Overview of the NTCIR-16 RCIR Task

Graham Healy*
Dublin City University
Ireland
graham.healy@dcu.ie

Tu-Khiem Le*
Dublin City University
Ireland
tukhiem.le4@mail.dcu.ie

Mai Boi Quach
Dublin City University
Ireland
mai.quach3@mail.dcu.ie

Minh-Triet Tran
University of Science, VNU-HCM
Vietnam National University
Vietnam
tmtriet@hcmus.edu.vn

Thanh-Binh Nguyen
University of Science, VNU-HCM
Vietnam National University
Vietnam
ngtbinh@hcmus.edu.vn

Cathal Gurrin
Dublin City University
Ireland
cathal.gurrin@dcu.ie

ABSTRACT

The NTCIR-16 RCIR pilot task aimed to motivate the development of a first generation of personalised retrieval techniques that integrate reading comprehension measures and eye tracker signals as a source of information when ranking text content. The dataset used in the challenge was newly generated by capturing eye movement measures while experimental participants read text passages on a computer screen. The RCIR challenge included two sub-tasks: a) the comprehension-evaluation task (CET) that involved predicting a measure of a reader's comprehension for text passages and, b) the comprehension-based retrieval task (CRT) that involved retrieving relevant passage texts ranked by comprehension score. The participating teams were ranked using Spearman's correlation coefficient (ρ) for the CET sub-task and normalised Discounted Cumulative Gain (*nDCG* score) for the CRT sub-task.

KEYWORDS

reading comprehension, eye movements, eye tracking, information retrieval (IR)

SUBTASKS

Comprehension-Evaluation Task (CET)
Comprehension-based Retrieval Task (CRT)
Insight Task (IT)

1 INTRODUCTION

Reading is a crucial skill as the vast majority of the knowledge of humankind is communicated in written form [16]. Although many are capable of reading [18], reading the same passage of text can often result in different levels of comprehension for different readers.

Previous efforts to estimate reading comprehension have mainly relied on a combination of one or more different assessment approaches that include interviews, questionnaires, oral retelling, freewriting and think-aloud procedures [9]. While the effectiveness of such assessment approaches is clear, these approaches are often only feasible when evaluation is carried out in a controlled setting e.g. where designed assessments for the material to measure comprehension are available. In more general scenarios where reading happens as part of daily activities (e.g. browsing websites on a computer, reading newspapers, joining seminars), it would be prohibitively burdensome to employ these techniques to measure comprehension in an individual. Hence, there is a need for automatic methods that allow reading comprehension to be estimated passively, using data from sensor sources that capture eye movements, facial expressions, and other biometrics.

There has been a growing body of research in computer science and cognitive psychology domains, investigating the relationship between eye movements and reading. There are a number of common eye movement behaviours that include fixations (moments when the eyes remain stable around a point of gaze), saccades (rapid eye movements between fixation points) and blinks [6, 15]. It has been shown that eye tracking systems can be used to estimate important measures related to reading. For example, Kunze et al [10] and Ishimaru et al [7] demonstrated algorithms that can be used to estimate the number of words read by an individual, Bulling et al. [1] developed an approach to recognise five classes of daily activities, and Yoshimura et al. [20] was able to estimate English reading proficiency skills.

The RCIR task involved a collaborative evaluation with the RCIR task participants[8, 12, 13], to explore how eye tracking signals can be used to estimate reading comprehension in individuals. In addition to exploring the prediction of reading comprehension measures, we also explore the integration of comprehension prediction models into the information retrieval process, as this could be highly beneficial in the development of personalised retrieval systems, especially for applications in lifelogging. Although the ultimate goal of lifelogging is to passively capture and form digital records of our life experience [5], the recent benchmarking dataset and challenges [3, 4] only emphasise the retrieval of visual data from lifelogs (i.e. images), whilst the indexing and retrieval of text that one has seen and comprehended in a lifelogging context has not been actively investigated as of yet.

In this paper, we describe the RCIR task at NTCIR-16, where a novel multimodal reading dataset was created to support a collaborative evaluation exercise to benefit the IR community e.g. in understanding the incorporation of reading comprehension measures into document ranking systems in retrieval tasks. This dataset was collected from experimental participants who followed a pre-defined protocol while their eye movements were measured, carrying out reading tasks under different reading constraints and manipulations (e.g., reading, skimming, scanning, and proofreading), and completing comprehension measurement assessments.

