collected a total of 4274 clicks for unknown words and 1232 unknown words after eliminating duplication.

3.1.2 Data Pre-processing. Since the gaze data is acquired with only a webcam without highly accurate devices, the original data is extremely noisy and could not be directly fed to the machine learning models. Meanwhile, to avoid potential distractions to users' reading, we do not ask the users to label their unknown words during their reading in real-time, which makes it a necessary step to align the gaze data with its corresponding context and unknown words.

As shown in Fig. 2, we first de-noise the gaze data by applying a moving average technique with a sliding window size of 50. Since each time step of the output data from WebGazer is uneven, we re-sample the gaze data with the sample rate of 20Hz and conduct linear interpolation between the raw data on both the x and y-axis. Afterward, to align each unknown word with its corresponding gaze data, we anticipate each word's time based on its relative position of the whole document assuming that people's reading speeds are likely to be uniform. We also chunk the gaze data into segments based on each piece of text's length for matching gaze and text data. We select the text read within 1 second by the user when they see the unknown word as its context. Finally, since we only want to prompt users with the unknown words they are reading, we add negative samples for each gaze data by combining it with the contents containing unknown words that the user was not reading.

3.2 Unknown Words Detection Model

Our model is designed to predict users' unknown words during reading based on gaze data and reading text information. To capture the positional information of texts and gaze data, we use LSTM layers to encode them into vector space. To improve the model's performance, we use RoBERTa [16] as the backbone language model for infusing text information. Meanwhile, we further enhance the text information by leveraging word-level knowledge. The overall architecture of our model is illustrated in Fig. 3.

3.2.1 Positional Data Encoding. At first, we illustrate our neural network that encodes the sequential 2-D data, including the word position in the reading document and the user's eye gaze data. Since both word position and eye gaze data are sequential, we use LSTM network to encode them since it can capture the long and short-term dependencies in sequential data. Overall, the two LSTM layers transform the 2-D gaze sequence $g \in \mathbb{R}^{2\times n_{gaze}}$ and word position sequence $d \in \mathbb{R}^{2\times n_{txt}}$ into $H_g \in \mathbb{R}^{n_p\times n_{gaze}}$ and $H_d \in \mathbb{R}^{n_p\times n_{txt}}$ (where n_p is the output dimension of LSTM layer, while n_{gaze} and n_{txt} are the length of the gaze data and texts), respectively. Afterward, we conduct a matrix multiplication between the encoded text positional data and the gaze data as the attention between the gaze and texts:

$$g = [g_x; g_y], H_g = \text{LSTM}(g)$$
$$d = [d_x; d_y], H_d = \text{LSTM}(d)$$
$$A_p = \delta(H_d^T \cdot H_g)$$

where δ is the activation function and g_x, g_y, d_x, d_y are the gaze data and word position on X and Y axis, respectively.

3.2.2 Context Information Capturing. In order to leverage the rich information from the reading materials, we use a pre-trained RoBERTa model to encode these texts in the vector space. RoBERTa is a pre-trained language model with 12 transformer layers, where each layer contains a self-attention module that captures the attention between word tokens and a feed-forward layer that maps the attention into a higher vector space. It transforms the input texts $s \in \mathbb{R}^{n_{txt}}$ into $Z \in \mathbb{R}^{n_{txt}\times dim}$, where dim is the hidden dimension of RoBERTa. Since the model has been trained on huge amounts of text data, it can accurately embed documents into vector space for downstream tasks. In general, RoBERTa calculates

