A Comparative Performance Evaluation of BERT, RoBERTa and LSTM based RNN for Complex Word Identification

Shreyas Ramachandran
School of Computer Science and Informatics, Cardiff University
ramachandrans3@cardiff.ac.uk

Abstract

Lexical simplification enhances text accessibility by identifying and replacing complex words with simpler alternatives. A critical sub-task in this process is Complex Word Identification (CWI), which determines word complexity in context. This study compares two distinct machine learning approaches: BERT, a transformer-based model, and an LSTM-based recurrent neural network—for predicting lexical complexity using the CompLex dataset, made available in the MSLP 2024 shared task resources (Shardlow et al., 2024). This project also explores training BERT for substitute word generation using the BenchLS (Paetzold and Specia, 2016) dataset and the trained model is evaluated on the TSAR-2022 dataset (Štajner et al., 2022). Our evaluation reveals that BERT achieves an accuracy of 0.79 for CWI, outperforming RoBERTa (0.44) and the LSTM-RNN (0.57). For Substitution Generation, the BERT model generates substitutes with an accuracy score of 0.59 and a semantic similarity of 0.65 to the gold standard. These results underscore the efficacy of Transformer models in Lexical Simplification and highlight areas for further improvement.

1. Introduction

Lexical simplification (LS) improves text readability by substituting complex words with simpler synonyms, benefiting diverse populations such as language learners and individuals with cognitive impairments. A foundational component of LS is Complex Word Identification (CWI), which assesses word complexity within context. Traditional CWI approaches, like those in SemEval-2016 Task 11 (Paetzold and Specia, 2016), treated complexity as binary, oversimplifying its nature. Recent datasets like CompLex (Shardlow et al., 2020) introduce continuous complexity scores, reflecting a spectrum of difficulty. This project aims to compare two models with distinct architectures—BERT (Devlin et al., 2019), a transformer-based model, and an LSTM-based model (Hochreiter and Schmidhuber, 1997), a recurrent neural network—for CWI using CompLex. For CWI, we assess the performance of BERT, RoBERTa, and an LSTM-based RNN model using the CompLex dataset, which provides annotated sentences with complex words. For Substitution Generation, we fine-tune a BERT model on the BenchLS dataset, which contains sentences with target words and their corresponding simpler substitutes. Our objectives are to determine which model performs best for CWI and to showcase the capability of BERT in generating contextually appropriate substitutes, contributing to the development of more effective Lexical Simplification systems.

2. Related Work

Early CWI efforts relied on heuristics like syllable counts (McLaughlin, 1969) or frequency thresholds (Shardlow, 2013). The SemEval-2016 CWI task (Paetzold and Specia, 2016) formalized binary classification, followed by CWI 2018 (Yimam et al., 2018), which added probabilistic scores. However, these approaches lacked granularity. CompLex (Shardlow et al., 2020) shifts to continuous prediction using a 5-point Likert scale across Bible, Europarl, and biomedical texts, revealing complexity variations by genre and context. Text simplification research, including SARI (Xu et al., 2016) and EASSE (Alva-Manchego et al., 2019), evaluates LS holistically but often overlooks individual word complexity.

This work bridges this gap by focusing on CWI with diverse model architectures, building on CompLex's baseline of linear regression with GloVe embeddings. This project builds upon existing research by providing a direct comparison between state-of-the-art Transformer models (BERT and RoBERTa) and a traditional LSTM-based RNN model for CWI, and by evaluating a fine-tuned BERT model for Substitution Generation on standard datasets.