*Two authors contributed equally to this research.

In the NTCIR-16 RCIR task, the participating teams developed and benchmarked their approaches to integrate multimodal signals (e.g., eye tracking data and text content) into the retrieval process for two sub-tasks:

- *A comprehension-evaluation task (CET)* that aimed to predict a person's comprehension level for various passages.
- *A comprehension-based retrieval task (CRT)* that aimed to retrieve and rank passages (on a variety of topics) by integrating text comprehension-evidence into the IR process.

Both of these sub-tasks were exploratory in nature, and designed to facilitate initial experimentation on the topic by the community.

2 DATASET

The dataset for RCIR was collected by recording the eye movements of experimental participants as they read a number of pre-defined passages (text excerpts) on a computer screen. After reading each passage, a set of MCQs (multiple-choice questions) were presented to experimental participants that measured their comprehension [2]. The MCQ responses for each passage were recorded and processed to calculate a comprehension score corresponding with the eye tracking measures for each passage. The details of the protocol, experimental participants, materials, data sources and the dataset's structure are provided below. Data collection was carried out with approval from Dublin City University's Research Ethics Committee (DCUREC / 2021 / 138).

3 DATASET COLLECTION

There were a total of $N = 9$ experimental participants ($Male = 4$, $Female = 5$) recruited for data collection, 5 of which were non-native English speakers and the others native English speakers. All experimental participants had normal or corrected-to-normal vision through glasses. Each experimental participant completed a total of 96 trials, where each trial comprised of reading a passage and answering MCQs for that passage. Experimental participants were instructed prior to each trial on how to read the passage, using one of the four reading conditions (sequential reading, skimming, scanning, and proofreading). The experiment was structured so that there was an equal number of trials (24) per reading condition (4). More information about the structure of the dataset can be found in Appendix 9.3.

The passages used were sampled from 12 frequently-occurring topics present in the RACE dataset [11] e.g. university/education, transportation, nature and animals, music, art, energy and climate change, sleep, stress and mental health, etc. Each experimental participant read an equal number of passages for each condition and topic. The MCQs used to measure comprehension were sampled from each passage's MCQ set (that excluded cloze-type questions). The passage sampling process is described in Appendix 9.2

Passages were presented on a 24-inch Phillips LCD monitor (model 240V5QDAB/00) with 1920 × 1080 resolution and controlled by a Dell Optiplex 5060 PC powered by the Windows 10 operating system. Experimental participants sat approximately 60cm from the screen and no chin rest was used. Eye movements were captured using the Gazepoint GP3 HD Eye Tracking device[1] with a sampling rate of $150Hz$ (one sample per 6.67 milliseconds). The data gathering process was driven by software written in Python using Psychopy [14].

The experimental participants were seated in a chair facing a computer monitor, with the eye tracker positioned below the monitor to capture eye movements. Calibration was performed using a 5-point grid, and the accuracy was checked with a 12-point grid (provided in the eye tracker's software). The data gathering process then started, comprising 96 trials divided into 4 sessions. Participants were allowed to rest for up to 15 minutes after each session was completed. Within one session, 24 passages were shown. After each passage the experimental participant answered 3 MCQs. Each trial had a short guideline at the beginning to indicate the required reading condition. This was followed by presentation of the passage on screen, where after a scene displaying a passage for reading, and finally three scenes for three multiple-choice comprehension questions.

4 RELEASED DATA AND RESOURCES

For each passage for each experimental participant, the following data was made available:

- Eye-movement measurements for the passage text with pre-extracted features from the eye tracking data;
- Identifying information for the passages and the associated MCQs used to measure comprehension;
- The calculated comprehension scores for all trials based on participant responses to the MCQs;
- And a variety of other useful and related features to support task participation (e.g., reading time, passage length and total words).

In total, there were 864 trials (96 trials/passages for $N = 9$ experimental participants), where the ground truth values for 216 trials (24 per participant) were kept aside as testing data, and the remaining 648 trials (72 per participant) were made available for training. The features in the RCIR dataset were made available in tabular format, and the passages and comprehension questions in a JSON format. A baseline approach in a Jupyter Notebook, and a guideline document were also made available.