$$Z = \text{RoBERTa}(s) \quad (1)$$

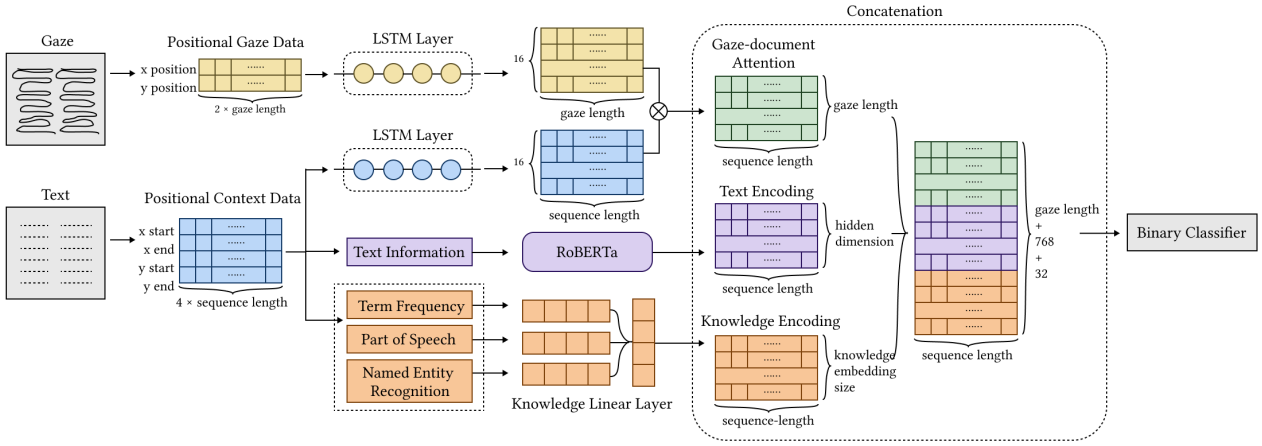

Figure 3: Overview of GazeReader model.

3.2.3 Knowledge Enhancement. During the processing of language models, the original documents are tokenized into sub-word tokens instead of words. Therefore, there exists word-level knowledge lost during tokenization which causes the degradation of model performance. Hence, for each token, we add its original word-level knowledge, including term frequency, part of speech [2], named entity recognition [12] to promote the model's performance. In general, we use learnable embeddings and layers to encode these features into a "knowledge matrix" $K \in \mathbb{R}^{n_{txt} \times n_k}$.

3.2.4 Training. We combine these features above and feed them to a linear classifier for unknown word prediction:

$$O = [A_p; Z; K]$$
$$a = \sigma(W_o \cdot O + b_o)$$

where W_o, b_o are learnable parameters, σ is the sigmoid function, and $a \in \mathbb{R}^{n_{txt}}$ is the output activation of the classifier. The model is trained with a binary entropy loss on all tokens

$$\mathcal{L} = -\sum_{i=1}^{n_{txt}} (1 - \tilde{y}_i) \cdot \log(1 - a_i) + \tilde{y}_i \cdot \log(a_i) \quad (2)$$

where $\tilde{y}_i$ is the label whether the i-th token is part of the unknown word of the user.

3.3 User Study

To demonstrate the variability of our approach, we used the Vocabulary Levels Test (VLT) to test users' vocabulary levels. The test was developed by Nation in 1987 and produced questions at different word frequency levels in 2000, 3000, 5000, 10,000, and academic groups [20]. The VLT is one of the most widely used tests to measure the vocabulary level of English learners [24, 25]. We chose an optimized version of the test [27]. Considering that the required vocabulary level for TOEFL reading as experimental material is about 4500 [3], we used a 5000-word frequency level group to test users' vocabulary levels.

We also explored future directions for the tool design. In order to understand users' needs, we designed a group of questions to test their subjective feelings and attitudes. The questions include perceived factors that influence English text reading, behaviors when checking unknown words in reading, concerns toward eye-tracking tools, and attitudes toward our proposed new features (eye-tracking reading assistance and vocabulary management). All the above questions and potential answers were grouped and conducted using 5-point Likert items.

Besides subjective ratings, we collected descriptive information from randomly selected users about their needs during reading. Users were asked to write down their potential needs as much as possible.

4 RESULTS AND FINDINGS

4.1 Prediction Accuracy

The main results of our model are shown in Table 1. We use precision, recall, and F-1 score as metrics for our unknown word detection task. We trained our models for 5 epochs, with the learning rate of 8e-5 for parameters in the RoBERTa and linear layers, and 0.1 for LSTM layers. Our best model can achieve the accuracy of 98.09% with 75.73% F-1 score. Moreover, we tried out different hidden dimensions for gaze-text attention n_p and knowledge embedding n_k. It shows that by increasing the knowledge dimension n_k, the overall performance of our model can be improved while increasing the gaze-document attention dimension n_p helps promote the recall.

PLM Backbone	n_p	n_k	Precision	Recall	F-1 Score
RoBERTa	16	16	68.31	79.20	73.97
RoBERTa	16	32	**71.21**	80.70	**75.73**
RoBERTa	32	16	67.29	84.81	75.11
RoBERTa	32	32	65.50	**86.55**	74.97

Table 1: Unknown word detection results of GazeReader.

