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

Multitask Summary Scoring with Longformers

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
Botarleanu, Robert-Mihai	University Politehnica of Bucharest	
Dascalu, Mihai	University Politechnica of Bucharest/Academy	
Allen, Laura K.	University of Minnesota/ Department of Educa	
Crossley, Scott Andrew	Georgia State University/Department of Applie	
McNamara, Danielle S.	Arizona State University/Department of Psych	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2022

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

AIED 2022, LNCS 13355

DOI or URL to published work (if available) https://doi.org/10.1007/978-3-031-11644-5_79

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]** R305A1801 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Arizona State University or the U.S. Department of Education.

Multitask Summary Scoring with Longformers

Robert-Mihai Botarleanu[1], Mihai Dascalu[1,2(✉)], Laura K. Allen[4], Scott Andrew Crossley[3], and Danielle S. McNamara[5]

[1] University Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania
robert.botarleanu@stud.acs.upb.ro, mihai.dascalu@upb.ro
[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania
[3] Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA 30303, USA
scrossley@gsu.edu
[4] University of Minnesota, Department of Educational Psychology, 250 Education Sciences Bldg, 56 E River Rd, Minneapolis, MN 55455, USA
lallen@umn.edu
[5] Department of Psychology, Arizona State University, 871104, Tempe, AZ 85287, USA
dsmcnama@asu.edu

Abstract. Automated scoring of student language is a complex task that requires systems to emulate complex and multi-faceted human evaluation criteria. Summary scoring brings an additional layer of complexity to automated scoring because it involves two texts of differing lengths that must be compared. In this study, we present our approach to automate summary scoring by evaluating a corpus of approximately 5,000 summaries based on 103 source texts, each summary being scored on a 4-point Likert scale for seven different evaluation criteria. We train and evaluate a series of Machine Learning models that use a combination of independent textual complexity indices from the ReaderBench framework and Deep Learning models based on the Transformer architecture in a multitask setup to predict concurrently all criteria. Our models achieve significantly lower errors than previous work using a similar dataset, with MAE ranging from 0.10–0.16 and corresponding R^2 values of up to 0.64. Our findings indicate that Longformer-based [1] models are adequate for contextualizing longer text sequences and effectively scoring summaries according to a variety of human-defined evaluation criteria using a single Neural Network.

Keywords: Natural language processing · Text summarization · Automated summary scoring · Multitask learning

1 Introduction

Summary scoring is a common task in education that requires a significant amount of attention and time, but that represents a crucial skill for students since it evaluates their ability to discern the primary message of texts. Evaluating a summary requires the rater to read both the source text and the summary and then evaluate the extent to which the summary captures the essence of the source text in a concise manner. Summarization has been shown to be an important aspect of reading comprehension and learning to read [2–4], for both first and second language learners [5, 6].

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13355, pp. 756–761, 2022.
https://doi.org/10.1007/978-3-031-11644-5_79

Automated summary scoring can help reduce the load on teachers and represents a method to provide immediate feedback to students on the quality of their summaries. Semi-automatic methods, such as having humans select portions of texts to be evaluated [7, 8], have been proposed; however, they do not significantly lower the time impact for teachers. Automated methods involve the use of various NLP techniques for text representation, including similarities using Latent Semantic Analysis [9], word embeddings and linguistic indices generated by the ReaderBench framework [10], or divergences among probability distributions [11]. However, much work on automated summary evaluation systems has focused on evaluation methods for automatically generated summaries, and not for scoring summaries written by humans.

The current study expands on Botarleanu, Dascalu, Allen, Crossley and McNamara [10] where textual complexity indices were used to train a summary scoring system that measured how well a human-written summary covers the main idea of the original source text. We build on this work by using a larger corpus of summaries and build regressors to predict seven different summary evaluation criteria. The regressors implemented in the current study can handle source texts of relatively large lengths, which was the main limitation of Botarleanu et al. [10]. This study aims to answer the following research questions:

1. How well do Longformer-based architectures, as compared to the linguistic indices used in Botarleanu et al. [10], perform in automatically scoring summary elements?
2. What is the performance of a multi-task learning model that predicts all 7 scoring criteria simultaneously in contrast to 7 individual models?

