Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Automated Assessment of Comprehension Strategies from Self-explanations Using Transformers and Multi-task

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Nicula, Bogdan	University Politehnica of Bucharest, Fotonatior	
Panaite, Marilena	University Politehnica of Bucharest, Romania	
Arner, Tracy	Arizona State University, USA	
Balyan, Renu	SUNY at Old Westbury, USA	
Dascalu, Mihai	University Politehnica of Bucharest, Academy	
McNamara, Danielle S.	Arizona State University, USA	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2023

Check type of content being submitted and complete one of the following in the box below:

- ☐ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☑ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Ir

DOI or URL to published work (if available) https://doi.org/10.1007/978-3-031-36336-8_107

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** National Center for Educat through **[Grant number]** R305A1900 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** National Center for Educa or the U.S. Department of Education.

Automated Assessment of Comprehension Strategies from Self-explanations Using Transformers and Multi-task Learning

Bogdan Nicula[1,2], Marilena Panaite[1], Tracy Arner[3], Renu Balyan[4], Mihai Dascalu[1,5(✉)], and Danielle S. McNamara[3]

[1] University Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania
{bogdan.nicula,marilena.panaite,mihai.dascalu}@upb.ro
[2] Fotonation SRL, Blvd. Timisoara, Nr. 4A, 061328 Bucharest, Romania
[3] Department of Psychology, Arizona State University, PO Box 871104, Tempe, AZ 85287, USA
{tarner,dsmcnama}@asu.edu
[4] Math/CIS Department, SUNY at Old Westbury, Old Westbury, NY 11568, USA
balyanr@oldwestbury.edu
[5] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania

Abstract. Self-explanation practice is an effective method to support students in better understanding complex texts. This study focuses on automatically assessing the comprehension strategies employed by readers while understanding STEM texts. Data from 3 datasets (N = 11,833) with self-explanations annotated on different comprehension strategies (i.e., bridging, elaboration, and paraphrasing) and an overall quality score was used to train various machine learning models in both single-task and multi-task setups. Our end-to-end neural architecture considers RoBERTa as an encoder applied to the target and self-explanation texts, combined with handcrafted features for assessing text cohesion and filtering out low-quality examples. The best configuration obtained a .699 weighted F1-score for the overall self-explanation quality.

Keywords: Self-explanations · Language models · Multi-task learning

1 Introduction

The ability to read and comprehend text is a critical skill for learners to be successful in their educational and future career goals. However, recent assessment data suggest that only 33% of fourth-graders and 31% of eighth-graders in the United States are at or above proficiency when reading and comprehending grade-level texts [11]. More importantly, students' level of proficiency has decreased significantly since the global pandemic forced instruction to occur in online environments [1].

N. Wang et al. (Eds.): AIED 2023, CCIS 1831, pp. 695–700, 2023.
https://doi.org/10.1007/978-3-031-36336-8_107

Reading comprehension is the ability to obtain meaning from text. Prior to deriving meaning from a text, readers must first decode letters and corresponding sounds to create words, access the meaning of the words, and then combine the words to create meaning from the text. As such, reading comprehension is a complex phenomenon that has been described with many theoretical frameworks [10] that vary in their explanation of these processes.

Explaining text to oneself can serve to scaffold inference generation and improve the coherence of the reader's mental representation of the text. As such, considerable evidence indicates that explicit self-explanation reading training (SERT) improves comprehension, particularly for low-knowledge and less skilled readers [8]. SERT includes instruction on five sub-strategies; comprehension monitoring, paraphrasing, bridging, prediction, and elaboration. Comprehension monitoring is a critical skill for students to recognize where they have gaps in knowledge. The use of prediction has been shown to be less common in the self-explanation of science texts. Therefore, in this study, we focus on paraphrasing, bridging, and elaboration.

Paraphrasing involves restating text in one's own words. This process is important in developing the reader's textbase level understanding by prompting them to think about the words and select words from their own lexicon that are more familiar and, thus, easier to understand and remember. *Bridging* is a critical strategy in the development of the reader's mental model of the text because authors cannot include every piece of relevant information. Therefore, readers must link ideas in the text to fill in the missing information. *Elaboration* is similar to bridging in that the reader is linking ideas from the text. However, elaborative inferences involve the reader connecting information from the text to their own knowledge base.

The aim of this work is to develop an automated model capable of evaluating the overall quality of self-explanations as well as the comprehension strategies employed by the readers.

2 Method

2.1 Corpus

The present study includes three different datasets collected at a large university in the Southwest United States. Participants in each of the studies were provided with training that was either direct instruction or worked examples of different methods of generating self-explanations (i.e., paraphrasing, bridging, elaboration). The direct instruction training consisted of short vignettes with descriptions and worked examples that demonstrated each method of generating self-explanations. The worked examples for training included a target sentence and an example response that highlighted each method participants could use to produce a self-explanation. In some studies, the training was presented in the context of a larger study using the iSTART Intelligent Tutoring System [9]. Following the training phase, participants were instructed to generate self-explanations of target sentences in relatively complex science texts.

