# Paraphrasing Academic Text: A Study of Back-translating Anatomy and Physiology with Transformers

Andrew M. Olney[0000−0003−4204−6667]

University of Memphis, Memphis TN 38152, USA
aolney@memphis.edu
https://olney.ai

**Abstract.** This paper explores a general approach to paraphrase generation using a pre-trained seq2seq model fine-tuned using a back-translated anatomy and physiology textbook. Human ratings indicate that the paraphrase model generally preserved meaning and grammaticality/fluency: 70% of meaning ratings were above 75, and 40% of paraphrases were considered more grammatical/fluent than the originals. An error analysis suggests potential avenues for future work.

**Keywords:** paraphrase · deep learning · natural language generation

## 1 Introduction

Paraphrasing is a core task in natural language processing (NLP) and has multiple educational applications, like essay grading [5], short answer assessment [11], text simplification [4] and plagiarism detection [1]. Recent developments in automated paraphrase have largely tracked advances in machine translation using neural networks, i.e., neural machine translation (NMT), primarily using the LSTM [8, 13, 17] and Transformer [10, 12, 14, 20] architectures. One approach to generating paraphrases is back-translation, by which a sentence is translated from a source language to a *pivot* language and back to the source language.

Paraphrasing academic text has its own challenges because it differs from normal text both in vocabulary and syntax, particularly in scientific domains [6, 16] and it is usually copyright-restricted and therefore difficult to obtain in quantities necessary for machine learning models. The present study addresses these problems through NMT back-translation and fine-tuning a recent Transformer variant called T5 [18]. Our primary research questions are therefore (1) how well the paraphrases preserve the meaning of the source text and (2) how grammatical and fluent are the paraphrases with respect to the source text.

## 2 Model & Human Evaluation

We conducted a small pilot study to determine the best pivot languages for paraphrasing anatomy and physiology. Randomly selected sentences (N=24) from

a textbook [19] were back-translated with different pivot languages using the Google Translate API. The paraphrases were evaluated by an expert judge on (1) the degree of change as none, word, or phrase (a measure of diversity) and (2) whether the paraphrase was disfluent or incorrect (a measure of acceptability). Results are presented in Table 1. Values are sentence counts except for weighted change, which weights word change counts by 1 and phrase change counts by 2.

An ideal pivot language would result in low unacceptability and high diversity. Our analysis suggests Czech introduces more changes at the word choice level, and Russian introduces marginally more changes at the phrasal level. On the intuition these properties may be additive, we conducted an additional evaluation using Czech and Russian as pivot languages together (English-Czech-Russian-English). As indicated by the results in the table, the combination appears to increase the weighted change above Czech and Russian individually without noticeably increasing error. Furthermore, the weighted change is comparable to most of the non-European pivot languages, which created substantially more unacceptable paraphrases. Based on these results, we back-translated the complete textbook (12,062 sentences) both with Czech as a pivot and with Czech-Russian as a double pivot, producing 24,124 source-paraphrase pairs.

Training and testing sets were prepared by aligning the two back-translations with the corresponding source and randomly selecting 90% of the 3-tuples for training and the remainder for test. These datasets were then augmented by permuting the 3-tuples to create combinations of all pairs in all orders. Pairs differing by less than 3 characters and sentences with less than 11 characters were excluded as noisy data. Augmentation resulted in 34,094 pairs in the training and 3,836 pairs in the test sets. The T5-BASE pre-trained model from the HuggingFace library [21] and fine-tuned using Pytorch with the training set for 8 epochs, though test set loss did not improve past epoch 4. The training process completed in approximately 3.5 hours using an NVIDIA 1080Ti GPU.

A human evaluation was conducted to determine the quality of the model-generated paraphrases, specifically (1) how well the paraphrases preserve the meaning of the source text and (2) how grammatical and fluent the paraphrases are with respect to the source text. Raters ($N = 29$) were recruited through the Amazon Mechanical Turk (AMT) marketplace between January and February of 2021, using the CloudResearch platform [15]. In this study, raters were required to be native English speakers and be employed as a nurse or physician. Raters were further required to have completed at least 100 previous AMT tasks with at least a 95% approval rating. Raters were paid $7.

A separate textbook on anatomy and physiology [2] from OpenStax was used as a source for sentences to paraphrase. The book was downloaded and preprocessed by splitting main body text into sentences, removing sentences that refer to figure and tables, removing parenthetical elements, performing Unicode to ASCII translation, and performing spelling correction. The final sentences contained ranges, slashes, formulas, and chemical symbols. Paraphrases of these sentences were then generated using the model.

