A World CLASSE Student Summary Corpus

Scott A. Crossley[1], Perpetual Baffour[2], Mihai Dascalu[3], Stefan Ruseti[3]
Vanderbilt University[1], The Learning Agency[2], University Politehnica of Bucharest[3]
scott.crossley@vanderbilt.edu, perpetual@the-learning-agency.com, mihai.dascalu@upb.ro, stefan.ruseti@upb.ro

Abstract

This paper introduces the Common Lit Augmented Student Summary Evaluation (CLASSE) corpus. The corpus comprises 11,213 summaries written over six prompts by students in grades 3-12 while using the CommonLit website. Each summary was scored by expert human raters on analytic features related to main points, details, organization, voice, paraphrasing, and language beyond the source text. The human scores were aggregated into two component scores related to content and wording. The final corpus was the focus of a Kaggle competition hosted in late 2022 and completed in 2023 in which over 2,000 teams participated. The paper includes a baseline scoring model for the corpus based on a Large Language Model (Longformer model). The paper also provides an overview of the winning models from the Kaggle competition.

1 Introduction

Many educational applications are interested in assessing student-generated knowledge to assess learning and development. In terms of assessing student comprehension of text, generation effects (Slamecka & Graff, 1978) that result from students writing about what they have read have been shown to substantially improve learning (Bertsch et al., 2007; McCurdy et al., 2020). A number of educational applications have taken advantage of generation effects to enhance students' reading comprehension skills, including Summary Street (Wade-Stein & Kintsch, 2004), the Interactive Strategy Training for Active Reading and Thinking (iSTART) tool (McNamara et al. 2004), the CommonLit online reading program (commonlit.org), and the intelligent Textbooks for Enhanced Lifelong Learning (iTELL) framework (Morris et al., in press).

The most common approach to assessing students' reading comprehension through text generation is likely through text summarization. Text summarization is a valuable tool to build and assess student knowledge (Graham & Harris, 2015; Head et al., 1989) because the process of summarization helps students build and consolidate their knowledge about reading materials (Silva & Limongi, 2019). Text summarization has also been shown to lead to stronger learning gains than other forms of comprehension assessment, including constructed responses (Carroll, 2008), long-form essays (Gil et al., 2010), and traditional assessments like multiple-choice and fill-in-the-blank questions (Mok & Chan, 2016).

While effective, many teachers hesitate to integrate summary assessments of reading in the classroom because manually grading summaries is resource-intensive (Lagakis & Demetriadis, 2021; Li et al., 2018). However, student text summarization can also be assessed automatically through the use of Natural Language Processing (NLP) techniques such as semantic similarity metrics (Crossley et al., 2019; Li et al., 2018; Wade-Stein & Kintsch, 2004) or contextualized word embeddings like those found in Transformer-based language models (Botarleanu et al., 2022; Morris et al., 2023).

To assess student summarization strength automatically, NLP models depend on the availability of large corpora of summaries that have been scored for quality. Unfortunately, previous research has depended on closed-source collections of summaries that are not available to the broader research community (Botarleanu et al., 2022; Crossley et al., 2019; Li et al., 2018; Wade-Stein & Kintsch, 2004), which limits the strength,

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 99–107

replication, and generalizability of summarization models. Additionally, many of the corpora used in previous research have included summaries written by crowdsourced workers and not students (Botarleanu et al., 2022; Crossley et al., 2019; Li et al., 2018)

The goal of this study is to introduce the Common Lit Augmented Student Summary Evaluation (CLASSE) corpus. The corpus comprises 11,213 summaries written over six prompts by students in grades 3-12. All summaries were written on the CommonLit website. Each summary was scored by expert human raters on analytic features related to summarization content and wording. The study also introduces a baseline NLP summary scoring model for the corpus as well as the winning models developed in a large-scale data science competition hosted for the corpus.

