ANU at MLSP-2024: Prompt-based Lexical Simplification for English and Sinhala

Sandaru Seneviratne[1] **and Hanna Suominen**[1,2]
[1]The Australian National University (ANU) / Canberra, ACT, Australia
[2]University of Turku / Turku, Finland
Firstname.Lastname@anu.edu.au

Abstract

Lexical simplification, the process of simplifying complex content in text without any modifications to the syntactical structure of text, plays a crucial role in enhancing comprehension and accessibility. This paper presents an approach to lexical simplification that relies on the capabilities of generative Artificial Intelligence (AI) models to predict the complexity of words and substitute complex words with simpler alternatives. Early lexical simplification methods predominantly relied on rule-based approaches, transitioning gradually to machine learning and deep learning techniques, leveraging contextual embeddings from large language models. However, the the emergence of generative AI models revolutionized the landscape of natural language processing, including lexical simplification. In this study, we proposed a straightforward yet effective method that employs generative AI models for both predicting lexical complexity and generating appropriate substitutions. To predict lexical complexity, we adopted three distinct types of prompt templates, while for lexical substitution, we employed three prompt templates alongside an ensemble approach. Extending our experimentation to include both English and Sinhala data, our approach demonstrated comparable performance across both languages, with particular strengths in lexical substitution.

1 Introduction

Lexical simplification, an essential component in making complex text more understandable, involves replacing complex words with simpler alternatives while preserving the meaning and syntax (Bott and Saggion, 2011; Seneviratne et al., 2022b). This task is specifically valuable for people with limited knowledge in certain languages or domains or for people with low literacy skills (Gooding and Kochmar, 2019). Lexical simplification can be composed as a cascade of complex word identification and lexical substitution. Addressing both these tasks is vital for improved language understandability.

Complex word identification task is the first step in lexical simplification (Shardlow, 2014). This task can be formulated as identifying the complex words in text or as predicting the level of lexical complexity for each word. Various techniques have been employed for this task, ranging from rule-based (Devlin, 1998; Carroll et al., 1999) through threshold-based (Zeng et al., 2005) to lexicon-based approaches (Miller, 1995). Following these methods, researchers have also explored feature-based machine learning methods (Wróbel, 2016; Malmasi et al., 2016) that also incorporate word embedding models and more sophisticated approaches like deep learning models such as long short-term memory (LSTM) networks, modelling the problem as a sequence labelling task (Gooding and Kochmar, 2019). Recently, contextual embedding models like Bidirectional Encoder Representation from Transformers (BERT) have gained attention for complex word identification due to their ability to capture nuanced contextual information (Qiang et al., 2021; Seneviratne, 2024).

Similar to complex word identification, lexical substitution is an important sub-task for lexical simplification. Early methods relied on lexical resources to generate simpler, suitable, relevant substitutes for complex or target words (Biran et al., 2011; Pavlick and Callison-Burch, 2016). This evolved with the introduction of word embedding models like Word2Vec (Mikolov et al., 2013), Global Vectors for Word Representation (GloVe) (Pennington et al., 2014), and Embedding from Language Models (ELMo) (Peters et al., 2018). More recently, contextual embedding models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019) have become popular for lexical substitution, sometimes combined with lexical resources for improved performance (Seneviratne et al., 2022a).

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 599–604
June 20, 2024

Even though many natural language processing tasks have relied on more complex or sophisticated methods based on deep learning models and contextual embeddings, with the emergence of generative Artificial Intelligence (AI), most of the methods have shifted to exploring simpler approaches based on prompt engineering (Aumiller and Gertz, 2022). Prompt engineering presents straightforward and effective approaches for a wide range of tasks, including for lexical complexity prediction and substitution. In this study, we leveraged prompt engineering for both tasks, focusing on improving the accuracy and efficiency.

2 Experiments

2.1 Datasets

Our experiments and evaluations used the English and Sinhala language datasets provided by the MLSP-2024 shared task (Shardlow et al., 2024b; North et al., 2024).

Lexical Complexity Prediction. The trial subset of both English and Sinhala lexical complexity prediction datasets comprised 30 sentences each, and consisted of samples with the context, target word, and their respective lexical complexities. The data from the trial subset of the data was used for one-shot and few-shot prompt template creation. The test subset of the dataset consisted of 600 samples each for both English and Sinhala, which was used for the evaluation of the proposed prompt-based methods.