3. CompLex Dataset

The CompLex dataset (Shardlow et al., 2024) contains 10,000 English sentences with binary annotations indicating complex words, designed for CWI tasks. Each sentence is sourced from diverse domains (e.g.: Wikipedia, news), with annotations reflecting perceived complexity for non-native speakers. This study performed a detailed analysis of this dataset to note the following points:

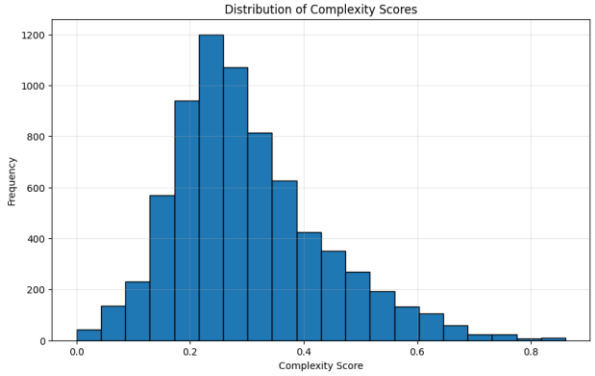

Fig 3.1: Distribution of Complexity scores in CompLex

- Sentences range from 5 to 50 words, with an average of 20 words, where approximately 15% of words are annotated as complex, with nouns and adjectives being the most frequent as shown in the word cloud below.

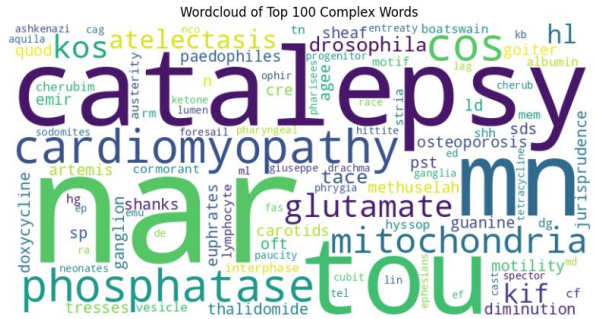

Fig 3.2: Top 100 complex words in CompLex

- Complex words often have higher syllable counts (average: 3.2 syllables) and lower frequency in general English corpora. Longer words are usually more complex.
- Sentences cover multiple topics, ensuring models must handle varied contexts, as seen in a word cloud of high-frequency terms.

These analyses confirm the dataset's suitability for training robust CWI models, capturing
diverse linguistic and contextual challenges

4. Methodology.

4.1 Data Collection

- For CWI, we used the CompLex dataset, comprising 9,476 sentences from Bible, Europarl, and biomedical domains. Each sentence contains a target word (nouns or multi-word expressions) annotated by approximately 7 crowd workers on a 5-point Likert scale (1=Very Easy, 5=Very Difficult), normalized to [0,1]. The dataset's diversity tests model robustness across genres.
- For Substitution Generation, we employed the BenchLS dataset, containing 1,000 sentences with target words and their simpler substitutes. The substitution model was then tested on the TSAR-2022 test dataset, containing 370 sentences with target words and a list of approximately 12 substitutes .

4.2 Preprocessing

For the CompLex dataset, we tokenized the sentences and aligned the annotations with the tokens. For the BenchLS dataset, we tokenized the sentences and prepared them for masked language modeling by masking the target words.

4.3 Model Implementation

For CWI, : We fine-tuned pre-trained BERT-base-uncased and RoBERTa-large models for token-level sequence classification. Each model takes tokenized sentences as input, outputting a binary classification (complex or not) per token. We used a learning rate of 2e-5, batch size of 16, and trained for 3 epochs with AdamW optimization. The LSTM-RNN model was trained from scratch using GloVe 300D (Pennington et al., 2014) embeddings

(100-dimensional), a bidirectional LSTM layer with 128 hidden units, and 2 layers with a dropout of 0.3, followed by a classification head.
For Substitution Generation, we fine-tuned a BERT model for masked language modeling, predicting possible substitutes for masked target words, and ranked them using semantic similarity computed with the 'all-MiniLM-L6-v2' model. Training used a learning rate of 2e-5, batch size of 8, and 5 epochs.

4.4 Evaluation Metrics

For CWI, we used accuracy, MSE, R^2, pearson correlation (r), and spearman correlation (r_s). For Substitution Generation, we computed Accuracy scores (comparing generated substitutes to gold standards in TSAR-2022) and semantic similarity (cosine similarity of embeddings).

5. Results and Evaluation

For CWI, model performance on the CompLex dataset is shown in Table 1. r = pearson correlation, r_s = spearman correlation.