1.1 Summary writing

Summarizing a reading involves two cognitive processes: comprehension and content production (Li et al., 2018). The reading process leads to the reader's comprehension of the source material. This process generally consists of readers identifying the text's main themes, the ideas that support these themes, and the structures and organization of the text (Spirgel & Delaney, 2016). After reading, summarization allows the student to reproduce the content of the source text that they read and involves the reader (now the writer) generalizing the main ideas contained in the text, synthesizing those ideas, organizing those ideas coherently within the summary, and selecting the proper words and sentence structures to represent the ideas (Brown & Day, 1983; van Dijk & Kintsch, 1983; Galbraith & Baaijen, 2018; León et al. 2006; Nelson & King, 2022). The cognitive demands entailed in summarizing help consolidate the knowledge gained from reading into long-term memory (Silva & Limongi, 2019).

Research indicates that reading to writing tasks like summarization can increase learning outcomes in various content domains (Graham et al., 2020; Silva & Limongi, 2019) and for different types of learners (Rogevich & Perin, 2008; Trabasso & Bouchard, 2002; Shokrpour et al., 2013). A meta-analysis of 56 experiments on the effect of reading on writing tasks found an average weighted effect size of Hedges's $g = 0.3$ ($p < .005$) between pre- and post-tests for students (Silva & Limongi, 2019). Additionally, compared to other methods to assess reading comprehension and knowledge development, like constructed responses, essays, and multiple-choice questions, research has found that summarizations are more effective (Carroll, 2008; Gil et al., 2010; Mok & Chan, 2016).

1.2 Automatic summary evaluation

Despite the effectiveness of having students summarize what they have read, providing feedback to students about the quality of summaries is time-consuming for educators (Gamage et al., 2021; Lagakis & Demetriadis, 2021; Li et al., 2018), thus making human-driven summary assessments difficult to scale.

Noting the importance of summarization in educational settings and the challenges of integrating it into the classroom, researchers have investigated the potential for automatic summary evaluation (ASE) to provide students with computational-derived feedback.

Initial methods for ASE predominantly involved assessing a student's summarization work by comparing it with model summaries crafted by experts. These methods have the advantage of relying on a single expert-derived summary to establish a benchmark for quality. Metrics like ROUGE (Lin & Hovy, 2003) were utilized to assign scores to summaries by examining the frequency of shared words and phrases between the student and expert summaries. Although ROUGE metrics align with the quality ratings given by experts and have been widely adopted in developing summarization tools (Ganesan, 2018; Scialom et al., 2019), the metrics tend to favor basic lexical attributes. This shortcoming can be overcome by employing more sophisticated NLP techniques, such as those involving word embeddings (Ng & Abrecht, 2015).

The earliest attempt at using a word embedding approach to score summaries was likely with the educational application Summary Street. Summary Street allowed students to produce multiple summary drafts and provided feedback to students based on Latent Semantic Analysis (LSA), an early word embedding model. Summary Street used LSA to uncover typical sentences in each section of a text. These sentences were then combined to form a typical summary. Semantic similarity between a student's summary and the typical summary was used to provide feedback to the student about the quality of their summary (Wade-Stein & Kintsch, 2004).

Li et al. (2018) also used LSA to provide scores for summaries written by crowdsourced workers on Mechanical Turk. The crowdsourced summaries were scored by graduate students on four criteria: thesis statement, content, mechanics and grammar, and signal words. Li et al. found that crowdsourced summaries were scored as well as summaries produced by experts using LSA. Li et al. argued that crowdsourced workers could produce a model summary similar to the model summaries produced by experts, which could make it easier to develop model summaries for automated scoring.

Other summarization scoring models have combined more advanced word embedding models and other NLP features to predict quality. For instance, Crossley et al. (2019) developed a summarization model to predict ratings of main idea integration in summaries collected on Mechanical Turk using lexical diversity features, a word frequency metric, and Word2vec semantic similarity scores between summaries and the corresponding source material. The model explained 53% of the variance in ratings.