In order to ease participation, the eye tracking signals for each trial were processed to extract representative eye movement features that have been commonly studied in the literature. These features included representative measures for fixations, saccades, blinks, and regressions, to name a few. For a more detailed description, please refer to Appendix 9.1). Participants' responses to MCQs were processed to obtain a single value (c_score) that captured their level of comprehension for each passage. This value (between 0 and 3), indicated the number of MCQs that the experimental participants could answer for each passage.

5 EVALUATION TASKS

5.1 Comprehension-evaluation sub-task (CET)

The CET sub-task aimed to explore how experimental participants comprehend on-screen text, and to what extent their comprehension level could be predicted by exploiting eye movement information captured by the eye tracking device. Given the provided

[1] https://www.gazept.com/

pre-extracted eye movement features of each passage, along with passage text, the task for participating researchers was to build a model to predict the comprehension score for each passage for each experimental participant in the testing data. Spearman's rank correlation coefficient (ρ) was then used to evaluate the prediction results for the participating teams.

5.2 Comprehension-based retrieval sub-task (CRT)

The CRT sub-task was an exploratory task that examined the potential of integrating comprehension evidence (from the CET sub-task) into an information retrieval system, in which retrieved passages were ordered based on their predicted comprehension score. 6 queries (topics) were given (each corresponding to a test topic as outlined in Appendix Table 5), and the participating researchers were expected to retrieve the passages that best match the query description and rank the passages in the descending order of comprehension level (from highest to lowest). More details of the 6 queries used in the evaluation process of this sub-task can be found in the Appendix 9.4. Normalised Discounted Cumulative Gain (nDCG) was used as the evaluation metric, which was formulated as follows:

$$nDCG(RL) = \frac{DCG(RL)}{DCG(GT\_RL)}$$

in which:

$$DCG(RL) = \sum_{i=1}^{N} \frac{2^{passage\_score(RL_i)} - 1}{log_2(i+1)}$$

where RL is the ranked list produced by the retrieval system while GT_RL is the true ranked list. The $passage\_score$ will justify the passage located at the i^{th} position in the ranked list by its *relevancy to the query* and *comprehension order*. In particular, the $passage\_score$ of a passage A is calculated as follow:

$$passage\_score(A) = (c\_score(A) + 1) * rel\_score(A)$$

where $c\_score(A)$ is the true comprehension score of the passage A (from 0 to 3 as in CET sub-task), and $rel\_score(A)$ is the passage's relevancy to the query and is either 0 (not relevant) or 1 (relevant).

6 EVALUATION RESULTS & ANALYSIS

Each team was allowed to submit up to 12 runs per sub-task in NTCIR-16 RCIR [2]. For the CET sub-task, there were a total of 16 official runs submitted by the three teams (HCMUS [12], KNUIR [8], and DCU [13]). Table 1 shows the Spearman's ρ scores (for the predicted score with respect to the true comprehension score for the test set) for different approaches investigated by the participating teams.

The DCU team [13] generated an additional 768 textual features extracted from the passages to predict the level of comprehension. Multiple ML configurations (i.e., subject-dependent, subject-independent and general/mixed training) were employed by the team to examine the variance in experimental participants' signals. A softmax-weighted aggregation of 3 models had the highest performance by the DCU team, with a Spearman's ρ score of 0.6. This used a Support Vector Classifier (SVC) trained on the top 30% of the most important non-textual features in a subject-dependent setting, and two Gradient Boosting Regressor (GBR) models trained on the top 50% (ranked most important) non-textual features in a subject-independent and pooled (general) setting. The second-ranked team (KNUIR) [8] scored a Spearman's ρ of 0.5706 using a random forest regressor that was trained on the pre-computed eye tracking features only and had its hyper-parameters tuned using grid searching. The HCMUS team [12] proposed 3 approaches to address the task that employed machine learning and neural network techniques. Additional features for the passages (i.e., the amount of numbers and uncommon English words) were created to generate additional information to train the models. The team came in third-place with a Spearman's ρ score of 0.5085 using the AutoML tool.