Six surveys were created on Qualtrics, an online survey tool, using randomly

selected source-paraphrase pairs, each containing 100 pairs, as is common for this type of evaluation [7, 9, 3]. Each pair was formatted on a single survey page where the source text was formatted above the paraphrase, followed by two questions with slider-format response on a 0-100 scale. The first was a meaning-assessment question, "The paraphrase conveys the same meaning of the original," and was anchored by "not at all" on the left and "perfectly" on the right. The second was a fluency-assessment question, "Which is more grammatical and fluent?", with "original" on the left and "paraphrase" on the right. The sliders had no numeric indicators and were initialized at the midpoint. Following the direct assessment methodology [9, 7], 12 of each 100 were control pairs were created by copying an existing item (a survey page) and then degrading the paraphrase on that page by deleting a random span of words, where $span_{length} = 0.21696 * word_{count} + 0.78698$, rounded down, which linearizes existing rules [9]. Twelve pairs are sufficient to detect a large (.8 SD) effect using a Wilcoxon signed-ranks test for matched pairs at $\alpha = .05$ and .80 power with a one-tailed test. If we do not detect a large effect between ratings of distinct items and their degraded versions, we infer the rater is not reliable. The degraded items were randomly positioned based on the position of their matched item, modulo 44.

## 3 Results & Discussion

Subsets of raters passed control checks for meaning ($n = 35$) and fluency ($n = 23$), with $p < .05$ on the signed-ranks test, except for a fluency check on the 2nd survey, $p = 0.06$, which was allowed because its control items were more difficult to distinguish. Cronbach's alpha for passing raters was high ($\alpha > .85$), except survey 6, $\alpha = .66$, until two raters were dropped to obtain high agreement, $\alpha = .77$. The mean meaning rating was high ($M = 78.78$, $SD = 16.89$, $CI_{95} = [77.33, 80.22]$), and the mean grammaticality/fluency rating was less than the midpoint of 50 ($M = 43.97$, $SD = 20.75$, $CI_{95} = [42.19, 45.74]$). The distribution of each rating may be examined in Figure 1. The distribution for meaning illustrates that most paraphrases are rated as highly meaning preserving. The meaning distribution peaks at the most frequent rating of 89, and approximately 70% of all meaning ratings are above 75. The distribution for fluency reflects its anchoring at 50, at which point both the original (0) and paraphrase (100) are considered equally fluent. The grammaticality/fluency distribution is symmetric and peaks at a rating of 38, and approximately 40% of all grammaticality/fluency ratings are above 50, indicating that the paraphrase was considered more grammatical/fluent than the original sentence approximately 40% of the time.

Paraphrases associated with the lowest 5% of ratings for meaning and grammaticality/fluency were examined to determine common error types, four of which accounted for 76% of errors. Most common was the substitution of a near neighbor for the target, e.g. "membrane" for "diaphragm," and it more negatively impacted grammaticality/fluency than meaning. Second was the use of the wrong word sense for the target, e.g. "adults' volumes" for "volumes in adults," and more evenly affected both metrics. The third arose when the text

| Language | Change | | | | Err |
|---|---|---|---|---|---|
| | No | Wd | Ph | Wt | |
| Czech | 3 | 15 | 6 | 27 | 4 |
| Russian | 7 | 9 | 8 | 25 | 2 |
| Cz-Ru | 3 | 11 | 10 | 31 | 4 |
| Chinese | 2 | 11 | 11 | 33 | 9 |
| Persian | 2 | 14 | 8 | 30 | 9 |
| Arabic | 2 | 13 | 9 | 31 | 11 |
| Hindi | 5 | 11 | 8 | 27 | 9 |
| Turkish | 0 | 8 | 16 | 40 | 8 |
| Welsh | 5 | 10 | 9 | 28 | 10 |

**Table 1.** Paraphrase change (None, Word, Phrase, Weighted) and error across pivot languages.

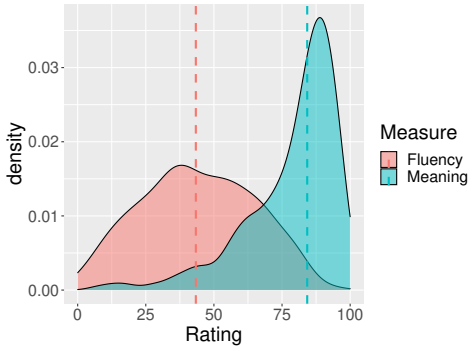


**Fig. 1.** Density plot for paraphrase ratings with indicated medians.

contained an acronym, chemical formula, time range, or malformed Unicode, e.g "Rh-abundant" for "Rh+," and adversely impacted meaning more than grammaticality/fluency. Forth was the replacement of a word with its antonym, e.g. "more mature" for "immature," and primarily impacted meaning. The other error types were approximately evenly represented and included pronoun insertion/deletion, replacement with a foreign word/phrase, insertion of a random word, and correct paraphrases that were misclassified. While some of these errors might be resolved with better or larger language models, we speculate that acronyms and chemical formulas may require a specialized approach.

