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 Paraphrase Quality Assessment using Recurrent Neural Networks and Language Models

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	
Dascalu, Mihai	Academy of Romanian Scientists	
Newton, Natalie	Arizona State University	
Orcutt, Ellen	University of Minnesota	
McNamara, Danielle	Arizona State University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) July 9, 2021

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

International Conference on Intelligent Tutoring Systems, July 9, 2021, Athens, Greece

DOI or URL to published work (if available) doi.org/10.1007/978-3-030-80421-3_36

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]** Institute of Education Scien through **[Grant number]** R305A19006 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute of Education Scie or the U.S. Department of Education.

Automated Paraphrase Quality Assessment using Recurrent Neural Networks and Language Models

Bogdan Nicula[1], Mihai Dascalu[1,2], Natalie Newton[3], Ellen Orcutt[4], Danielle S. McNamara[3]

[1] University Politehnica of Bucharest, 313 Splaiul Independentei, 060042, Bucharest, Romania
bogdan.nicula@stud.acs.upb.ro, mihai.dascalu@upb.ro
[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044, Bucharest, Romania
[3] Arizona State University, Department of Psychology, PO Box 871104, Tempe, AZ 85287
nnnewton@asu.edu, dsmcnama@asu.edu
[4] University of Minnesota, Department of Educational Psychology, 56 East River Road, Minneapolis, MN, 55455
orcut039@umn.edu

Abstract. The ability to automatically assess the quality of paraphrases can be very useful for facilitating literacy skills and providing timely feedback to learners. Our aim is twofold: a) to automatically evaluate the quality of paraphrases across four dimensions: lexical similarity, syntactic similarity, semantic similarity and paraphrase quality, and b) to assess how well models trained for this task generalize. The task is modeled as a classification problem and three different methods are explored: a) manual feature extraction combined with an Extra Trees model, b) GloVe embeddings and a Siamese neural network, and c) using a pre-trained BERT model fine-tuned on our task. Starting from a dataset of 1998 paraphrases from the User Language Paraphrase Corpus (ULPC), we explore how the three models trained on the ULPC dataset generalize when applied on a separate, small paraphrase corpus based on children inputs. The best out-of-the-box generalization performance is obtained by the Extra Trees model with at least 75% average F1-scores for the three similarity dimensions. We also show that the Siamese neural network and BERT models can obtain an improvement of at least 5% after fine-tuning across all dimensions.

Keywords: Paraphrase Quality Assessment, Natural Language Processing, Recurrent Neural Networks, Language Models

1 Introduction

A paraphrase is a restatement, generated with different words, of the meaning of a text, generally with the aim of clarifying a sentence or a small group of sentences. Paraphrasing is useful for a number of purposes and applications. For example, in Natural Language Generation, automated paraphrases are a method to increase diversity of generated text [1] and recognition of queries [2]. By contrast, our focus is on developing algorithms to assess the quality of human-generated paraphrases in order to provide

feedback to students who are learning how to paraphrase more effectively and efficiently. Encouraging readers to transform a source text into more familiar words and phrases helps the reader to better understand the text by activating relevant prior knowledge. Learning to paraphrase facilitates both reading comprehension and writing ability, particularly for less skilled readers and writers [3, 4, 5]. Thus, paraphrase assessment is used in Intelligent Tutoring Systems aimed at improving reading and writing.

Our overarching objective is to develop feedback for a new version of iSTART (Interactive Strategy Training for Active Reading and Thinking; [6]), called iSTART-Early for young developing readers (ages 9-11). iSTART provides adaptive instruction and practice to use comprehension strategies (e.g., elaboration, bridging), while self-explaining and reading science texts to improve low-knowledge and less skilled readers' comprehension of challenging texts and performance in science courses.

The aim of this work is to assess the generalization capability of these models. First, we analyze the performance obtained by an Extra Trees model, a Siamese neural network model [7], and a BERT-based model [8] when trained on the ULPC dataset and evaluated on a different dataset. Second, we assess the importance of fine-tuning in improving results for the Siamese neural network and BERT models.

2 Related Work

One of the most well-known datasets for paraphrase identification is the Microsoft Research Paraphrase Corpus (MSRP) [9]. Given its relatively small size (5801 sentence pairs out of which 66.5% are positive examples), some of the best results on this dataset were obtained by small models. One example consists of using SWEMs (Simple-Word-Embedding-Models) [10], which rely on aggregating word embeddings via simple pooling operations (e.g., max pooling, average pooling). Another successful approach by Ji and Eisenstein [11] was to use a combination of fine-grained overlap features (e.g., unigram, bigram and dependency relation overlap metrics) and latent sentence-level features extracted using matrix factorization and a term weighting approach based on KL-divergence, called TF-KLD.