Table 1: Spearman's ρ for all official runs for the CET sub-task submitted by the three participating teams. Higher the Spearman's ρ scores indicate better comprehension score prediction.

Team	Run ID	Spearman's ρ
DCU [13]	0	0.4038
	1	0.5529
	2	0.5600
	3	0.5119
	4	0.5993
	5	0.5165
	6	0.5233
	7	**0.6000**
	8	0.3389
KNUIR [8]	0	0.5319
	1	**0.5706**
	2	0.0502
	3	0.3112
HCMUS [12]	0	0.4024
	1	0.4918
	2	**0.5085**

Table 2: Normalised Discounted Cumulative Gain (nDCG) of all submissions for the CRT sub-task (higher is better).

Team	Run ID	nDCG
DCU [13]	0	0.6929
	1	0.5856
	2	0.7178
	3	0.7245
	4	0.7215
	5	0.7215
	6	0.7153
	8	0.7149
	9	0.7271
	10	**0.7296**
	11	0.7164

[2] The submission of predictions for both tasks was achieved by using an online evaluation system that allowed participating teams to send their submissions using a HTTP GET / POST request to an online server. The evaluation result for each submission for a participating team was then automatically made available on a separate web page shortly after receiving the submission.

For the CRT sub-task, the DCU team submitted a total of 12 runs where 11 runs were reported as official submissions. The team focused mainly on evaluating different SBERT [17] structures (Base, Mini, Fast-Mini) and different keyword generation methods for the retrieval task. In particular, four levels of keyword extraction from the given queries (Appendix 9.4) were considered, from specific (more information, similar to the queries) to abstract (only nouns or most frequently-occurred words). The retrieved passages were then ranked using their best-performing model from the CET sub-task – SVC trained in a subject-dependent setting – without any further training or fine-tuning. The DCU team achieved a *nDCG* score of 0.7296 using the Fast-Mini type of SBERT with the input generation methods that only extract nouns from the queries. Only the DCU team participated in the CRT sub-task.

7 DISCUSSION

The evaluation results on the CET sub-task suggested that the highest Spearman's ρ scores were produced by Support Vector Machine, Random Forest Regressor, and Gradient Boosting approaches. Neural networks, in contrast, tended to perform worse due to the small size of the training dataset. Additionally, techniques to reduce features to avoid the curse of dimensionality problem, and techniques for tuning the models' hyper-parameters, namely grid search, were employed by the teams to make the models generalise better to the data, which consequently boosted the overall respective Spearman's ρ scores.

In terms of features used, both the HCMUS and DCU teams extracted additional features from the passages and aggregated them with eye movement features to predict comprehension. Despite the expectation that extended features would improve the model's prediction, the DCU team found that their textual features obtained from the ERNIE [19] model decreased the Spearman's ρ score. In the case of the textual features employed in the HCMUS team approaches, there were no explicit experiments conducted to evaluate the effectiveness of these features individually, and hence no conclusion could be drawn. The KNUIR team, as opposed to the other two teams, focused on eye movement features only. Notwithstanding that experimental participants' eye movement behaviours are different, the team discovered that horizontal regressive movements (feature F12 in Table 3), forward and backward movement distances (feature F11 in Table 3), and blink rate (feature F10 in Table 3) were important features for predicting comprehension.

For the CRT sub-task, the DCU team demonstrated a successful integration of their comprehension prediction model constructed for CET sub-task into their retrieval pipeline. Their highest nDCG score was 0.7296 (Table 2).

8 CONCLUSION

In this paper, we have outlined the motivation behind the RCIR task for NTCIR-16, that comprises two sub-tasks focused on reading comprehension prediction and comprehension-based retrieval. The creation of the novel multimodal reading dataset used in the RCIR was also described along with information about the experimental recording protocol, materials, pre-computed features from eye movement data, participating teams and evaluation results for same.

Three teams submitted runs along with a paper to the NTCIR-16 RCIR task. One team participated in both sub-tasks while the other teams took part in the CET sub-task only. The final ranking of the teams in CET sub-task is DCU, KNUIR and HCMUS with Spearman's ρ scores of 0.6000, 0.5706 and 0.5085, respectively. The best comprehension prediction model in CET sub-task was employed by the DCU team in the CRT sub-task without re-training, to rank the retrieved passages. The team achieved a normalised Discounted Cumulative Gain score of 0.7296 on the withheld test set.