With the rise of Transformer-based language models, new methods of automated summary evaluation have been evaluated. For instance, Botarleanu et al. (2022) used the summaries of Crossley et al. (2019) to train a Longformer model (Beltagy et al., 2020) to predict overall summarization scores derived from an analytic rubric; their model explained ~55% of the score variance. Morris et al. (in press) used an extended dataset of the one used by Crossley et al. (2019). In addition to crowdsourced summaries, the extended dataset also included summaries written by high school and university students. Morris et al. used the dataset to predict two aspects of summarization quality: content and wording. Using a Longformer, they explained .82 of the variance in the content scores and .70 of the variance in the wording scores.

2 The CLASSE Corpus

While research ASE has gained traction and shown improvements over the last 20 years, the work is somewhat fragmented. A major reason for this is that researchers do not have a large-scale open-source summarization corpus to develop, test, and validate ASE models. Other reasons include the use of different NLP approaches to model summarization quality, the sampling of different populations of writers, and the use of different scoring metrics.

The Common Lit Augmented Student Summary Evaluation (CLASSE) corpus is meant to help address this fragmentation by providing researchers with a gold-standard corpus of open-source summaries written by students. The corpus is freely available in the following repository: https://github.com/scrosseye/CLASSE.

2.1 Summaries

The corpus of summaries found in CLASSE was provided by CommonLit, an online content library and writing platform. The initial corpus comprised 11,353 summaries. Within the CommonLit interface, students read texts and write summaries on those texts. Students also have the opportunity to write essay responses, complete vocabulary quizzes, and answer multiple-choice questions about the text. The final CLASSE corpus after pruning (see section 2.2) comprises 11,213 summaries written over six prompts by students in grades 3-12.

Grade	N	Length (M)	Length (SD)
3	2	172.00	49.50
4	12	77.92	49.19
5	248	87.51	70.17
6	1072	82.58	57.61
7	1177	78.92	58.66
8	1844	76.30	46.06
9	2531	71.62	43.82
10	2247	75.92	50.73
11	1942	73.61	51.15
12	138	80.86	57.22

Table 1: Grade Level

Prompt	N	Length (M)	Length (SD)
Third-Wave	1103	73.88	47.31
Tragedies	2057	63.87	44.93
Jungle	1996	80.52	56.16
Greek	2021	73.72	38.31
Egyptian	2009	85.71	62.58
Nature Nurture	2027	77.10	48.67

Table 2: Prompt Information

The majority of the summaries were written by students in the 6th to 11th grade, with smaller numbers of 3rd, 4th, 5th, and 12th grade students (see

Table 1 for details). English language learning (ELL) status is also available for the students (n = 661). The six prompts were related to the topics of the third wave, poetic tragedies, the novel *The Jungle*, Greek society, Egyptian Society, and the nature/nurture debate (see Table 2 for details). The mean length of the summaries was 75.90 (SD = 50.94, min = 22, max = 651). Text length by grade and prompt is reported in Tables 1 and 2. No demographic information beyond grade and ELL status is available for the students.

2.2 Summary scoring

Summaries were scored by expert raters using a standardized scoring rubric and procedure. An outside agency specialized in providing performance assessment scoring services was hired to score the summaries and initial selection of summaries. Two expert raters scored each summary using a 0-4 scaled analytic rubric to score six criteria important in understanding the quality of summarizations. The rubric was developed based on research into language elements related to essay quality reported by Taylor (2013) and Westby et al. (2010). The initial rubric was revised based on feedback from a panel of teachers and a panel of researchers who specialize in the teaching of summaries. The finalized rubric included analytic ratings for main point/gist (did the summary contain the ideas of the source text), details (did the summary contain all the main ideas of the source text), organization (were the ideas logically presented and linked to each other to support comprehension), voice (was language impartial and objective in the summary), word/paraphrasing (did the summary appropriately paraphrase the source text), and language beyond the source text (did the summary show a range of lexical and syntactic features). The scoring rubric is available at this link. Raters also flagged any summaries that included offensive or emotionally charged language or personally identifiable information (PII). While no PII was reported, 127 summaries were removed for language use.