At the opposite end of the spectrum is the Quora Question Pair dataset (QQP), which consists of 400,000 question pairs with a binary annotation for paraphrasing. This dataset represents a good fit for data-hungry deep learning NLP models. A significant number of the current top performing models are based on the highly successful Bidirectional Encoder Representations from Transformers (BERT) model. Some approaches focus on reducing the size of the BERT model, while extracting the maximum performance from it [12], while others introduce innovative masking techniques for improving BERT performance [13]. Lastly, there are also models with similarly good performance that do not rely on BERT at all [14] as they create a custom neural network, that is considerably faster and uses much fewer parameters than classic BERT models with GloVe [15] word embeddings.

Despite the differences in style and content quality, both types of datasets (i.e., MSRP vs. QQP) share one shortcoming: they provide very little information regarding

the *quality* of the paraphrase, as they solely indicate whether a given pair of sentences constitute paraphrases of one another. To our knowledge, the sole dataset that includes rubric scores regarding quality is the User Language Paraphrase Corpus (ULPC) [16], which scores paraphrases on 10 aspects using a point range from 1 to 6. We leverage this corpus in order to develop and test algorithms to assess paraphrase quality, and then to test the far transfer of these algorithms to paraphrases generated by young developing readers ages 9-11.

3 Method

3.1 Corpus

Two datasets were used as part of this work: the ULPC dataset consisting of 1998 source text – paraphrase pairs, and one smaller dataset, containing 115 paraphrases generated by children aged 9-11. The two datasets will be referred to as ULPC and the children dataset. The ULPC dataset consists of source texts – paraphrase pairs that were extracted from the input that users provided for the iSTART intelligent tutoring system (ITS). The children dataset is composed of paraphrase responses from a group of 13 3rd and 4th grade children participating in a summer school program. Notably, all students participants were English Language Learners. The paraphrase – sentence pairs in both datasets were scored in terms of the following four dimensions: semantic similarity, syntactic similarity, lexical similarity, and paraphrase quality.

For the ULPC dataset, the raters assigned scores ranging between 1 and 6 for each dimension. The four dimensions were then categorized into binary (1.00-3.49 vs 3.5-6.00), or tripartite (1.00-2.66, 2.67-4.33, 4.33-6.00) evaluations. For the children dataset, the four dimensions were originally scored on a binary system, except for paraphrase quality, which was scored on a tripartite scale. In order to have the same approach for both datasets, the problem was modeled as binary classification for semantic, syntactic and lexical similarity, whereas a tripartite classification was used for paraphrase quality.

3.2 Classification Models

Three different classification models were used for these experiments: An Extra Trees model combined with manually engineered features (ET), a Siamese neural (SN) network, and a BERT-based model. Out of the many possible options, these three alternatives were chosen to establish a strong baseline comparing systems relying heavily on manually engineered features versus deep learning systems, as well as lightweight (SN) versus resource intensive (BERT) models.

For the ET model, several types of features based on the sentence-paraphrase pairs were used: a) complexity indices related to surface, lexical, syntactic, and semantic properties of the texts were computed using the ReaderBench framework [17]; b) complexity indices outlining text cohesion were computed on the concatenation of the source text and the paraphrase, and c) Levenshtein distance [18] at word level between the source and paraphrase, as well as simple overlap indices for both words and part-

of-speech (POS) tokens. For the ReaderBench complexity features, the difference between the value of the same index computed for both source and paraphrase was used. The resulting 2368 features were filtered in order to eliminate constant values and features with high intercorrelation. The filtered features were used as input for several ML classifiers from the SciKit Learn library [19] to predict one of the targeted four dimensions. In all four cases, the best results were obtained by the Extra Trees model.

For the Siamese neural network [7], Bidirectional Long Short-Term Memory (BiLSTM; [20]) layers were used with pretrained 300-dimensional GloVe or Word2Vec [21] word embeddings at the entry point in the architecture. Both the source sentence and the paraphrase were converted into an array of indices, each index pointing to an embedding representing the meaning of the corresponding word in the text. This representation was then processed separately (once for the source, once for the paraphrase) by the BiLSTM layers, and after a set of pooling operations, the two processed results were combined, and a prediction was made. The results are reported for the model using GloVe embeddings, as that model obtained a better performance.

For the BERT-based model, a pretrained version of BERT from the Huggingface library [22] was considered. In terms of the architecture, the source and paraphrase texts were passed as a text pair to the pretrained BERT model, delimited by a special BERT separator. The combined input was truncated if longer than a threshold of 75 words, and then converted into embeddings and passed through the BERT pipeline. The output of the BERT model went through a Dropout layer with a conservative p=0.2 dropout rate, and then a fully connected (FC) layer was used to make the final prediction. Different learning rates for the BERT model (lr_BERT=1e-5) and the FC layers (lr_FC=2e-2) were considered to make the fine tuning feasible.