In total, the three datasets contain 11,833 entries. Each datapoint consists of the participant's self-explanation, the target sentence, and scores for the presence of a paraphrase, bridge, or elaboration, as well as overall quality. These datasets have been split into train/dev/test by a ratio of 54.5%/27.5%/18%.

The 4 dimensions are represented as categorical variables having either 3 or 4 classes. The 0 class contains poor SEs, 1 contains acceptable SEs, whereas 2 and 3 contain good SEs. In the case of bridging and elaboration presence, the final 2 classes (marked with bold) were merged to minimize class imbalance (see Table 1).

Table 1. Class distribution per task.

Dimension	Num classes	Class 0	Class 1	Class 2	Class 3
Paraphrase presence	3	1487	1992	8354	
Bridge presence	4	4869	981	**4569**	**1414**
Elaboration presence	3	9382	**777**	**1674**	
Overall	4	799	4207	5093	1734

2.2 Neural Architecture

Our model consists of a combination of deep learning Natural Language Processing (NLP) [5] techniques and feature engineering. The core component of the model is the Transformer [13] block which processes the SE and source text pairs, thus generating a set of contextualized embeddings. Transformer-based architectures achieve state-of-the-art results in most NLP tasks. We consider both BERT [4] and RoBERTa [6] models as building blocks for our neural end-to-end architecture. Pretrained versions were loaded using the Huggingface library [14].

The embeddings generated by the Transformer block are combined with a set of handcrafted features that extract lexical, syntactical, and semantic information from both the SE and the source text. The aggregated features are processed via a fully connected layer. A smaller set of specific filtering features is generated separately to function as a mask to filter out part of the aggregated features and, in some cases, label the entire example as a 0. The resulting set of filtered features is used as input for the task classifiers in the multi-task learning setup, each having a single fully-connected layer. In the case of single-task training, the model has only one classifier.

A set of handcrafted features was generated to complement the representations generated from the Transformer block. These features were computed using the ReaderBench framework [3] for both the source text and the self-explanation and consisted of: lexical and semantic distance metrics, lexical and part-of-speech n-gram overlap, as well as surface, lexical, syntactic, semantic complexity indices.

The previously developed workflows for automatic self-explanation scoring [12] included a separate filtering step in the pipeline for poorly scored predictions (label 0). Based on the zero prediction rules from this filtering step, a smaller set of handcrafted features were generated to filter out irrelevant SEs. The filtering rules penalize SEs that are short, or contain copy-pasted text or frozen expressions.

RoBERTa models consistently outperformed the BERT alternatives. As such, our final configuration considered RoBERTa. Out of the 124.8M parameters, 180,912 parameters were trained at a max learning rate of 2e−4, while the rest of the 124.6M parameters (all related to the pretrained Transformer model) were trained with a smaller LR of 1e−5. A linear learning rate scheduler with warmup was used. The learning rate was gradually increased in the first five epochs up to the maximum value of 2e-4 and then gradually decreased to 0. The model was trained for 25 epochs using an AdamW optimizer [7] and a CrossEntropy loss for each task. A weighted CrossEntropy loss is used in order to account for the class imbalances present in all 4 tasks, as seen in Table 1.

Multi-task learning [2] is an approach by which a neural network can be trained simultaneously on multiple related tasks while using a shared representation. This type of training can improve generalization and is more efficient resource-wise than training one model per task. For the multi-task setup, multiple weighting schemes were considered for the individual task losses. Only the best-performing weighting scheme was used in the final experiment. This scheme considered the overall dimension to be as important as the sum of the other 3, and the paraphrase and bridging presence dimensions being twice as important as elaboration presence.

3 Results

The experiments done as part of this study were meant to determine the best model for assessing the 4 tasks. A second objective was that of determining whether a multi-task model could obtain similar or better performances than the 4 single-task models while training with similar resource constraints. For this reason, both the multi-task and the single-task models were trained for the same amount of time (25 epochs), using the same batch size, learning rate scheduler, optimizer, and regularization settings. The models are trained using a CrossEntropy loss, or a weighted sum of CrossEntropy losses, in the case of the multi-task model. They are evaluated throughout the run on the test set by computing the F1-score (see Table 2).

One aspect that is made clear when looking at the results is that the weighted Cross-Entropy loss does not completely address the issue of class imbalance. In the case of elaboration, we see that there is a 1-to-13 ratio between class 1 and class 0 on the test set. This offers an explanation for the large discrepancy in F1- scores for both models on this task, with the F1-score for class 0 being close to 0.9, while the F1-score for class 0 is barely 0.2. A similar but less extreme phenomenon can be observed for other tasks (e.g., class 1 has considerably fewer samples than classes 0 and 2 for bridging presence).