Raters were provided with ground truth example summaries that had been previously scored. As well, raters went through extensive norming prior to independent rating. After norming, each summary was read by at least two raters and, in some cases, three raters (if there was substantial disagreement). Ratings were conducted by prompt, and rater final scores were averaged such that scores of 3 and 2 were averaged to 2.5.

Score distributions were generally normal except for the details, organization, and wording items, which were positively skewed, indicating a greater number of 1s than 1.5s. Strong correlations were reported among the analytic items, with the highest correlation between organization and voice and the lowest correlation between detail and word (see Figure 1 for a correlation heat map). The exact agreement among analytic items hovered around 70% (see Table 3 for details). Quadratic weighted kappa (QWK) scores for inter-rater reliability were substantial (QWK < .60) for all items except wording, which reported a moderate QWK = .532 (see Table 4).

Significant differences were noted between ELL students and non-ELL students for both content scores (t = 3.993, p < .001) and wording scores (t = 5.684, p < .001). Descriptive statistics for content and wording scores by ELL and non-ELL students are reported in Table 5. No significant correlations were reported between grade level and content score (r = -0.036, p > .050) and wording scores (r = -0.049, p > .050). Descriptive statistics for content and wording scores by grade are reported in Table 6.

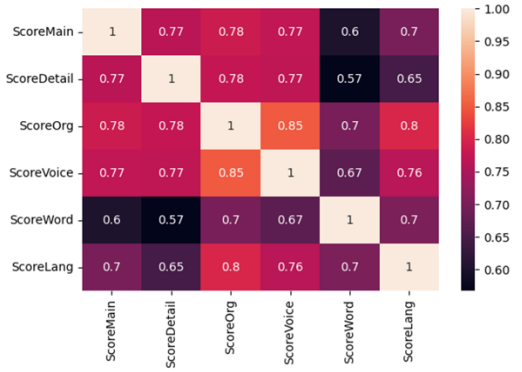

Figure 1: Heatmap for correlations among analytic item scores

Item	Adjacent Low	Exact	Adjacent High
Main Idea	13.2	73.0	13.2
Details	13.9	72.0	13.9
Organization	15.1	69.0	15.1
Voice	15.4	69.0	15.4
Wording	16.9	65.0	16.9
Language	11.8	76.0	11.8

Table 3: Exact and adjacent percentages

Item	QWK
Main Idea	0.617
Details	0.673
Organization	0.694
Voice	0.683
Wording	0.532
Language	0.653

Table 4: Quadratic Weighted Kappa (QWK) for inter-rater reliability

Group	Content M (SD)	Wording M (SD)
Non-ELL	0.016 (1.002)	0.023 (0.999)
ELL	-0.136 (0.950)	-0.186 (0.910)

Table 5: Descriptive statistics for content and wording scores for ELL and non-ELL students

Grade	Content M (SD)	Wording M (SD)
3	1.593 (2.015)	1.041 (1.419)
4	-0.201 (1.131)	-0.359 (0.759)
5	-0.056 (1.115)	-0.14 (0.964)
6	0.036 (1.067)	-0.039 (0.939)
7	-0.063 (1.054)	-0.071 (0.953)
8	0.008 (0.985)	0.076 (0.963)
9	0.025 (0.923)	0.098 (0.955)
10	0.081 (1.01)	0.084 (1.057)
11	-0.061 (1.002)	-0.142 (1.04)
12	-0.073 (1.008)	-0.146 (0.967)

Table 6: Descriptive statistics for content and wording scores by grade

2.3 Dimensionality reduction

Since the rubric consisted of six criteria, many of which were related, we conducted a Principal Component Analysis (PCA) to assess the potential to reduce the dimensionality of the six analytic scores into a smaller number of related constructs.

