Comparing Generic and Expert Models for Genre-Specific Text Simplification

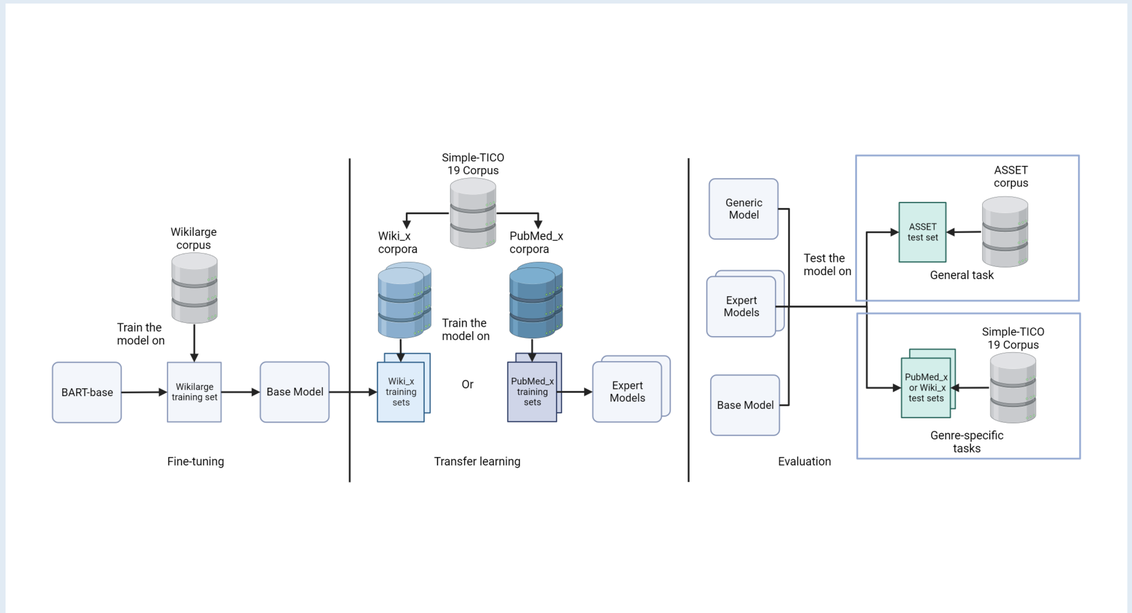
Zihao LI and Matthew Shardlow1 and Fernando Alva-Manchego
Manchester Metropolitan University
School of Computer Science and Informatics, Cardiff University
21443696@stu.mmu.ac.uk, m.shardlow@mmu.ac.uk, alvamanchegof@cardiff.ac.uk

INTRODUCTION

Most text simplification corpora focus on one text genre, such as Wikilarge (Zhang and Lapata, 2017) and Newsela (Xu et al., 2015). We investigate how text genre influences the performance of models for controlled text simplification. Regarding datasets from Wikipedia and PubMed as two different genres, we compare the performance of genre-specific models trained by transfer learning and prompt-only GPT-like large language models.

In this paper, we leveraged a newly published text simplification dataset Simple-TICO 19 (Shardlow and Alva-Manchego, 2022), designed a test scenario for controlled text simplification with different genres, proved the effects of transfer learning on the genre-specific datasets, compared the performance of generic and expert models in SARI score and BERTScore, and discussed the cost-effectiveness between expert models and generic models.

METHODS



Corpora:

- Wikilarge (Zhang and Lapata, 2017) is used to train the base model;
- ASSET (AlvaManchego et al., 2020a) is used as general test bench;
- Simple-TICO 19 (Shardlow and Alva-Manchego, 2022) is split into two genre-specific subsets based on the data source. For the two subsets, *Wikipedia* and *PubMed*, we create the training, validation and test sets in permutations. Each permutation is used to build and test corresponding genre-specific models, such as *Wikipedia0* and *PubMed0*.

Models:

- **Base model** is built on Wikilarge (Zhang and Lapata, 2017), following the MUSS (Martin et al., 2020) paradigm with customized settings.
- **Genre-specific models** are generated by fine-tinning the base model on training set of different subsets or permutations of Simple-TICO 19.
- We choose zero-shot GPT-3 and ChatGPT as our **General models.**

Task:

- General task : we test the 3 types of models on ASSET (Alva-Manchego et al., 2020) test set.

Genre-specific task: we test the 3 types of models on the 60 test sets from *Wikipedia0* to *PubMed29*.