Model	Accuracy	MSE	R^2	r	r_s
BERT	0.7942	0.0104	0.5349	0.7323	0.7141
RoBERTa	0.4437	0.0229	0.1515	0.1282	0.0969
LSTM-RNN	0.5705	0.0169	0.2414	0.5205	0.4903

Table 1: Evaluation Report for CWI

From the report, we can clearly identify that BERT has the best performance with an accuracy of approx. 80%.

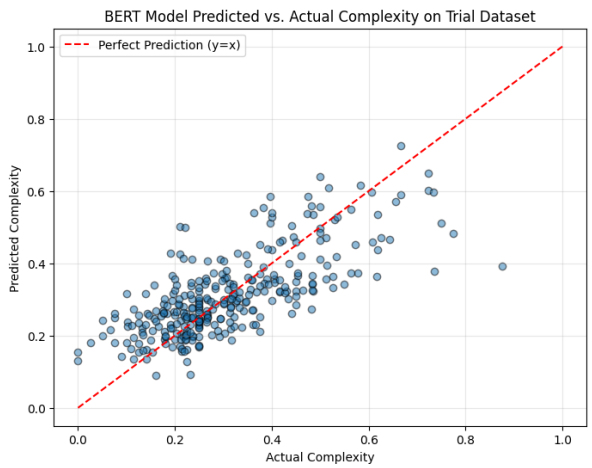

Fig 5.1: Predicted vs Actual Complexity for BERT

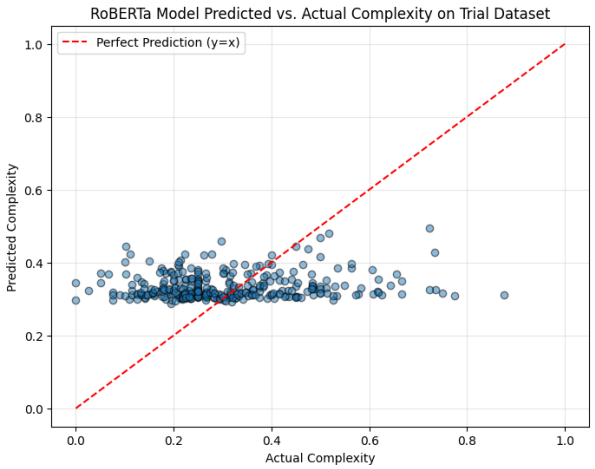

Fig 5.2: Predicted vs Actual Complexity for RoBERTa

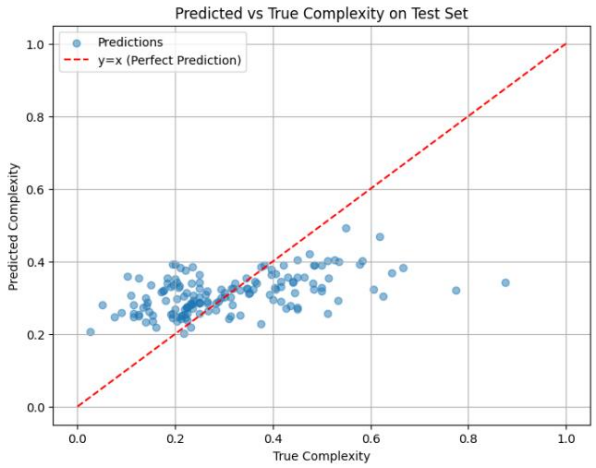

Fig 5.3: Predicted vs Actual Complexity for LSTM-RNN

BERT outperformed the LSTM-RNN in CWI, likely due to its contextual understanding, while RoBERTa's slightly lower performance may reflect overfitting on the dataset, or due to lower epochs in training.

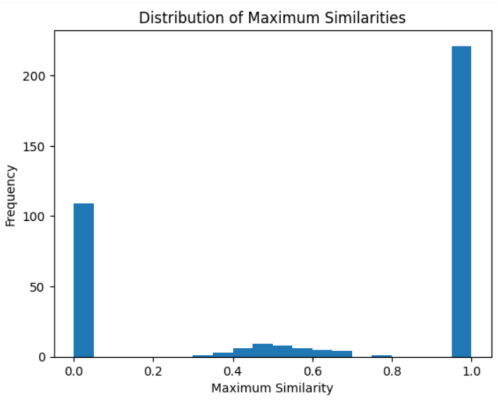

Fig 5.4: Similarity Distribution for Substitution Generation

For Substitution Generation, the BERT model was evaluated on the BenchLS dataset using BLEU

score and semantic similarity. The average accuracy score was 0.5925, and the average semantic similarity to gold substitutes was 0.6524.