Before conducting the PCA, the human scores were standardized using z-score normalization. An initial PCA was performed with all possible factors (n = 6). A Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy indicated that no variables need to be removed (i.e., all KMO values were above .5), and the overall KMO score = .918 indicated a "meritorious" sample (Kaiser & Rice, 1974). The PCA reported a Bartlett's test of sphericity, $\chi^2 = 61{,}533.87$, $p < .001$, indicating that correlations between the analytic scores were sufficiently large for the PCA. Within the components, there was a break in the cumulative variance explained between the second and the third components. Considering this break, we decided on a 2-component solution when developing the PCA. These 2 components explained approximately 86% of the shared variance in the data from the initial PCA.

The first component was related to *content* (i.e., Component 1), and the analytic items details, main point, voice, and organization were combined into a weighted score. The analytic items wording/paraphrasing and language beyond the source were combined into a weighted score designated as *wording* (i.e., Component 2). The component scores were z-score normalized and rescaled such that zero represents the mean for each principal component, and one unit represents one standard deviation.

2.4 Final dataset

The final dataset comprises 11,213 summaries and metadata in tabular format and is available at this link. The dataset contains student ID numbers (anonymous), the prompt ID for each summary, the text of the summary, the average content and wording scores for the summary, the student grade level, and ELL classification, along with the data split that was used in the Kaggle competition (see section 4 for details). The data was split into a training set (n = 7,165), a validation set used as a test set for the public leaderboard on Kaggle (n = 2,021), and a test set used for the private leaderboard on Kaggle (n = 2,027). The splits were selected so that the difference in scores across the splits was similar to demographic information (grade and ELL classification). The training set comprised four prompts (Third Wave, Tragedies, The Jungle, and Egyptian Society). The validation set included a single prompt (Greek Society), as did the test set (Nature versus Nurture).

3 Baseline prediction model for CLASSE corpus

We developed a simple baseline model for the CLASSE by finetuning a Longformer model (Beltagy et al., 2020) to predict the content and wording scores, given the original text and the summary. The baseline model is not meant to extend the technical boundaries of summary classification models but rather provide a simple metric from which to measure scoring gains.

3.1 Model description

An encoder architecture was chosen for the baseline model over a decoder model because the prediction task is a regression that involves continuous values. Since a decoder model is used to generate text, the output values would have to be expressed in words. This does not imply that a decoder cannot be used for this task, but an encoder model seemed a better fit for the data.

The input for the model consisted of both the summary and the source text, separated by the "sep" token. Given the length of the input exceeding 512 tokens, a Longformer model was chosen as a baseline encoder.

Several options were tested for the final summary embedding: pooled output, average of all tokens, and average of summary tokens. Adding a hidden layer between the embedding and the decision layer was also considered. The best configuration used the average of the summary tokens followed by a dropout layer of 20%, no hidden layer or output activation, and a learning rate of 1e-5 using the Adam optimizer. The mean squared error sum for the two tasks was used as a loss function. The lowest validation loss was obtained after three epochs, and the corresponding model was used for evaluation. The model was trained on the training set, validated on the validation set, and tested on the test set used in the Kaggle competition.

3.2 Prediction performance

The metric used for the Kaggle competition was Mean Columnwise Root Mean Squared Error (MCRMSE), which is the average of the RMSE for the two scoring components (content and wording). RMSE is a general error metric used for numerical predictions that punishes large errors in predictions. An RMSE score of zero represents a perfect fit between the model and the outcome variables (in this case, content and wording scores). Thus, a lower RMSE represents a better model.

The results for the baseline model for each partition, each component, and the average scores are presented in Table 6. The model performed well on the training and validation sets for content, but it performed less accurately on the wording scores. Model performance dipped for the content scores in the test set and fell for the wording scores. The overall scores for MCRMSE were strong for the training set but fell in the validation and test sets. The final MCRMSE reported for the test set was 0.582.

Partition	Content RMSE	Wording RMSE	MCRMSE
Train	0.375	0.427	0.401
Validation	0.415	0.614	0.515
Test	0.480	0.683	0.582

Table 6: Baseline model performance

4 Kaggle Competition

The CLASSE dataset was the subject of a recently completed Kaggle competition (CommonLit - Evaluate Student Summaries). The goal of the competition was for data scientists to assess the quality of summaries in the CLASSE corpus in terms of content and wording. The winning models provide state-of-the-art techniques for modeling summary scoring in student data and demonstrate the potential for the CLASSE corpus to inform student learning and interventions.

