Judging Text Simplicity with Large Language Models

...
University
{...}@uni.edu

Abstract

TODO...

1 Introduction

TODO...
We introduce a panel of language models as a reference-free metric for text simplification. Unlike existing model-based metrics, our panel does not need fine-tuning. Thus, it understands texts from a variety of subjects and is not limited by the availability and domain of specialized training data. In this paper, we show that our metric has higher correlation with human judgement and matches or exceeds current metrics in all categories.

2 Panel of Language Models??

TODO...

2.1 Task description/Panel of Language Models

The task is as follows: Given some source sentence c, target simplification t, and if needed some set of references $r_1, ..., r_n$, compute a real-valued scalar that represents the quality of the simplification q. We do not use $r_1, ..., r_n$ as our metric is referenceless, although other metrics often do. Instead, we first construct a prompt $P = [I_{pre}; c; I_{mid}; t; I_{post}]$ by concatenation (;). While I_{mid} and I_{post} are primarily formatting and punctuation, I_{pre} can be any natural language instruction that elicits some score s from the model. We take these instructions and query multiple language models $J_1, ..., J_n$, k times each, to collect a set of scores $S = s_{1,1}, ..., s_{n,k}$, and aggregate the results into some final score $s_f = f(S)$. We experiment with multiple instructions, model sizes, and aggregation methods.

2.2 Instructions

As we are asking the model to evaluate a pair of simplifications through inference, the instructions, particularly I_{pre}, heavily influences the performance of the model. We focus on three aspects of design: The main task description, the examples, and the provided rationale.

First, we consider the main task description. While the primary component of this is a straightforward description of the text simplification evaluation task, we find that using a persona - "You are an expert professional linguist" - results in slightly more accurate readings. We also find that including detailed notes on what constitutes a "good simplification" and potential mistakes generally improves performance. While we could optimize our prompt further, for the sake of fair comparison we instead keep it similar to existing literature and the instructions they provide to human editors. We do make some minor edits based on the exact dataset we are evaluating on, particularly with respect to the aspects that the model is supposed to emphasize.

The second aspect is the examples we provide to the model. The vast majority of existing literature agree that few-shot learning generally improves the performance of model outputs. We therefore provide examples of different simplifications and their corresponding scores. To do this, we use examples provided in the human annotation instructions from the Newsela-LikeRT dataset. These examples are identical for all datasets, with only the score changing. In some cases, this does result in a less diverse distribution of examples, but we find that this is somewhat mitigated by providing the model with similar instructions to human raters.

The last design aspect we focus on is the rationale. In order to provide an accurate score, we ask the model to first note down what each simplification does well and what it does poorly, and then reason through the rating in a manner similar to chain-of-thought. We provide sample rationale for each of the examples given to the model. While the examples were from prior work, we find that they are not directly usable due to formatting. They also do not explain as much as would be necessary for the model to make a strong decision. We therefore elaborate on each of the ratings, noting down problems such as grammar mistakes, inaccurate connotations, and the use of difficult words. These explanations are also written such that the details are first and build towards the final score; the model is therefore encouraged to start with raw observations and build towards a final evaluation.

Template	SimpEval2022 τ_{all}
{source} can be simplified as:	0
a simpler version of {source} is:	0
a simpler and semantically identical way to say {source} is:	0

Table 1: Comparison of various templates, GPT-2 XL

2.3 Model Selection

Because we are using language models in inference, model choice is much more open than previous work. As we do not need access to weights, we could use the most cutting-edge models available, such as GPT-4o or Claude 3. However, metrics should be relatively accessible, and we therefore limit our selection to smaller open-weight models. Recent studies have also shown that multiple small models can perform better than larger ones in an evaluation context, while being cheaper. With these criterion in mind, we ultimately settle on Ministral 8B (Mistral AI Team, 2024), Llama 3.1 8B (Dubey et al., 2024), and Gemma 2 9B (Team, 2024) as our panel of judges. While we do experiment with larger models and find marginal improvements, we find that they are not worth the performance cost.

2.4 Aggregation

In our experiments, we generally run each of the 3 judging language models twice on the same point, for a total of 6 scores from 1 through 5. These generations were done with a temperature of 0.1 to decrease randomness while leaving room for different reasoning each time. We experiment with both the mean and median of the scores, and find that the mean generally performs better; the median is in 0.5-point intervals and does not differentiate between scores well. Thus, we have $n = 3$, $k = 2$, and $f(S) = \Sigma S/(n \cdot k)$.

Model	SimpEval2022 τ_{all}
GPT-2	0.477
GPT-2 Medium	0.520
GPT-2 Large	0.508
GPT-2 XL	0.471

Table 2: Comparison of GPT-2 Models on SimpEval2022. Final template used.