For lexical complexity prediction, since the dataset had samples where the same sentence had been associated with different target words, we first grouped the sentences together and obtained lexical complexities for each target word in a sentence. This facilitated a comparative perspective on the complexity levels of the target words relative to one another. Moreover, this enhanced the information included in the prompt template allowing a better understanding of the distinctions and variations in lexical complexity.

Lexical Substitution. For the lexical substitution task, we employed datasets in both English and Sinhala, each consisting of context sentences with words requiring simplification, along with sets of alternative words. Similar to the complexity prediction task, the trial subset of the both the datasets consisted of 30 samples, which were used for prompt template creation. The test subset of the data, that was used for evaluation, comprised 570 samples for English and 600 samples for Sinhala, respectively.

2.2 Methods

We relied on prompt-based methods for both lexical complexity prediction and lexical simplification through substitution generation. We relied on Generative Pre-trained Transformer– GPT3.5-turbo-instruct model with a *temperature* of 0.5 and *top_p* value of 1 for our experiments. This specific model has a context window of 4, 096 tokens.

Lexical Complexity Prediction. For lexical complexity prediction, we explored the following three distinct prompt templates to study how varying levels of additional information can affect the final prediction: zero-shot, one-shot, and few-shot. Each of these widely recognized templates provided unique information as to how additional contextual information influences lexical complexity prediction. Namely, the zero-shot template, which only used the given sample input to determine lexical complexity of the target word, served us as a baseline to compare with the other two prompt-template methods. For the one-shot approach, we selected a single random sample from the processed trial dataset. Conversely, the few-shot approach involved incorporating three examples from the trial dataset into the prompt. Since we processed the dataset to consolidate the same contexts and their target words, the samples included in the prompt consisted of context sentence along with their target words and the lexical complexity values.

Context: {context}
Question: Given the above context, give the lexical complexity for each word as a value between 0 and 1. The words are {words}
Lexical complexity:

Table 1: Zero-shot prompt template used for lexical complexity prediction. For one-shot and few-shot prompt templates, examples were incorporated.

Lexical Substitution for Simplification. Similar to the lexical complexity prediction task, we relied on three prompt templates for the initial generation of simpler, relevant, and suitable substitutes for a given target work. While our zero-shot approach only included the given context and the target word for substitution generation, we incorporated in the one-shot and few-shot prompt templates one and three samples from the trial dataset, respectively. In the latter two approaches, our prompt included the given context sentence, target word, and their possible substitutes for the generation process. In each prompt template, we asked the model to provide ten simpler substitutes for the target word.

Context: {context} **Question**: Given the above context, list ten alternative words for {word} that are easier to understand. **Alternative susbtitutes**:

Table 2: Zero-shot prompt template used for lexical substitution. For one-shot and few-shot prompt templates, examples were incorporated.

We further filtered the results obtained from the three prompt templates. To combine the results from the prompt templates, we followed (Aumiller and Gertz, 2022), where the authors computed a combination score V (Eq. 1) for each distinct prediction, where $\text{rank}_{Sp(s)}$ is the ranked position of a possible substitute s for a given prompt p.

$$V(s) = \sum_{p=1}^{3} \max(5.5 - 0.5 \times \text{rank}_{Sp(s)}, 0). \quad (1)$$

2.3 Evaluation metrics

We based the evaluation of the proposed methods on the metrics used in the MLSP-2024 shared task (Shardlow et al., 2024a). For lexical complexity prediction, Pearson's R, Spearman's Rank, *Mean Absolute Error* (MAE), and *Mean Squared Error* (MSE) were used. For the lexical substitution task, we relied on Accuracy@K ($K \in \{1, 2, 3\}$), *Potential@K* ($K \in \{1, 3\}$), and *Mean Average Precision@K* (MAP@K) ($K \in \{3, 5\}$).

3 Results

The results of the prompt-based lexical complexity prediction methods did not reach the performance levels of the top submissions in the lexical complexity prediction task (Table 3). While the best submission achieved Person's R of 0.8497, the best system from our experiments — the zero-shot approach — had Person's R of 0.3358. Among our prompt-template-based methods for Sinhala, the one-shot approach yielded the most promising results. However, its Pearson's R of 0.0534 was placed fifth among the submissions for Sinhala.