6. Error Analysis

In CWI, models struggled with ambiguous words. For instance, the LSTM-RNN misclassified "bank" as complex in a financial context but correctly identified it in a riverbank context. For the target word "*sectarian*" in the sentence *"Lebanon is sharply split along sectarian lines, with 18 religious sects."*, BERT classified the complexity as 0.3601, while RoBERTa predicted 0.3189 and LSTM-RNN predicted 0.3488. The actual complexity of the word was 0.3913, making BERT's prediction the closest. RoBERTa severely underperformed on the test set for the same training settings, indicating possible overfitting to the noise in the training data. These errors highlight challenges in capturing nuanced semantics

For Substitution Generation, the model often provided inappropriate substitutes, and in many cases was unable to provide any substitutes, returning [UNK] instead when we provide an example input that is not part of the BenchLS dataset. [UNK] character is returned by the model whenever it is not able to generate any substitutes. This can be attributed to the fact that BenchLS only has 929 examples for training. This small dataset size is insufficient to train the model to generate substitutions for all the words in various contexts.

7. Ethical Implications

CompLex's crowd-sourced annotations may reflect biases from annotators' backgrounds (e.g., familiarity with religious texts skewing Bible scores). Pipeline design errors, such as misjudging complexity, could oversimplify texts for advanced readers or present advanced text as substitutes for novices, risking loss of meaning or patronization. LS systems must prioritize inclusivity, adapting to user needs and avoiding stereotypes in substitutions. Privacy concerns are minimal, but fairness across domains and demographics requires the ongoing scrutiny. Diverse datasets, rigorous evaluation, and human oversight to ensure ethical and accurate simplification are recommended.

8. Conclusion

This project compared BERT, RoBERTa, and an LSTM-RNN model for CWI, with BERT achieving the highest accuracy (0.79). showcasing very good contextual understanding in CWI, a vital task in Lexical Simplification. The fine-tuned BERT model for Substitution Generation showed promising results (Accuracy: 0.59, Similarity: 0.65). The CompLex dataset analysis and substitute generation demonstration underscore the models' capabilities and limitations.

8.1 Future Work

Future research could explore:

- **Hybrid Models:** Combining Transformer and RNN architectures to balance contextual understanding and sequential processing.
- **Contextual Features:** Incorporating more nuanced semantic features to improve CWI and substitute selection.
- **Multilingual Simplification:** Extending the use of models to non-English languages using multilingual Transformers.
- **Larger and Diverse Datasets:** BenchLS only had 929 examples for training. We need to explore using larger datasets with multiple sources to provide our models rich and diverse data to train with.

References

[1] Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, 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. 2024. 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 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

[2] Sanja Štajner, Horacio Saggion, Daniel Ferrés, Matthew Shardlow, Kim Cheng Sheang, Kai North, Marcos Zampieri, and Wei Xu. 2022. *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual).

[3] Gustavo Paetzold and Lucia Specia. 2016. Benchmarking Lexical Simplification Systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3074–3080, Portorož, Slovenia. European Language Resources Association (ELRA).

[4] Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier Automatic Sentence Simplification Evaluation. 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): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

[5] 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.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] McLaughlin, G., 1969. SMOG grading–A new readability formula in the journal of reading.

[8] 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.

[9] Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words.. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

[10] Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

[11] 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.

[12] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

[13] Goran Glavaš and Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora?. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

[14] Paetzold, G.H. and Specia, L., 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, *60*, pp.549-593.

[15] Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. Deep learning approaches to lexical simplification: A survey. J. Intell. Inf. Syst. 63, 1 (Feb 2025), 111–134. https://doi.org/10.1007/s10844-024-00882-9