## 4 Conclusion

Results from this study indicate that relatively high-quality paraphrases may be generated using a Transformer-based model fine-tuned with back-translated academic text. By leveraging a pre-trained Transformer like T5, researchers can construct a paraphrase model for a new domain in about a day, given available text in electronic format. An important limitation of these results is that only one domain was investigated, anatomy and physiology, raising the question of whether these results will generalize to other domains. Furthermore, while our results seem promising, we did not have a dataset to allow direct comparison to human performance, as is often the case in machine translation. Two important targets for future research are to replicate these findings in other domains and to conduct an evaluation directly comparing model-generated paraphrases with paraphrases generated by humans on the same source sentences.

## Acknowledgements

This material is based upon work supported by the National Science Foundation (1918751, 1934745) the Institute of Education Sciences (R305A190448).

## References


1. Barrón-Cedeño, A., Vila, M., Martí, M.A., Rosso, P.: Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics **39**(4), 917–947 (2013). https://doi.org/10.1162/COLI_a_00153
2. Betts, J.G., Desaix, P., Johnson, E., Johnson, J.E., Korol, O., Kruse, D., Poe, B., Wise, J.A., Womble, M., Young, K.A.: Anatomy and Physiology. OpenStax (2017)
3. Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., Monz, C.: Findings of the 2018 conference on machine translation (WMT18). In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 272–303. Association for Computational Linguistics, Belgium, Brussels (Oct 2018). https://doi.org/10.18653/v1/W18-6401
4. Botarleanu, R.M., Dascalu, M., Crossley, S.A., McNamara, D.S.: Sequence-to-sequence models for automated text simplification. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) Proceedings of the 21st International Conference on Artificial Intelligence in Education. pp. 31–36. Springer International Publishing, Cham (2020)
5. Burstein, J., Flor, M., Tetreault, J., Madnani, N., Holtzman, S.: Examining Linguistic Characteristics of Paraphrase in Test-Taker Summaries. ETS Research Report Series **2012**(2), i–46 (2012)
6. Fang, Z.: The language demands of science reading in middle school. International Journal of Science Education **28**(5), 491–520 (2006). https://doi.org/10.1080/09500690500339092
7. Federmann, C., Elachqar, O., Quirk, C.: Multilingual whispers: Generating paraphrases with translation. In: Proceedings of the 5th Workshop on Noisy User-Generated Text. pp. 17–26. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-5503
8. Fu, Y., Feng, Y., Cunningham, J.P.: Paraphrase generation with latent bag of words. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Proceedings of the Thirty-third Annual Conference on Neural Information Processing Systems. pp. 13623–13634 (2019)
9. Graham, Y., Baldwin, T., Moffat, A., Zobel, J.: Is machine translation getting better over time? In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 443–451. Association for Computational Linguistics, Gothenburg, Sweden (Apr 2014). https://doi.org/10.3115/v1/E14-1047
10. Hu, J.E., Khayrallah, H., Culkin, R., Xia, P., Chen, T., Post, M., Van Durme, B.: Improved lexically constrained decoding for translation and monolingual rewriting. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 839–850. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1090
11. Koleva, N., Horbach, A., Palmer, A., Ostermann, S., Pinkal, M.: Paraphrase detection for short answer scoring. In: Proceedings of the third workshop on NLP for computer-assisted language learning. pp. 59–73. LiU Electronic Press, Uppsala, Sweden (Nov 2014)
12. Krishna, K., Wieting, J., Iyyer, M.: Reformulating unsupervised style transfer as paraphrase generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 737–762. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.55

13. Li, Z., Jiang, X., Shang, L., Li, H.: Paraphrase generation with deep reinforcement learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3865–3878. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1421
14. Li, Z., Jiang, X., Shang, L., Liu, Q.: Decomposable neural paraphrase generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3403–3414. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1332
15. Litman, L., Robinson, J., Abberbock, T.: TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behavior Research Methods **49**(2), 433–442 (2017). https://doi.org/10.3758/s13428-016-0727-z
16. Nagy, W., Townsend, D.: Words as tools: Learning academic vocabulary as language acquisition. Reading Research Quarterly **47**(1), 91–108 (2012). https://doi.org/10.1002/RRQ.011
17. 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. 3173–3182. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1313
18. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
19. Shier, D., Butler, J., Lewis, R.: Hole's Human Anatomy & Physiology. McGraw-Hill Education, 15th edn. (2019)
20. Witteveen, S., Andrews, M.: Paraphrasing with large language models. In: Proceedings of the 3rd Workshop on Neural Generation and Translation. pp. 215–220. Association for Computational Linguistics, Hong Kong (Nov 2019). https://doi.org/10.18653/v1/D19-5623
21. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: 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. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6