4 Results and Discussions

Our first experiment involved examining accuracy of the models on the children dataset. ET, SN and BERT models were trained on the entire ULPC dataset (training + validation + testing) and tested on the entire children dataset (115 paraphrase pairs). The tripartite split (into low 1-2, mid 3-4, and high 5-6) was used for the Paraphrase Quality dimension, as it was available for both datasets.

The results provided in Tables 1 and 2 indicate that the Extra Trees model obtained the best results in 3 out of 4 cases (in terms of average weighted F1-score). On the three binary dimensions (semantic, syntactic, lexical similarity), the ET model consistently outperformed the SN and BERT models. Overall, the high performance of extra trees is beneficial, given the interpretability of the models relying on linguistic features reflective of writing style and on semantic relatedness between the paraphrase and the source text. The interpretability is beneficial because the features can guide feedback.

The poor overall performance obtained on the Paraphrase Quality task might be caused by the fact that the children paraphrases are more difficult to be split up into three classes, given the simplicity of the text (i.e., most answers are either fair paraphrase attempts or not paraphrases at all, and there is less room to be vague).

Table 1. Performance on ULPC models tested on Children dataset (Semantic similarity, Syntactic similarity and Lexical similarity).

Dimension	Model	Support low	Support high	Low F1	High F1	Avg F1
Semantic Similarity	ET	22	93	.706	.916	**.875**
	SN	22	93	.371	.725	.657
	BERT	22	93	.575	.802	.758
Syntactic Similarity	ET	35	80	.688	.776	**.749**
	SN	35	80	.444	.327	.362
	BERT	35	80	.530	.367	.416
Lexical Similarity	ET	31	84	.806	.929	**.895**
	SN	31	84	.422	.629	.573
	BERT	31	84	.689	.811	.778

Table 2. Performance on ULPC paraphrase quality models tested on children dataset.

Model	Support low	Support mid	Support high	Low F1	Mid F1	High F1	Avg F1
ET	24	60	31	.610	.708	.244	**.562**
SN	24	60	31	.333	.337	.205	.300
BERT	24	60	31	.454	.712	.000	.466

In the second experiment we evaluated the benefits of fine-tuning for the Siamese Network and BERT-based models. The models trained on the ULPC dataset were trained for a small number of epochs on 67 pairs from the children dataset and tested on the remaining 48 pairs. All examples containing the same source text were added to either the test or the training set, but not both. Because of the nature of the dataset (i.e., for a given source text there are a variable number of paraphrases), the children dataset could not be split into equal halves. In all the cases, the slightly larger half was used for training and the smaller one for validation.

The F1 scores for all the classes, as well as weighted average of the F1 scores, are reported in Table 3 and Table 4. This metric was computed for predictions made by a) the initial pretrained models (e.g., SN and BERT), and b) the pretrained models that were fine-tuned for a short number of epochs.

Table 3. Results obtained after fine-tuning on the children dataset (Semantic similarity, Syntactic similarity and Lexical similarity).

Dimension	Model	Pretrained			Finetuned		
		Low F1	High F1	Weighted Avg	Low F1	High F1	Weighted Avg
Semantic Similarity	SN	.303	.635	.572	.348	.795	.711
	BERT	.545	.761	.720	.5	.833	**.770**
Syntactic Similarity	SN	.370	.190	.238	.378	.610	.547
	BERT	.464	.25	.307	.619	.703	**.680**
Lexical Similarity	SN	.457	.689	.631	.435	.822	.725
	BERT	.666	.800	.766	.733	.878	**.841**

When looking at the individual F1 scores we tend to see improvements after fine-tuning in most cases. One notable exception is the High class for Paraphrase Quality for which both models have difficulties without pretraining, and BERT does not manage to obtain an F1 of over 0 even after fine-tuning, despite having non-zero scores on the training set.

When comparing the weighted F1 scores an improvement of at least .05 can be observed after fine-tuning. In the 2-class setting, the most dramatic improvements were made for the Syntactic similarity class. On this dimension, the distributions for the children dataset where almost inverted versions of the ULPC distribution. This could mean that the model had learned useful features in the initial training phase, but it relied on a bad estimate of the distribution for the classes.

Table 4. Results obtained after fine-tuning on the children dataset (Paraphrase Quality).