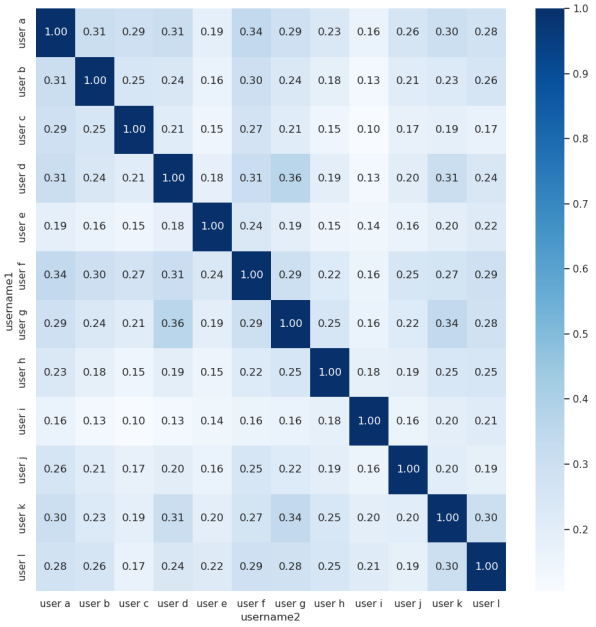

Figure 4: Jaccard similarity of unknown words between users.

4.2 Transfer Learning Analysis

4.2.1 Transfer Learning Setting. Table 2 indicates our model's transferability between users and reading materials. For the cross-user transfer learning experiment, we trained the model 12 times in total, and for each running, we selected one user's data as the test set and took other users' data as the training set. For the cross-document experiment, we also conducted 12 pieces of training and evaluations. Under each running, we selected the data generated from three documents as the test set, and the rest of the data were combined as the training set. Therefore, our relative data sizes between training and testing data are similar under cross-user and cross-document settings.

4.2.2 Cross-User Transferability. We find out that when the model is applied to new users' data, its performance is still on par with its original setting. Therefore, our model can be effectively adapted to new user groups without specific calibrations or tuning, which successfully fits our goal of ubiquitous unknown word detection. We further analyze the unknown words of different users, and their Jaccard similarity matrix is shown in Fig 4. The average of cross-user unknown words Jaccard similarity score is 0.23 and most of their similarities are below 0.3, meaning that different users' unknown words are distinct even given the same article. It further proves that our model does not simply remember the unknown words of users, but predicts the unknown words based on the user's behaviors. Meanwhile, the model's performance can become significantly better on some users, and a possible reason is that these users' behaviors are more general, i.e., their language skills might be the average among all people.

4.2.3 Cross-Document Transferability. In order to test whether our model only memorizes the difficult words from the text, we conducted the experiment to see how the model would perform if the words in the test data were not presented for training. We find that the model's performance dropped in this setting, while the recall is still acceptable.

4.3 Ablation Study

We tried to exclude three different modalities of data presented in our model respectively, which are text information, gazing data, and word-level knowledge. After the text information is removed, the model's performance drops significantly, which proves that the use of a pre-trained language model can effectively promote the model's ability to detect unknown words of users. Then we removed the gaze data, in which case the model cannot obtain the gaze behavior related to the text. The model's precision drops and the recall increases. It indicates that both the true positive and the false positive increase. In other words, without gaze, the model tends to predict a word as an unknown word. The gaze may help identify whether a difficult word is an unknown word or not. Finally, the result of ablating knowledge enhancement proves that infusing word-level knowledge can promote the model's performance.

4.4 Design Guidance

Our results of the 5000-level VLT show that users can be divided into a high-level group and a low-level group. A Mann-Whitney U test was run to determine if there were differences in the test score between them. Scores for the high-level group (mean rank = 9.5) and low-level group (mean rank = 3.5) were statistically significantly different, U = .000, z = -2.887, p = .002, using an exact sampling distribution for U. The results demonstrate that our users covered learners of different levels.

As for factors affecting the reading of academic texts, compared to a neutral attitude score of 3, the scores were 3.67 for new words (SD = 0.98, $t(11)$ = 2.345, $p < .05$), 3.83 for the ambiguous meaning of words (SD = 0.72, $t(11)$ = 4.022, $p < .01$), 4.25 for long and complex sentences (SD = 0.87, $t(11)$ = 5.000, $p < .001$), all of which were perceived by users as having a significant effect on comprehension of the text. This result suggests the importance of contextualizing the interpretation of unknown words in reading.