2 Method

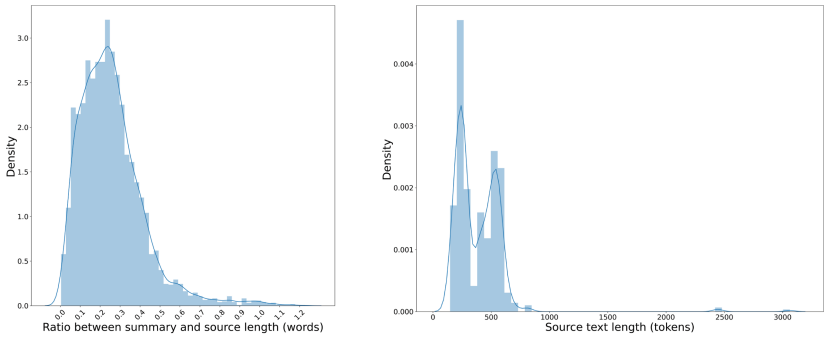
2.1 Corpora

Our corpus is an expanded version of the corpus considered by Botarleanu et al. [10] and includes 5,037 summaries (instead of the 2,976 previously used) corresponding to 103 source texts (instead of the 87 found in the aforementioned work). Our corpus was rated on a 1 to 4 Likert scale by expert raters for seven different scoring criteria: the cohesiveness of the summary text ("cohesion"), the appropriate use of objective language ("Objective Language"), the appropriate use of new paraphrasing ("Paraphrasing"), the use of language beyond that found in the source text ("Language Beyond Source Text"), how appropriate the length of the summary is in relation to the source text ("Summary Length"), the degree to which important details are captured from the source text ("Details") and whether the summary succeeds in capturing the main point of the reference text ("Main Point").

The corpus consists of summaries collected from a mix of unrelated studies: a) summaries collected in a study on Adult Literacy on general topics such as seat belt laws, disability services, and patients' rights, b) summaries collected using Amazon's Mechanical Turk service (MTurk) on science texts related to biology and climatology, c) summaries on heart disease and red blood cells collected in a study on Adult Literacy using MTurk, d) summaries on science texts collected using MTurk from primarily speakers of English as a second language, e) summaries on cellphone risk and climate

change were collected as part of a study on multiple text comprehension and f) summaries on science and history written by undergraduate students.

In order to filter out summaries that were malformed, we searched for summaries that were either as long as the source text or significantly shorter than it. A plot showing the ratio between the source and summary text lengths is presented in Fig. 1.a. We elected to remove all summaries with lengths below 10% or above 80% of the source text length. These values were chosen to remove the tails of the distribution from Fig. 1.b which depicts the ratios between summary length and source text length. This pro-cess reduced the number of summaries from 5,037 to 4,233 without removing any of the 103 source texts.



(a) (b)

Fig. 1. a) Ratio between the number of words between summaries and source texts. b) Source text lengths in tokens.

Another consideration in corpus development is the length of the sequences that are used as inputs for transformer-based models. Due to internal constraints, models such as BERT (Devlin et al. 2018) are only suitable for sequences of up to 512 tokens, whereas models such as Longformer can work well with longer sequences, typically of up to 4,096 tokens. We utilized the pre-trained Longformer tokenizer provided by the "transformers" package [12] with the "allenai/longformer-base-4096" pretrained model to evaluate the lengths of the texts in our corpus (see Fig. 1.b). Indeed, a significant proportion of source texts in our corpus exceeded 512 tokens; however, the majority did not exceed 1000 tokens in length making the texts suitable for the Longformer model, but not for BERT.