Table 2. Classification results for single-task, multi-task, and legacy models.

Strategy	Support	Model	F1-score				
			Weighted	Class 0	Class 1	Class 2	Class 3
Paraphrase presence	(90/141/1911)	STL	0.843	0.32	0.14	0.92	
		MTL	0.818	0.21	0.23	0.89	
Bridge presence	(859/24/1259)	STL	0.785	0.71	0.08	0.85	
		MTL	0.765	0.66	0.03	0.85	
Elaboration presence	(1996/146)	STL	0.899	0.95	0.21		
		MTL	0.882	0.93	0.72		
Overall	(61/688/926/467)	STL	0.694	0.24	0.73	0.69	0.71
		MTL	**0.699**	0.25	0.73	0.7	0.71

* STL - single-task learning; MTL - multi-task learning

When looking at both per-class and weighted F1-scores, we observe that there are very few differences between the multi-task and single-task models (i.e., the difference between the 2 models is smaller than 0.025 F1-score in all 4 cases). The single-task model seems to perform slightly better on the 3 tasks involving comprehension strategies, while the multi-task model has a better performance on the overall task.

4 Conclusions and Future Work

The aim of this paper was to develop a machine learning model to assess self-explanations in terms of an overall quality score and three individual scores that denote the presence of specific comprehension strategies. The model receives the raw text of the SE and of the source sentence and computes the four scores. Furthermore, the model is designed to automatically classify empty, illegible, or bad examples as class zero without requiring a separate filtering stage in its pipeline. Two separate approaches for meeting these criteria were developed. Firstly, 4 single task models were trained, each for predicting one of the scores. Secondly, a multi-task model was trained to solve the four tasks simultaneously, using an analogous architecture and training regimen.

The best result when predicting the overall quality score was a 0.699 F1-score obtained by the multi-task model, which was only slightly better than the 0.694 F1-score obtained by the single-task model. On the other three tasks, the single-task models obtained slightly better results, but the difference between single-task and multi-task training was never larger than 0.025.

Both single-task and multi-task models consisted of a similar number of parameters (i.e., 124.8M) and were trained using the same optimizer, the same learning rate scheduler, the same batch size, and an identical number of epochs. This underlines the value of the multi-task model, which managed to obtain similar or even better performance as the 4 single-task models while requiring only a quarter of the resources.

The current study is a significant advance in the ability to provide feedback to short responses to multiple science texts (i.e., any text, any time), both in terms of the specific strategy used by the reader, as well as the overall quality of the self-explanation. This was achieved by leveraging a larger dataset than available previously combined with recent advances in the field of AI.

Acknowledgements. This work was supported by the Ministry of Research, Innovation, and Digitalization, project CloudPrecis, Contract Number 344/390020/06.09.2021, MySMIS code: 124812, within POC, the Ministry of European Investments and Projects, POCU 2014–2020 project, Contract Number 62461/03.06.2022, MySMIS code: 153735, the IES (NSF R305A130124, R305A190063), the U.S. Department of Education, and the NSF (NSF REC0241144; IIS-0735682).

References

1. Adedoyin, O.B., Soykan, E.: Covid-19 pandemic and online learning: the challenges and opportunities. In: Interactive Learning Environments, pp. 1–13 (2020)
2. Caruana, R.: Multitask learning. Mach. Learn. **28**, 41–75 (1997)
3. Dascalu, M., Crossley, S., McNamara, D., Dessus, P., Trausan-Matu, S.: Please ReaderBench This Text: A Multi-dimensional Textual Complexity Assessment Framework, pp. 251–271. Nova Science Publishers Inc., New York (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the NAACL, vol. 1, pp. 4171–4186 (2019)
5. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Prentice Hall, Upper Saddle River (2009)
6. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019), http://arxiv.org/abs/1907.11692
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
8. McNamara, D.S.: SERT: self-explanation reading training. Discourse Process. **38**(1), 1–30 (2004)
9. McNamara, D.S., Levinstein, I.B., Boonthum, C.: iSTART: interactive strategy training for active reading and thinking. Behav. Res. Methods Instrum. Comput. **36**(2), 222–233 (2004)
10. McNamara, D.S., Magliano, J.: Toward a comprehensive model of comprehension. Psychol. Learn. Motiv. **51**, 297–384 (2009)
11. National Assessment of Educational Progress: The nation's report card: mathematics and reading at U.S. department of education, institute of education sciences. National Center for Education Statistics (2022). https://www.nationsreportcard.gov/reading/nation/scores/
12. Panaite, M., et al.: Bring It on! Challenges encountered while building a comprehensive tutoring system using *ReaderBench*. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 409–419. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93843-1_30
13. Vaswani, A., et al.: Attention is all you need. CoRR abs/1706.03762 (2017). http://arxiv.org/abs/1706.03762
14. Wolf, T., et al.: Transformers: State-of-the-art natural language processing. In: EMNLP (2020)