The scores of the two proposed features are 4.35 for eye-tracking reading assistance (SD = 0.64, $t(11)$ = 7.244, $p < .001$) and 3.93 for vocabulary management (SD = 0.99, $t(11)$ = 3.260, $p < .01$), respectively. Both are significantly higher than a neutral score of 3, demonstrating that users hold a positive attitude toward our proposed features. The results show that none of the privacy and power consumption is rated significantly, which indicates the possibility of future tool design.

By analyzing words users wrote about expecting features, we found that the three most frequently mentioned features are translating proper nouns, translating long and complex sentences, and accurately identifying multi-meaning words. Two other users also mentioned the need to avoid pop-ups from affecting their reading. We believe these goals can all be achieved using our webcam-based tool and NLP model.

Based on the results mentioned above, applying the webcam to unknown words detecting and reading assistance by our approach of eye tracking and NLP model has theoretical and practical value.

User dependency	Document dependency	Precision	Recall	F-1 Score
Dependent	Dependent	71.21	80.70	75.73
Independent	Dependent	73.65 ± 5.83	82.77 ± 4.41	78.26 ± 4.53
Dependent	Independent	46.56 ± 4.26	68.39 ± 5.80	56.31 ± 3.38

Table 2: Transferability analysis of unknown word detection by GazeReader.

Model	Precision	Recall	F-1 Score
GazeReader	**71.21**	80.70	**75.73**
GazeReader w/o context-awareness	4.78	59.03	10.00
GazeReader w/o gaze	66.35	**86.67**	75.59
GazeReader w/o knowledge enhancement	69.96	80.32	74.93

Table 3: Ablation study of GazeReader.

Interaction design strategies should also be addressed in developing applications to improve the user experience.

5 DISCUSSION AND FUTURE WORK

The accuracy of our model is high by combining the information from the gaze, text, and knowledge, but the F1-score is only 75.73%. The reason could be that unknown words are sparse compared with other words. It causes an unbalance in the dataset. In addition, the size of the dataset is relatively small compared with the parameter in the neural network might be the other reason. More case studies are needed to find out the reason.

The ablation shows that the F1-score doesn't significantly decrease when we remove the gaze from the model. One of the potential reasons is the inaccuracy of the webcam-based eye-tracking algorithm we used. WebGazer [21] uses the coordinates of clicks to calibrate the position of gaze points, so tracking accuracy becomes worse when participants read without using a mouse. The optimization of the eye-tracking algorithm is needed to make gaze contribute more to the performance. Multi-modal data obtained by earphones and smart glasses can be used to infer head orientation and position [28, 29] to facilitate the eye-tracking algorithm.

Our data is collected on a commodity laptop using an eye-tracking library that can work on any website [22]. Therefore, our method should be able to generalize to other models of computers, though we only used one type of laptop to collect data in this paper. In addition to the current web-based implementation, the same algorithm should be able to migrate to various applications.

We took TOEFL reading passages as reading materials in the experiment, considering that unknown words in academic text create more difficulty in reading. The TOEFL readings cover a variety of topics, so our method has the potential to work on different types of articles. In the future, we will use more diverse materials to make our approach more generalizable. Participants in our experiment are all graduate students in the United States. We regard them as our target users because they have a more urgent need to read more efficiently. Our participants consisted of learners with different vocabulary levels, and we will also include more users of different learning stages in the future.

Applying our approach to languages other than English is another aspect of future work. We believe that this work may be generalized to other alphabetic languages because alphabetic languages have similar word structures and some of them even have the same alphabet. It remains to be seen whether gaze data can be used to detect unknown words in non-alphabetic languages.

For applications, future work could explore how to provide proper assistance to second language users during reading according to our exploratory user study. We aim to design a simple webcam-based tool across different levels of users that has high efficiency and usability at the same time. Exploring ways to properly present the information that users need would be a compelling future study.

6 CONCLUSION

We propose GazeReader, a novel unknown word detecting method using the webcam for ESL learners. GazeReader leverages the gaze data obtained from the webcam and introduces the transformer-based model to enhance the performance. GazeReader can detect unknown words with the accuracy of 98.09% and the F1-score of 75.73%. We also conducted a user study to explore the design scope based on GazeReader. The result shows that the design focus should be proper nouns, multi-meaning words, and long and complex sentences. Users show positive attitudes towards the design features based on unknown word detection. As the webcam becomes more ubiquitous with the prevalence of online working and studying, GazeReader can enable a wide range of applications and benefit a large population of ESL learners in reading.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (NSFC) under Grant No. 62132010 and No. 62002198, Young Elite Scientists Sponsorship Program by CAST under Grant No. 2021QNRC001, Tsinghua University Initiative Scientific Research Program, Beijing Key Lab of Networked Multimedia, Institute for Artificial Intelligence, Tsinghua University.