The competition started in July of 2023 and ended in October of 2023. Over 2,000 teams comprising ~2,500 competitors entered the competition, creating over 40,000 summary scoring models. All winning models are freely available for use through an MIT license and provided on the Kaggle website. The Kaggle website also provides the training and validation data used in the competition.

5 Kaggle competition results

As mentioned earlier, success in the Kaggle competition was demonstrated through a model's mean column-wise root mean squared error (MCRMSE), which represented the average Root Mean Squared Error (RMSE) across the content and wording scores.

The top 17 teams reported an MCRMSE below .46, with the first-place team reporting an MCRMSE of .452. These models thus outperformed our baseline model (MCRMSE = 0.582). Within the top five entrants, the most common approach used when modeling the summary scores was an ensemble model using the DeBERTa encoder. This approach was used with the second through fifth place teams, with all teams except the fifth place team using only DeBERTa models (the fifth place team used DeBERTa v3 large and a LightGBM ensemble model). The first-place team used a single

DeBERTa model (v3 large), but critically, they augmented the training set by creating 1000 new prompts with associated sources using generative AI. For each prompt, they also created 21 summaries and pseudo-labeled those summaries. Other common approaches used to improve the models included using a head mask for only the student summaries instead of a normal attention mask, using generative AI models to generate varieties of the existing prompts, hyperparameter searches, extending the inference max length, and using all of the input (summary, prompt, source, and title) in the training models.

6 Discussion and conclusion

This paper has introduced the CLASSE corpus, the scoring metrics for the corpus, and a baseline model for summary scoring based on a DeBERTa Transformer-based encoder. The paper also introduced the winning summarization models from the Kaggle competition held in support of the CLASSE corpus.

The CLASSE comprises 11,213 summaries written over six prompts by students in grades 3-12 while using the CommonLit website. Each summary was scored by expert human raters on analytic features related to summarization content and wording.

Reliability metrics for the human scoring indicated substantial reliability in all items except paraphrasing/wording, which reported moderate reliability. Paraphrasing is the restatement of a passage such that the propositional meaning is similar, but the words and structures differ. Recognizing when words differ between passages is relatively easy, but recognizing the alteration of clauses is a difficult task (Barzilay & Lee, 2003), which may explain the moderate reliability reported by human raters.

The analytic scores were aggregated into components using a principal component analysis (PCA) to better represent the underlying structure of the human ratings. The PCA reported two components related to content and wording. Content included features related to main ideas, details for those ideas, the organization of those ideas, and the objectivity of how those ideas were presented. The content component provides an overall assessment of how the ideas in the source text are distilled into a coherent and objective framework in the student summaries. Wording includes features related to paraphrasing and the use of language beyond the source. This component was concerned with the manner in which the summary presented the ideas from the source text, specifically, did the summary use original wording (paraphrasing) and whether this wording was lexically and syntactically complex.

The baseline model introduced in this paper used a Longformer model that used both the summary and the source text as input for model predictions. The Longformer performed well on the training data but reported drops in the validation and test data. This is the result of the Longformer model learning the patterns of successful summarization specific to the four prompts in the training set but not learning how to extend scoring beyond those prompts to the two unique prompts in the validation and test sets.

The results of the subsequent Kaggle competition showed a number of innovations that helped competitors produce winning models, many of which addressed the limitations of the baseline model. The winning model used a single Transformer encoder (DeBERTa v3 large), but, importantly, they augmented their training data to include a much larger number of prompts and summaries written on those prompts. Extending the number of prompts and summaries allowed the model to generalize better to the unique prompts found in the validation and test set. Other innovations in summary scoring that resulted from the Kaggle competition included pseudo-labeling of AI generated summaries for content and wording scores, the use of head masks, and extending the inference max length.