2.2 Regression Models

We elected to construct our regression models around the Longformer model [1] to handle the source and summary texts that were often too long for BERT. The Longformer model employs an attention mechanism that combines local windowed attention with a global attention mechanism, that is designed to encode inductive bias about the task that the model is being trained to solve. Given the distribution in Fig. 1.b, we opted to use padded sequences of 2048 tokens formed by tokenizing and trimming both the

source and summary lengths down to 1024 tokens. The overview of this architecture is illustrated in Fig. 2. The architecture is evaluated under two setups: one where the model predicts only one of the seven summary scoring criteria at a time, and a second where the model is tasked with predicting all seven objectives at once.

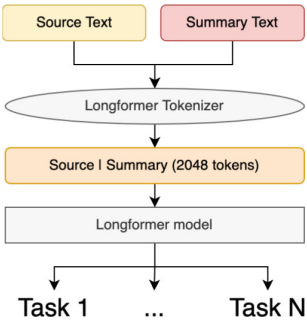


Fig. 2. The architecture of the Longformer-based model.

3 Results

We present the normalized Mean Absolute Error (nMAE) and the corresponding R^2 coefficients for the Longformer model used in both single-task and multi-task settings (see Table 1). We compare our results to those measured using the model presented by Botarleanu et al. [10], where a network with a single hidden layer with 256 units is applied to an input consisting of the textual complexity indices generated by the ReaderBench framework for both the summary and the source texts. This was trained using the same Once-Cycle Policy described in Botarleanu et al. [10], for 50 epochs with a batch size of 8.

The first observation is that the trained Longformer models outperform the models relying on the ReaderBench indices, with R^2 coefficients having values that are between .07 and .13 higher. Second, the multi-task model matches the performance of the individual models on average, and even exceeds the single-task models for the "Cohesion", "Language Beyond Source Text", and "Details" criteria. The most significant degradation in performance between the single-task and the multi-task setting is for the "Paraphrasing" score with the R^2 coefficient falling from .55 to .42, which is still higher than the performance of the model that uses ReaderBench indices.

Finally, the seven scoring criteria appear to have a relatively similar difficulty in terms of the models' capability to learn them. The lowest R^2 coefficient for the Longformer-based models is measured on the "Details" objective (.37 for the single task model), whereas the highest coefficient is observed for the "Summary Length" objective (.67 for the single task model). In contrast, the multi-task model appears to have a narrower variation in performance, with R^2 coefficients ranging from .42 for the "Paraphrasing" objective up to .64 for the "Summary Length" criterion, which may be explained by the fact that the multi-task model was trained with all 7 criteria being seen as equal, and no objective-specific weights were applied to the loss function.

Table 1. Normalized MAE and the R^2 score for the seven evaluation criteria.

Scoring criterion	Single task models using ReaderBench indices [10]		Multi-task model using ReaderBench indices [10]		Single task Longformer models		Multi-task Longformer model	
	nMAE	R^2	nMAE	R^2	nMAE	R^2	nMAE	R^2
Cohesion	.14	.49	.14	.46	.13	.52	**.12**	**.58**
Objective language	.13	.51	.14	.48	**.11**	**.59**	.13	.50
Paraphrasing	.15	.45	.15	.50	**.13**	**.55**	.16	.42
Language beyond ource text	.13	.43	.13	.47	.10	.59	**.10**	**.60**
Summary length	.13	.54	.14	.51	**.11**	**.67**	.12	.64
Details	.15	.46	.16	.39	.15	.37	**.13**	**.53**
Main point	.15	.53	.14	.52	.11	**.64**	**.12**	.59
Average	*.14*	*.49*	*.14*	*.48*	***.12***	***.58***	*.13*	*.55*