REFERENCES

[1] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence Classification with Human Attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Brussels, Belgium, 302–312. https://doi.org/10.18653/v1/K18-1030

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[3] Kiyomi Chujo and Kathryn Oghigian. 2009. How many words do you need to know to understand TOEIC, TOEFL & EIKEN? An examination of text coverage and high frequency vocabulary. *Journal of Asia TEFL* 6, 2 (2009).

[4] Charles Clifton Jr, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. *Eye movements* (2007), 341–371.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[6] Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. 2010. Personalized Reading Support for Second-Language Web Documents by Collective Intelligence. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 51–60. https://doi.org/10.1145/1719970.1719978

[7] Utpal Garain, Onkar Pandit, Olivier Augereau, Ayano Okoso, and Koichi Kise. 2017. Identification of Reader Specific Difficult Words by Analyzing Eye Gaze and Document Content. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01. 1346–1351. https://doi.org/10.1109/ICDAR.2017.221

[8] Riku Higashimura, Andrew Vargo, Motoi Iwata, and Koichi Kise. 2022. Helping Mobile Learners Know Unknown Words through Their Reading Behavior. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 249, 5 pages. https://doi.org/10.1145/3491101.3519620

[9] Rui Hiraoka, Hiroki Tanaka, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. 2016. Personalized Unknown Word Detection in Non-Native Language Reading Using Eye Gaze. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 66–70. https://doi.org/10.1145/2993148.2993167

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[11] Nora Hollenstein and Ce Zhang. 2019. Entity Recognition at First Sight: Improving NER with Eye Movement Information. In *North American Chapter of the Association for Computational Linguistics*.

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python. (2020).

[13] Stephen Hutt and Sidney K. D'Mello. 2022. Evaluating Calibration-Free Webcam-Based Eye Tracking for Gaze-Based User Modeling. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 224–235. https://doi.org/10.1145/3536221.3556580

[14] Aulikki Hyrskykari. 2006. *Eyes in attentive interfaces: Experiences from creating iDict, a gaze-aware reading aid.* Tampere University Press.

[15] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[17] Yunfei Long, Rong Xiang, Qin Lu, Chu-Ren Huang, and Minglei Li. 2021. Improving Attention Model Based on Cognition Grounded Data for Sentiment Analysis. *IEEE Transactions on Affective Computing* 12, 4 (2021), 900–912. https://doi.org/10.1109/TAFFC.2019.2903056

[18] Tobias Lunte and Susanne Boll. 2020. Towards a Gaze-Contingent Reading Assistance for Children with Difficulties in Reading. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 83, 4 pages. https://doi.org/10.1145/3373625.3418014

[19] Ahmad Azman Mokhtar and Rafizah Mohd Rawian. 2012. Guessing word meaning from context has its limit: Why. *International Journal of Linguistics* 4, 2 (2012), 288–305.

[20] Paul Nation. 1990. Teaching and learning vocabulary. In *Handbook of Practical Second Language Teaching and Learning*. Routledge, 397–408.

[21] Alexandra Papoutsaki, James Laskey, and Jeff Huang. 2017. SearchGazer: Webcam Eye Tracking for Remote Studies of Web Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 17–26. https://doi.org/10.1145/3020165.3020170

[22] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, 3839–3845.

[23] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

[24] John Read. 1988. Measuring the vocabulary knowledge of second langauge learners. *RELC journal* 19, 2 (1988), 12–25.

[25] John Read. 2000. *Assessing vocabulary*. Cambridge university press.

[26] Pat Rigg. 1991. Whole language in TESOL. *Tesol Quarterly* 25, 3 (1991), 521–542.

[27] Norbert Schmitt, Diane Schmitt, and Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing* 18, 1 (2001), 55–88.

[28] Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. FaceOri: Tracking Head Position and Orientation Using Ultrasonic Ranging on Earphones. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 290, 12 pages. https://doi.org/10.1145/3491102.3517698

[29] Yuzhou Zhuang, Yuntao Wang, Yukang Yan, Xuhai Xu, and Yuanchun Shi. 2021. ReflecTrack: Enabling 3D Acoustic Position Tracking Using Commodity Dual-Microphone Smartphones. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 1050–1062. https://doi.org/10.1145/3472749.3474805