Given the success of the pilot NTCIR-16 RCIR task, a future revised version of this task will focus on a larger longitudinal dataset with more experimental participants and trials. This will enable exploration of the consistency and stability of eye movement features that quantify one's reading comprehension over time. Moreover, our goal is to introduce more modalities into the dataset, namely electrooculogram signals (EOG), electrocardiography (ECG), galvanic skin response (GSR), electroencephalography (EEG) and facial expression, to further exploit the relationship between them and how they contribute to the prediction of measures of text comprehension. Future version of the RCIR task will incorporate a new sub-task which will exploit the use of eye movement data and comprehension prediction models into the indexing and retrieving process for information that a person has seen in daily living.

ACKNOWLEDGMENTS

This work was funded as part of Dublin City University's Research Committee, the Insight Centre for Data Analytics (which is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289_2, co-funded by the European Regional Development Fund) and with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183.

REFERENCES

[1] Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2011), 741–753. https://doi.org/10.1109/TPAMI.2010.86

[2] Richard R Day and Jeong-suk Park. 2005. Developing Reading Comprehension Questions. *Reading in a foreign language* 17, 1 (2005), 60–73.

[3] Cathal Gurrin, H. Joho, Frank Hopfgartner, Liting Zhou, Tu Ninh, Tu-Khiem Le, Rami Albatal, D.-T Dang-Nguyen, and Graham Healy. 2019. Overview of the NTCIR-14 Lifelog-3 task.

[4] Cathal Gurrin, Klaus Schoeffmann, Bjorn Thor Jonsson, Duc Tien Dang Nguyen, Jakub Lokoc, Luca Rossetto, Minh-Triet Tran, Wolfgang Hurst, and Graham Healy. 2021. An Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *ICMR '21, The 2021 International Conference on Multimedia Retrieval.* ACM, Taipei, Taiwan.

[5] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Found. Trends Inf. Retr.* 8, 1 (jun 2014), 1–125. https://doi.org/10.1561/1500000033

[6] Kenneth Holmqvist. 2011. *Eye Tracking : A Comprehensive Guide to Methods and Measures.* Oxford University Press, United Kingdom.

[7] Shoya Ishimaru, Kai Kunze, Koichi Kise, and Andreas Dengel. 2016. The Wordometer 2.0: Estimating the Number of Words You Read in Real Life Using Commercial EOG Glasses. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16).* Association for Computing Machinery, New York, NY, USA, 293–296. https://doi.org/10.1145/2968219.2971398

[8] Yumi Kim, Aluko Ademola, Jeong Hyeun Ko, and Heesop Kim. 2022. KNUIR at the NTCIR-16 RCIR: Predicting Comprehension Level using Regression Models based on Eye-Tracking Metadata. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16).* Tokyo, Japan.

[9] Janette K. Klingner. 2004. Assessing Reading Comprehension. *Assessment for Effective Intervention* 29, 4 (2004), 59–70. https://doi.org/10.1177/073724770402900408

[10] Kai Kunze, Hitoshi Kawaichi, Kazuyo Yoshimura, and Koichi Kise. 2013. The Wordometer - Estimating the Number of Words Read Using Document Image Retrieval and Mobile Eye Tracking. In *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*. IEEE Computer Society, 25–29. https://doi.org/10.1109/ICDAR.2013.14

[11] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 785–794. https://doi.org/10.18653/v1/D17-1082

[12] Kim-Nghia Liu, Vinh Dang, Thanh-Son Nguyen, and Minh-Triet Tran. 2022. HCMUS at the NTCIR-16 RCIR Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)*. Tokyo, Japan.

[13] Manh-Duy Nguyen, Nguyen Thao-Nhu, Thanh-Binh Nguyen, Caputo Annalina, and Cathal Gurrin. 2022. DCU Team at the NTCIR-16 RCIR Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)*. Tokyo, Japan.

[14] John Peirce, Jeremy R. Gray, Sol Simpson, Michael R. MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik K. Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51 (2019), 195 – 203.