In lexical simplification for English, our proposed few-shot approach showed strong performance, achieving comparable results with respect to the best submissions for the task across all metrics (Table 4). The proposed method gave the Accuracy@1 score of 0.5105, while the best submission gave 0.5245. However, for Sinhala, our submission (which was the ensemble approach) did not show satisfactory performance.

4 Discussion

In this paper, we have explored the applicability of prompt-templates for both lexical complexity prediction and lexical substitution for simplification in English and Sinhala. Our investigation primarily focused on three prompting methodologies: zero-shot, one-shot, and few-shot. The experiments demonstrated diverse performance levels across the two tasks and languages under consideration.

The most effective approach of our experiments for predicting complexity in English relied on the zero-shot method, while for Sinhala, the one-shot approach gave the best results. This difference may stem from differences in language data availability and the complexity of each language and task. Compared to Sinhala, English has more language data available, providing the model with a more extensive information base. This could be a reason why for English the zero-shot approach performed better, as the model could leverage enough contextual information. However, Sinhala, being less extensively studied, likely has fewer linguistic resources and data available for training. Therefore, the one-shot approach, which provides additional context, may be better suited to capture the patterns and dependencies in the language.

Considering the performance of the prompt-based methods for complexity prediction in Sinhala, the few-shot approach did not perform as well as the one-shot approach, even though more additional information was provided. This discrepancy could be attributed to the quality of the samples included in the prompt template. If the chosen examples fail to adequately represent the lexical features and patterns of the language, it may lead to a degradation in performance, resulting in poorer results compared to the one-shot approach.

The results from the lexical substitution for simplification indicated varied performance. In English, out of our experiments, the few-shot approach gave the best results, closely followed by the ensemble approach, which combined results from all three prompt templates. This suggests that the few-shot approach provided good example instances that helped in capturing the lexical intricacies of the language. Therefore, while the ensemble approach gave comparable performance,

Team Name	Run ID	Pearson's R	Spearman's Rank	Mean Absolute Error	Mean Squared Error
English					
GMU	1	0.8497	0.7984	0.1137	0.0175
TMU-HIT	2	0.8198	0.7552	0.1108	0.0178
SDJZUandUU	3	0.8123	0.7754	0.1071	0.0175
RETUYT-INCO	2	0.5502	0.4923	0.1561	0.0328
ANU	1	0.3358	0.3591	0.3484	0.1478
GMU	A	0.3118	0.3183	0.1389	0.0346
CocoNut	1	0.1972	0.2160	0.4150	0.2263
Sinhala					
TMU-HIT	A	0.3081	0.3343	0.1666	0.0422
TMU-HIT	A	0.2482	0.3261	0.2126	0.0661
RETUYT-INCO	A	0.1344	0.1094	0.3355	0.1340
GMU	1	0.1246	0.1303	0.1018	0.0168
ANU	2	0.0534	0.0866	0.1741	0.0539
SCaLAR	A	0.0450	0.0279	0.1576	0.0321
Archaeology	2	0.0437	0.0298	0.1239	0.0236
GMU	A	0.0263	0.0284	0.1066	0.0180

Table 3: Results of the experimented approaches on the test subsets of the English and Sinhala datasets provided at the MLSP-2024 shared task for lexical complexity prediction.