6.1 Limitations

While the CLASSE corpus is the largest corpus of student summaries, with individual human scores assigned to each summary, there are limitations to the corpus. An important limitation is that there are only six source texts and prompts for the corpus. As noted, the first-place solution on Kaggle augmented the CLASSE dataset by creating 1,000 new prompts and source text along with pseudo-labeling these summaries, all of which are available in the winning model. However, augmenting data is different from collecting real data, and future developments of CLASSE or newer summarization datasets should include a greater number of prompts.

Another limitation of the CLASSE corpus is that certain grades (i.e., 6^{th}-11^{th} grades) were over-

represented in the corpus. Greater representation of lower and upper grades, including college-level students, is warranted. Finally, while the CLASSE corpus includes some individual difference metrics, little information is known about the writers in terms of gender, race/ethnicity, or socio-economic status, all of which are important student-oriented variables that may influence human ratings.

6.2 Future directions

The goals of the Kaggle competition were to publicize and make freely available a large-scale corpus of student-written summaries and advanced models of assessing summarization quality. Future directions include integrating the models developed in the Kaggle competition into educational applications to help students receive feedback on summaries written within these applications. Knowing the strength of generation effects on learning (Bertsch et al., 2007; McCurdy et al., 2020) and the strengths of summarization tasks in general (Carroll, 2008; Gil et al., 2010; Mok & Chan, 2016), the integration of CLASSE corpus scoring models into educational applications will ensure students quickly receive formative feedback about their summaries, allowing for deliberative practice during the revision process and increased learning.

Acknowledgments

The authors would like to thank Michelle Brown at commonlit.org along with Ulrich Boser, Meg Benner, and Alex Franklin at the Learning Agency for their help in releasing the corpus and hosting the Kaggle competition.

References

Barzilay, R., & Lee, L. 2003. Learning to paraphrase: an unsupervised Approach Using Multiple-Sequence Alignment. *Proceedings of HLT-NAACL*. 16-23. Edmonton, Canada.

Beltagy, I., Peters, M. E., & Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. 2007. The generation effect: A meta-analytic review. *Memory & cognition, 35*, 201-210.

Brown, A. L., & Day, J. D. 1983. Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*, 1–14.

Botarleanu, R.-M., Dascalu, M., Allen, L. K., Crossley, S. A., & McNamara, D. S. 2022. Multitask Summary Scoring with Longformers. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (Vol. 13355, pp. 756–761). Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_79

Crossley, S. A., Kim, M., Allen, L., & McNamara, D. 2019. Automated summarization evaluation (ASE) using natural language processing tools. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20* (pp. 84-95). Springer International Publishing.

Carroll, D.W. 2008. Brief report: A simple stimulus for student writing and learning in the introductory psychology course. *North American Journal of Psychology, 10*, 159–164.

Galbraith, D., & Baaijen, V. M. 2018. The Work of Writing: Raiding the Inarticulate. *Educational Psychologist*, *53*(4), 238–257. https://doi.org/10.1080/00461520.2018.1505515

Gamage, D., Staubitz, T., & Whiting, M. 2021. Peer assessment in MOOCs: Systematic literature review. *Distance Education*, *42*(2), 268–289. https://doi.org/10.1080/01587919.2021.1911626

Ganesan, K. 2018. *ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks*. https://doi.org/10.48550/ARXIV.1803.01937

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. 2010. Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology, 35,* 157–173.

Graham, S., & Harris, K. R. 2015. Common Core State Standards and Writing: Introduction to the Special Issue. *The Elementary School Journal*, *115*(4), 457–463. https://doi.org/10.1086/681963

Graham, S., Kiuhara, S. A., & MacKay, M. 2020. The Effects of Writing on Learning in Science, Social Studies, and Mathematics: A Meta-Analysis. *Review of Educational Research*, *90*(2), 179–226. https://doi.org/10.3102/0034654320914744

Head, M. H., Readence, J. E., & Buss, R. R. 1989. An examination of summary writing as a measure of reading comprehension. *Reading Research and Instruction*, *28*(4), 1–11. https://doi.org/10.1080/19388078909557982

Kaiser, H. F., & Rice, J. 1974. Little jiffy, mark IV. *Educational and psychological measurement*, *34*(1), 111-117.