RESULTS

Model		SARI	BERTScore
Base	BART-base	44.05	0.777
Generic	GPT-3	41.73	0.703
	ChatGPT	**46.42**	0.731
Expert	**Wikipedia0**	43.24	**0.835**
	PubMed0	43.67	0.812

Table 1: SARI and BERTScore on ASSET

Model		SARI	BERTScore
Base	BART-base	40.78	0.741
Generic	GPT-3	29.03	0.530
	ChatGPT	31.12	0.542
Expert	Average corresponding expert **Wikipedia** models	**44.30**	**0.756**
	Average **PubMed** models	42.75	0.741

Table 2: Average SARI and BERTScore on all ***Wikipediax***

Model		SARI	BERTScore
Base	BART-base	40.56	0.723
Generic	GPT-3	30.72	0.547
	ChatGPT	31.55	0.515
Expert	Average Corresponding expert **PubMed** models	**45.05**	**0.741**
	Average **Wikipedia** models	43.38	0.726

Table 3: Average SARI and BERTScore on all ***PubMedx***

.

- ChatGPT reaches the highest SARI score, while the expert model ***Wikipedia0*** obtains the highest BERTScore.
- BERTScore of generic models is lower than the base model.
- The performance gap between ChatGPT and the GPT-3 aligns with the model structure and scale.
- After transfer learning, expert models attain a marginally lower SARI scores but a higher BERTScore.

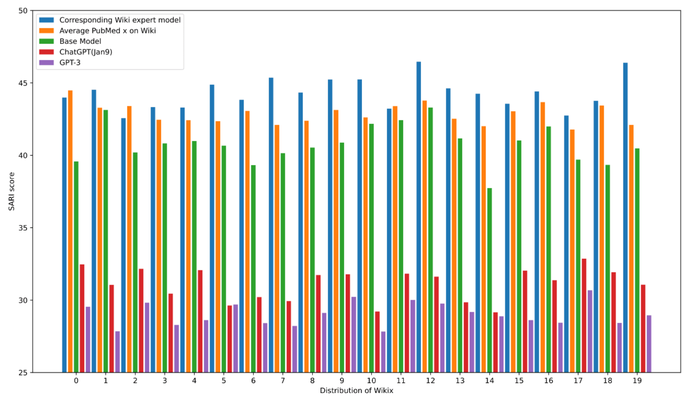
In both Table 2 and 3:

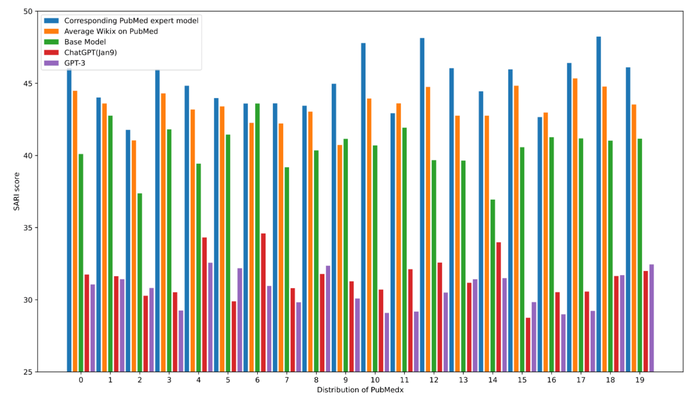
- The corresponding expert models show the highest average score in SARI and BERTScore.
- The overall performance gap between the two generic models is aligned to the gap in Table 1.
- The expert models have a higher SARI score, but the actual performance needs further exploration.
- Both kinds of expert models show improvement, caused by the sharing characteristics in the two subsets (both related to Covid-19).

Model	Simplicity	Meaning Preservation
Generic	3.55	3.86
Expert	3.46	4.17

Table 4 :Human evaluation score on test set of Wiki0 and PubMed0 (out of 5)

- For **simplicity** , the generic model (ChatGPT) obtains a similar, but marginally higher score than the expert models
- For **meaning preservation**, ChatGPT had worse performance than the expert models.





CONCLUSION

- The type of text (known as its genre) does affect the performance of text simplification models targeting general corpus;
- The zero-shot large language models are competitive but require tweaks to reach the same level of performance as the customized models;
- The smaller customized models may still hold their position as the best model.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoˆıt Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4668–4679, Online. Association for Computational Linguistics.
Louis Martin, Angela Fan, Eric de la Clergerie, An- ´ toine Bordes, and Benoˆıt Sagot. 2020. Multilingual unsupervised sentence simplification. CoRR, abs/2005.00352.
Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple TICO-19: A dataset for joint translation and simplification of COVID-19 texts. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 3093–3102, Marseille, France. European Language Resources Association.
Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 584– 594, Copenhagen, Denmark. Association for Computational Linguistics.
Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics, 3:283–297.