	Pretrained				Finetuned			
Model	Low F1	Mid F1	High F1	Weighted Avg	Low F1	Mid F1	High F1	Weighted Avg
SN	.359	.432	.200	.353	.385	.696	.500	**.578**
BERT	.461	.736	.000	.479	.476	.721	.000	.474

5 Conclusions

The aim of this study was to develop three ML algorithms to assess paraphrase quality leveraging the ULPC dataset and to evaluate how well these models generalize when presented with a new dataset of paraphrases generated by children. When tested on the children dataset, the Extra Trees model obtained the best results. The SN and BERT models also provided improved results after fine-tuning on the children dataset.

In the first generalization task, the Extra Trees model was shown to generalize better. This indicates that the manually extracted features might have a more general character than the ones automatically extracted by Siamese Networks or BERT-base models, making them more robust on new data. For the Semantic, Syntactic and Lexical similarity dimensions, the Extra Trees model generalized well on the children dataset. However, the results were slightly worse for the Paraphrase Quality dimension. This could indicate that it is more difficult to meaningfully separate poor, satisfactory, and good paraphrases for children, or it could indicate an issue with how the dataset was annotated (e.g., raters had difficulties separating the 3 levels of the dimension). In this case, results could be improved by adding a larger paraphrase dataset with similar characteristics (short source and paraphrase sentences) for initially training the models, followed by a finetuning on the ULPC dataset.

Fine-tuning helped improve results in all cases with differences ranging from 0.05 to 0.20. The BERT model fared better than the SN model in the binary classification tasks, while it underperformed when classifying paraphrase quality. Its poor performance was caused by its difficulties in predicting the High class.

This study provides promising evidence that our approach can generate models that generalize to texts that differ in reading ease and to individuals who vary in age, reading skill, and language abilities. While more evidence is needed to further test this approach and these models, these models provide a strong starting point in iSTART-Early for providing automated feedback on paraphrase quality to young developing readers.

Acknowledgments. The work was funded by a grant of 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". This research was also supported in part by the Institute of Education Sciences (R305A190063 and R305A190050) and the Office of Naval Research (N00014-17-1-2300 and N00014-19-1-2424). The opinions expressed are those of the authors and do not represent views of the IES or ONR.

References

1. Qian, L., Qiu, L., Zhang, W., Jiang, X., Yu, Y.: Exploring diverse expressions for paraphrase generation. 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), pp. 3164-3173 (2019)
2. Fader, A., Zettlemoyer, L., Etzioni, O.: Open question answering over curated and extracted knowledge bases. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1156-1165 (2014)
3. McNamara, D.S.: SERT: Self-Explanation Reading Training. Discourse Processes, 38, 1–30 (2004)
4. McNamara, D.S., Ozuru, Y., Best, R., O'Reilly, T.: The 4-pronged comprehension strategy framework. Reading comprehension strategies: Theories, interventions, and technologies, pp. 465-496. Erlbaum, Mahwah, NJ (2007)
5. Hawes, K.: Mastering academic writing: Write a paraphrase sentence. University of Memphis, Memphis: TN, (2003)
6. Jackson, G.T., McNamara, D.S.: Motivation and performance in a game-based intelligent tutoring system. Journal of Educational Psychology, 105(4), 1036 (2013)
7. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. arXiv, preprint arXiv:1404.2188 (2014)
8. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dolan, W.B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005) (2005)
10. Shen, D., Wang, G., Wang, W., Min, M.R., Su, Q., Zhang, Y., Li, C., Henao, R., Carin, L.: Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. arXiv preprint arXiv:1805.09843, (2018)
11. Ji, Y., Eisenstein, J.: Discriminative improvements to distributional sentence similarity. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 891-896 (2013)
12. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351, (2019)
13. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8, 64-77 (2020)
14. Yang, R., Zhang, J., Gao, X., Ji, F., Chen, H.: Simple and effective text matching with richer alignment features. arXiv preprint arXiv:1908.00300, (2019)
15. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), Vol. 14. ACL, Doha, Qatar (2014)

16. McCarthy, P.M., McNamara, D.S.: The user-language paraphrase challenge. Retrieved January, 10, 2008 (2008)
17. Dascalu, M., Crossley, S.A., McNamara, D.S., Dessus, P., Trausan-Matu, S.: Please ReaderBench this Text: A Multi-Dimensional Textual Complexity Assessment Framework. In: Craig, S. (ed.) Tutoring and Intelligent Tutoring Systems, pp. 251–271. Nova Science Publishers, Inc., Hauppauge, NY, USA (2018)
18. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady, 10(8), 707–710 (1965)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.J.J.o.m.l.r.: Scikit-learn: Machine learning in Python. 12(Oct), 2825–2830 (2011)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation, 9(8), 1735–1780 (1997)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representation in Vector Space. In: Workshop at ICLR, Scottsdale, AZ (2013)
22. Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45 (2020)