Lagakis, P., & Demetriadis, S. 2021. Automated essay scoring: A review of the field. *2021 International*

Conference on Computer, Information and Telecommunication Systems (CITS), 1–6. https://doi.org/10.1109/CITS52676.2021.9618476

León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. 2006. Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods, 38*, 616–627. https://doi.org/10.3758/BF03193894

Li, H., Cai, Z., & Graesser, A. C. 2018. Computerized summary scoring: Crowdsourcing-based latent semantic analysis. *Behavior Research Methods*, *50*(5), 2144–2161. https://doi.org/10.3758/s13428-017-0982-7

Lin, C.-Y., & Hovy, E. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology NAACL'03*, *1*, 71–78. https://doi.org/10.3115/1073445.1073465

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. 2020. Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, *27*(6), 1139–1165. https://doi.org/10.3758/s13423-020-01762-3

McNamara, D. S., Levinstein, I. B., & Boonthum, C. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 222-233.

Mok, W. S. Y., & Chan, W. W. L. 2016. How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science, 44*, 567–581.

Morris, W., Crossley, S., Holmes, L., Ou, C., McNamara, D., Dascalu, M. 2023 Using Transformer Language Models to Provide Formative Feedback in Intelligent Textbooks. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education (AIED)* (pp. 484–489). Springer Nature Switzerland.

Morris, W., Crossley, S., Holmes, L., Ou, Chaohua, Dascalu, M., & McNamara, D. in press. Formative Feedback on Student-Authored Summaries in Intelligent Textbooks using Large Language Models. *Journal of Artificial Intelligence in Education*.

Nelson, N., & King, J. R. 2022. Discourse synthesis: Textual transformations in writing from sources. *Reading and Writing*. https://doi.org/10.1007/s11145-021-10243-5

Ng, J.-P., & Abrecht, V. 2015. *Better Summarization Evaluation with Word Embeddings for ROUGE* (arXiv:1508.06034). arXiv. http://arxiv.org/abs/1508.06034

Rogevich, M., & Perin, D. 2008. Effects on science summarization of a reading comprehension intervention for adolescents with behavior and attention disorders. *Exceptional Children, 74*, 135–154.

Scialom, T., Lamprier, S., Piwowarski, B., & Staiano, J. 2019. *Answers Unite! Unsupervised Metrics for Reinforced Summarization Models*. https://doi.org/10.48550/ARXIV.1909.01610

Shokrpour, N., Sadeghi, A., & Seddigh, F. 2013. The effect of summary writing as a critical reading strategy on reading comprehension of Iranian EFL learners. *Journal of Studies in Education, 3,* 127–138. https://doi.org/10.5296/jse.v3i2.2644

Silva, A., & Limongi, R. 2019. Writing to Learn Increases Long-term Memory Consolidation: A Mental-chronometry and Computational-modeling Study of "Epistemic Writing." *Journal of Writing Research*, *11*(vol. 11 issue 1), 211–243. https://doi.org/10.17239/jowr-2019.11.01.07

Slamecka, N. J., & Graf, P. 1978. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4* (6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Spirgel, A. S., & Delaney, P. F. 2016. Does writing summaries improve memory for text? *Educational Psychology Review, 28*, 171–196.

Taylor, D. M. 2013. Writing rubrics as formative assessments in an elementary classroom. *Education and Human Development Master's Theses, Paper*, *258*.

Trabasso, T., & Bouchard, E. 2002. Teaching readers how to comprehendtexts strategically. In C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 176–200). New York, NY: Guilford Press.

van Dijk, T. A., & Kintsch, W. 1983. *Strategies of discourse comprehension* (pp. 11–12). New York, NY: Academic Press.

Wade-Stein, D., & Kintsch, E. 2004 Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333–362.

Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. 2010. Summarizing expository texts. *Topics in Language Disorders, 30*, 275–287.