3 Experiments

We evaluate our panel on Newsela-LikeRT. Other common datasets, such as SimpEval2022 and WikiDA, were publicly released prior to the training cutoffs of the models and the results could be contaminated; on the other hand, Newsela is private and likely not included in training data.

3.1 Newsela-LikeRT

Metric	Fluency	Meaning	Simplicity
FKGL	0.193	0.306	-0.051
BLEU	0.332	0.261	0.118
SARI	0.234	0.124	0.094
BERTScore	0.384	0.274	0.215
LENS$_{k=3}$	**0.624**	0.428	**0.359**
Templated Perplexity*	0.389	**0.589**	0.120

Table 3: Correlation with human scores on Newsela

The Newsela dataset is composed of around 440 simplification pairs. Each pair has three separate scores: fluency/grammar, meaning, and simplicity. Multiple human raters annotated each pair along each of the three dimensions, with the average in each dimension used as the final rating. This dataset is additionally of a relatively high quality, as they are sourced from news articles.

3.2 Results

We report our method's correlations with human scores in each of the dimensions HERE. Depending on the target dimension, we provide slightly different instructions to the models: The instructions mention word difficulty when measuring simplicity and run-on sentences when measuring fluency. This targeted approach focuses the model on the aspects that are most important. Our panel shows stronger correlation with human judgment than existing metrics on simplicity, which is generally understood to be the most important of the three. Our panel also performs well on meaning, and is competitive with current best metrics on fluency.

3.3 Model selection

We find that while choice of model has a generally large impact on correlation, model size does not. The table below shows the correlation of various models, ordered by size; while variation is large, there is no clear relation between size and performance. One potential caveat, however, is that our experiments used more samples for smaller models ($k = 3$ or more) compared to larger ones ($k = 2$) due to computational constraints; it is possible that this made larger models more susceptible to noise or that the more discrete nature of the score - which by necessity can only be in increments of 0.167 as opposed to the 0.0833 of smaller models - made the linear fit less accurate.

3.4 Instructions

In addition to this, we experiment with zero-shot, one-shot, and few-shot instructions. We find that zero-shot often results in outputs that do not follow the instructions, showing that the human annotation directions do not adapt well to LLM-style instructions and may need to be rewritten. On the other hand, we saw minimal differences between 1-shot and few-shot outputs. The lack of difference between providing a single high-scoring simplification and providing a range of different scores shows that the extra samples do not provide information that the model does not already know. Instead, it may be possible to gain increased performance by designing samples specifically for the aspects that the model often fails to consider.

4 Related Work

Existing text simplification metrics broadly fall into two categories: model-based and model-based metrics. Traditional metrics were non-model based, and often depended on word or n-gram occurrence. Examples of this include SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002). Even earlier approaches included FKGL (Flesch, 1948), which is still commonly used to this day. This computes text simplicity based solely on syllable and word counts in sentences. However, these metrics are not designed for the text simplification task, having been adapted from other fields; recent work has shown that this has its limitations (Sulem et al., 2018; Tanprasert and Kauchak, 2021).

More recently, work has been done on using language models to measure text simplicity. While initially designed for semantic similarity, BERTScore (Zhang et al., 2019) has been used to measure some aspects of text simplification. More recent work, such as LENS (Maddela et al., 2023), REFeREE (Huang and Kochmar, 2024), and SLE (Cripwell et al., 2023), have trained smaller models (such as RoBERTa) to predict scores. While they perform relatively well, they are also limited by the need to collect datasets with human ratings. SLE circumvents this by using a combination of Newsela data - already labeled by difficulty - and interpolation with FKGL, but this dataset is also constrained by the generalization of the former and performance of the latter.

Our work builds on language model inference techniques. We base our reasoning on chain-of-thought (Wei et al., 2022), adopted to a classification task, and use few-shot learning (Brown, 2020); in particular, one-shot learning greatly improves performance. Lastly, we use models as evaluators, which have previously shown performance competitive with, and in some cases superior to, human judgement (Bohnet et al., 2022). Additionally, pretrained models are able to generalize better than their fine-tuned counterparts (Huang et al., 2024). However, one main drawback is that these models tend to prefer their own outputs (Panickssery et al., 2024). To counteract this, we use juries as proposed by Verga et. al. to improve performance while decreasing hardware requirements and costs (Verga et al., 2024).

5 Conclusion

TODO Summarize...

5.1 Future Work

TODO Summarize...

6 Acknowledgements

TODO...

References

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity level estimate (sle): A learned reference-less metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Hui Huang, Yingqi Qu, Xingyuan Bu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4.

Yichen Huang and Ekaterina Kochmar. 2024. REFeREE: A REference-FREE model-based metric for text simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13740–13753, Torino, Italia. ELRA and ICCL.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Mistral AI Team. 2024. Un ministral, des ministraux.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix?

TODO...