Team Name	Run ID	Accuracy@1	Accuracy@2	Accuracy@3	Potential@1	Potential@3	MAP@3	MAP@5
English								
TMU-HIT	1, A1	0.5245	0.6807	0.7456	0.7982	0.9035	0.5762	0.4142
GMU	1, A1	0.5157	0.635	0.6894	0.7491	0.8754	0.513	0.3691
ANU	3	0.5105	0.6175	0.6649	0.7684	0.8824	0.5324	0.3744
ANU	1	0.4684	0.5929	0.6561	0.735	0.8684	0.5069	0.3652
ISEP_Presidency	1	0.4684	0.607	0.6754	0.7649	0.8859	0.5351	0.3877
ANU	2	0.4631	0.5807	0.6421	0.7228	0.8614	0.4978	0.3524
TMU-HIT	2	0.4438	0.6298	0.7456	0.7333	0.9035	0.5595	0.4042
RETUYT-INCO	3	0.3789	0.5105	0.5701	0.5947	0.7824	0.3832	0.2634
RETUYT-INCO	2	0.3438	0.4701	0.5526	0.5789	0.7666	0.3718	0.2542
Sinhala								
GMU	A1	0.2284	0.2829	0.3163	0.311	0.4165	0.1387	0.0894
GMU	1	0.2283	0.2866	0.32	0.3116	0.4183	0.14	0.0902
TMU-HIT	A2	0.2214	0.3286	0.3585	0.3198	0.4903	0.1673	0.108
TMU-HIT	A1	0.2144	0.304	0.3585	0.3444	0.4903	0.1709	0.1101
GMU	A2	0.13	0.2372	0.3057	0.195	0.3848	0.1147	0.0759
TMU-HIT	A3	0.1195	0.2759	0.3585	0.2249	0.4903	0.1469	0.0957
Archaeology	1	0.0466	0.0633	0.0783	0.0666	0.1383	0.0359	0.0242
ANU	1	0.0133	0.015	0.0166	0.0133	0.0183	0.0074	0.0045
RETUYT-INCO	A1	0.0017	0.0017	0.0017	0.0123	0.0123	0.0041	0.0024
RETUYT-INCO	A2	0	0	0	0.0087	0.0105	0.0032	0.0019

Table 4: Results of the experimented approaches on the test subsets of the English and Sinhala datasets provided at the MLSP-2024 shared task for lexical susbtitution for simplification.

it did not filter the best predictions as effectively as the few-shot method. However, for Sinhala lexical substitution, we only employed the ensemble approach. Unfortunately, the results indicated subpar performance. This suggests that the ensemble approach did not effectively capture the lexical patterns, dependencies of Sinhala language, that resulted in unsatisfactory outcomes.

The findings indicate the importance of investigating the influence of the factors such as data availability, language complexity, and sample quality on the outcomes of lexical simplification tasks. Additionally, refining prompt tuning methods could enhance the effectiveness and outcomes.

5 Conclusion

In this work, we have used prompt-based methods for both lexical complexity prediction and lexical substitution for simplification, focusing on exploring the applicability of generative AI methods. The results from the different methods indicate varied performance levels across the two tasks and languages, giving evidence of challenges related to data availability, representations, quality of the samples, language complexity, and adaptability of the models for the lexical simplification task. This encourages further investigations that could potentially improve the performance differences.

6 Limitations

The experiments were conducted using GPT-based models, which posed challenges primarily due to their significant resource requirements (Aumiller and Gertz, 2022). Thus, to facilitate these experiments, we accessed the GPT model through an Application Programming Interface (API), which costed approximately $8 for all experiments. Furthermore, the utilization of these models raises ethical concerns surrounding data privacy and transparency limitations. Additionally, our findings highlighted variations in results based on the prompt template, the examples included in the prompts, and the parameters used, highlighting the need for further investigation on the usability of these models for NLP tasks.

Acknowledgments

This research was funded by and has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalized health technologies and solutions in collaboration with patients, clinicians and healthcare providers.

References

Dennis Aumiller and Michael Gertz. 2022. UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26, Portland, Oregon. Association for Computational Linguistics.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*.

Sian Gooding and Ekaterina Kochmar. 2019. Recursive context-aware lexical simplification. 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)*, pages 4853–4863, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. LTG at SemEval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 996–1000, San Diego, California. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. Multils: A multi-task lexical simplification framework. *arXiv preprint arXiv:2402.14972*.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021. LSBert: Lexical simplification based on BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Ishanee Sandaru Seneviratne. 2024. *Text Simplification Using Natural Language Processing and Machine Learning for Better Language Understandability*. Ph.D. thesis, The Australian National University.

Sandaru Seneviratne, Elena Daskalaki, Artem Lenskiy, and Hanna Suominen. 2022a. CILex: An investigation of context information for lexical substitution methods. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4124–4135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sandaru Seneviratne, Elena Daskalaki, and Hanna Suominen. 2022b. CILS at TSAR-2022 shared task: Investigating the applicability of lexical substitution methods for lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 207–212, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024a. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024b. An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*.

Krzysztof Wróbel. 2016. PLUJAGH at SemEval-2016 task 11: Simple system for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 953–957, San Diego, California. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *Biological and Medical Data Analysis: 6th International Symposium, ISBMDA 2005, Aveiro, Portugal, November 10-11, 2005. Proceedings 6*, pages 184–192. Springer.