[15] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124 3 (1998), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

[16] Keith Rayner, Alexander Pollatsek, Charles Clifton, and Jane Ashby. 2012. *The Psychology of Reading, 2nd Edition.*

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[18] Max Roser and Esteban Ortiz-Ospina. 2016. Literacy. *Our World in Data* (2016).

[19] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 8968–8975. https://doi.org/10.1609/aaai.v34i05.6428

[20] Kazuyo Yoshimura, Koichi Kise, and Kai Kunze. 2015. The eye as the window of the language ability: Estimation of English skills by analyzing eye movement while reading documents. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. 251–255. https://doi.org/10.1109/ICDAR.2015.7333762

9 APPENDIX

9.1 Eye-tracking Feature Extraction

In order to ease participation, the raw eye-tracking data for each volunteer was processed to extract meaningful features/measures i.e. fixations, saccade, blink, and pupil size. The full list of pre-computed features is available in Table 3.

To encode the time series features F1, F2, F5, F8, F11, F13, F14 in Table 3, two approaches were used:

- Bag-of-features (BOF): The trimmed_max, trimmed_min, standard deviation, mean, max-min, interquartile range, skewness, and kurtosis were calculated from the time-series features.
- Histogram (HIST): A histogram was obtained from the features to capture the distribution of data values (the number of bins and ranges were kept the same across all features and all experimental participants).

The encoded features are concatenated with the scalar data (F3, F7, F8, F9 in Table 3) to form the final set of features as provided in the dataset.

9.2 Materials for Reading Tasks

9.2.1 Overview. The passages and comprehension questions used in RCIR were extracted from the RACE dataset [11]. This dataset contained passages collected from online websites that span many different topical domains. The comprehension questions were constructed by the dataset experts to assess individuals' comprehension of each text. The questions are in the form of MCQs with two types: embedded (cloze type) and normal (non-cloze type). In addition, the RACE dataset is divided into 2 levels (middle and high school text content). In RCIR, we used the high-school level passages, as our targeted experimental participants were undergraduates, post-graduate students, and staff members within the department.

9.2.2 Topic-modelling. Since the passages in RACE dataset are not categorised, we employed a topic-modelling process to group the passages into topics. First, we filtered the 18,728 high-school-level passages down to 4,275 filtered texts to include the candidate passages that had at least three questions in the normal answering style (questions without cloze). Then a vocabulary was built for the filtered passages using the five most frequent words of each text. Next, all candidate passages were TF-IDF vectorised to fit a NMF clustering model to group these passages into multiple clusters. Twelve clusters were selected from these and formed the topics based on the passages within the cluster.

9.2.3 Topic Validation. Prior to experimental participant data collection, we conducted a topic validation process for the passage texts, where we had two annotators confirm for each text passage that it belonged to the topic. A summary of topics is described in Table 5.

9.3 Dataset Structure

The data was provided as an archive (in ZIP format), that included 9 directories (from **0000** to **0008**), each of which contained the reading data for one experimental participant and the other associated metadata (for model training). The training and test data were available in the ***train.csv*** and ***test.csv*** files. The format of the training data is illustrated in Table 4.

The testing data had a similar structure, but instead the columns *c_score* and *topic_id* were removed, as these are the prediction targets (dependent variable(s)) of the CET and CRT tasks. In addition to training and testing data, in each experimental participant's directory, there was a ***text.json*** file that contained the passage's content and MCQs.

9.4 Queries for the CRT sub-task

The queries used in the withheld test set for the CRT sub-task are as follows:

- Query 1: Find texts that describe/discuss the teaching strategy and the learning process in school. Texts that capture the students' opinion about their school and stories of students' school life are also relevant.
- Query 2: Find texts that talk about animals in general. Texts that discuss their life, habit, abilities, benefit, and endangerment are also considered relevant.
- Query 3: Find the texts that describe/discuss public transport. Texts can also be about analysing pros and cons, announcements and notices when traveling, and incidents around public transports, to name a few.
- Query 4: Find the texts that are related to the arts. Texts might also discuss the history of art, different genres of art, art exhibitions, galleries, and similar.
- Query 5: Find the texts that describe/discuss climate change and global warming. Texts that describe/discuss environmental effects and wildlife are also relevant.
- Query 6: Find the texts that describe/discuss mental and emotional health. Texts can also include research studies on stress and how to cope with it.