4 Conclusions and Future Work

In this paper, we analyzed the effectiveness of using Longformer-based regression models to perform automated summary scoring. Our models achieved significantly better results than the previous models in Botarleanu et al. [10]. Our results also indicate that a model trained in a multi-task setting achieved a performance that was on par with training seven different networks. With an average normalized mean absolute error of .13 and a corresponding R^2 of .55, our model predicts the human rating of a summary with an average deviation of 13%. The capability to perform automated summary scoring in a multi-task setting has several advantages. First, it reduces the computational load and supports the development of automated summary scoring systems that can analyze summaries more effectively. Second, it more closely matches the human expert scoring method because it forces the model to perform a holistic analysis of the text, instead of relying on patterns captured for each of the scoring criteria individually. One method of improving the performance of the model might be to combine the ReaderBench indices with the Longformer inferences into an ensemble model. Moreover, part of the ReaderBench indices might also benefit from the use of Longformer models to predict their values (e.g., intra- and inter- paragraph cohesion scores).

A potential avenue for future research lies in performing an interpretability analysis of the multi-task model. Through this, one might explore the degree to which the different summary scoring criteria presented in this work may complement each other. Additionally, studying the way in which the most relevant blocks of the summaries and source texts are selected by the model, and aligning these segments with human rater

observations may provide valuable insight into what humans look for in summaries, which can help in providing targeted feedback to students.

Finally, the principal measure of the usefulness of such a system lies in the impact it has on the summarization skills of real students. An important future research direction would be the use of the model described in this work to help students improve their summary writing skills. Notably, the success of this model in predicting scores on seven different attributes, for such a wide range of source texts, bodes well for the eventual utility of this model within an automated tutoring system.

Acknowledgments. This research was supported by a grant from the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number TE 70 *PN-III-P1-1.1-TE-2019-2209,* ATES – "Automated Text Evaluation and Simplification", the Institute of Education Sciences (R305A180144 and R305A180261), and the Office of Naval Research (N00014-17-1-2300; N00014-20-1-2623; N00014-19-1-2424, N00014-20-1-2627). The opinions expressed are those of the authors and do not represent the views of the IES or ONR.

References

1. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: the long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
2. Bensoussan, M., Kreindler, I.: Improving advanced reading comprehension in a foreign language: summaries vs. short-answer questions. J. Res. Read. **13**(1), 55–68 (1990)
3. Brown, A.L., Campione, J.C., Day, J.D.: Learning to learn: on training students to learn from texts. Educ. Res. **10**(2), 14–21 (1981)
4. Bean, T.W., Steenwyk, F.L.: The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. J. Read. Behav. **16**(4), 297–306 (1984)
5. Karbalaei, A., Rajyashree, K.S.: The impact of summarization strategy training on university ESL learners' reading comprehension. Int. J. Lang. Soc. Cult. **30**, 41–53 (2010)
6. Pakzadian, M., Rasekh, A.E.: The effects of using summarization strategies on Iranian EFL learners' reading comprehension. Engl. Linguist. Res. **1**(1), 118–125 (2012)
7. Nenkova, A., Passonneau, R.J.: Evaluating content selection in summarization: the pyramid method. In: The Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Boston, Massachusetts, USA, pp. 145–152. ACL (2004)
8. Van Halteren, H., Teufel, S.: Examining the consensus between human summaries: initial experiments with factoid analysis. In: HLT-NAACL 03 Text Summarization Workshop, Edmonton, Canada, pp. 57–64. ACL (2003)
9. Steinberger, J., Jezek, K.: Using latent semantic analysis in text summarization and summary evaluation. In: Proceedings of the 7th International Conference ISIM, pp. 93–100 (2004)
10. Botarleanu, R.-M., Dascalu, M., Allen, L.K., Crossley, S.A., McNamara, D.S.: Automated summary scoring with ReaderBench. In: Cristea, A.I., Troussas, C. (eds.) Intelligent Tutoring Systems. Lecture Notes in Computer Science, vol. 12677, pp. 321–332. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80421-3_35
11. Torres-Moreno, J.-M., Saggion, H., Cunha, I.D, SanJuan, E., Velázquez-Morales, P.: Summary evaluation with and without references. Polibits **42**, 13–20 (2010)
12. Facebook Inc.: Transformers (n.d.). https://huggingface.co/docs/transformers/index. Accessed 20 Jan 2022