Table 3: Summary of the pre-computed eye tracker features. The features were divided into six groups (GID), and each feature had a unique identifier (FID). Labels in parentheses under each feature's name are the abbreviations used in the dataset.

GID	FID	Features	Description
G1	F1	Raw Movement Data (RAW_X, RAW_Y)	The original horizontal and vertical eye movements captured by eye-tracker (X–horizontal, Y–vertical).
G2	F2	Fixation Durations (FIXA_DUR_NORM)	Time in seconds for each fixation point, divided by the total reading time for the entire text.
	F3	Number of Fixations (NUM_FIXA)	Total number of fixations that a person made.
	F4	Fixations Rate (RATE_FIXA)	Number of fixations divided by the total words in the text.
G3	F5	Saccade Durations (SACC_DUR_NORM)	Time in seconds for each eye movement from one fixation point to another fixation point, divided by the total reading time for the entire text.
	F6	Number of Saccades (NUM_SACC)	Total number of saccades that a person made.
	F7	Saccades Rate (RATE_SACC)	Number of saccades divided by the total words in the text.
G4	F8	Blink Durations (BLINK_DUR_NORM)	Time in seconds for each blink, divided by the total reading time for the entire text.
	F9	Number of Blinks (NUM_BLINK)	Total number of blinks that a person made.
	F10	Blinks Rate (RATE_BLINK)	Number of blinks divided by the total words in the text.
G5	F11	Forward and Backward Movement Distances (FIXA_X_FWD, FIXA_X_BWD, FIXA_Y_FWD, FIXA_Y_BWD)	The L1 fixation distances made by the eyes when reading forward and regressing, vertically (X) and horizontally (Y).
	F12	Regression Rate (RATE_X_BWD, RATE_Y_BWD)	Frequency of the eyes moving back and fixating on certain points.
	F13	Speed (SPEED_X, SPEED_Y)	Horizontal (X) and vertical (Y) movement speed between two consecutive fixations.
G6	F14	Pupil Diameters (LP_SIZE, RP_SIZE)	The diameter of the left (L) and right (R) eye pupil in millimeters.

Table 4: Format of the training data

Column name	Description
c_score	The comprehension score based on the experimental participant's responses to the MCQs.
topic_id	The topic the passage belongs to.
text_id	The identifier of the passage associated with the trial. This can be used to obtain the passage's content and question set from the provided ***text.json*** file)
time_reading	The time it takes for an experimental participant to read a passage in a trial. (The time limit for each trial was set to be 60 seconds).
total_words	The number of words in the passages.
remaining columns	The pre-extracted features from experimental participant's eye tracking data.

Table 5: Description of the topics used in the RCIR dataset

Topic	Description	Top 5 Keywords	Test set
1	The passages mainly focus on different topics related to university, students and education.	students, college, education, student, university	No
2	The passages are about students’ school life, teaching and learning.	school, high, teacher, teachers, schools	Yes
7	The passages are related to animals e.g. their life, their abilities.	animals, animal, elephants, wild, zoo	Yes
9	The passages mainly focus on public transportation, especially trains.	train, london, station, travel, bus	Yes
11	The passages are about music, in which most of the passages describe singers, composers, and bands.	music, songs, song, festival, listening	No
16	The passages are related to energy in general e.g., green energy, clean energy, source of energy.	energy, pollution, air, oil, wind	No
19	The passages mainly discuss sleep with most of the passage about the study conducting sleep research.	sleep, night, sleeping, hours, bed	No
24	The passages are about cars and driving cars.	car, cars, road, driving, traffic	No
29	The passages are mainly related to the arts, spanning different genres of art, history, galleries and exhibitions.	art, paintings, artists, painting, artist	Yes
37	The passages focus mainly on discussing climate change, global warming, and how wildlife are affected.	ice, sea, scientists, antarctica, climate	Yes
40	The passages are mainly about stress, mental health, and emotional health.	stress, health, mental, anxiety, life	Yes
41	The passages are related to pets, their stories and their abilities.	dog, dogs, cat, pet, pets	No