أكاديمية كاوست
KAUST ACADEMY

AI-Driven Text Simplification Unraveling Complex Sentences

Muath Bamuqabil

Jawad Alraimi

Table of Contents

Introduction 4
Methodology 4
Dataset Preparation 5
Model Selection 5
Fine-tuning Process: 5
Evaluation Metrics: 6
Post-Processing: 6
Results and Analysis 6
Challenges Identified 7
Conclusion 7

Abstract

Text simplification is an important task in the field of natural language processing (NLP), aimed at transforming complex sentences into more accessible and easy-to-understand forms. This project investigates the use of transformer-based models for automatic text simplification. The notebook details a workflow that includes data preparation, model implementation, and performance evaluation. The model is trained to simplify complex sentences, and its performance is assessed using the BLEU score, a metric widely used to evaluate the quality of machine-generated text. The results show the potential of AI-driven models in simplifying text, though further optimization is needed to improve accuracy and generalization.

Introduction

As the volume of written content grows, the need for making information accessible to a wider audience becomes increasingly important. Text simplification, a subset of text generation tasks, aims to reduce the complexity of text without losing the original meaning. This has broad applications, from making educational materials more accessible to improving content for people with language impairments or non-native speakers.

The idea behind this project is to assist non-expert readers in understanding complex material within specific fields. By simplifying complex sentences, the model helps bridge the gap between technical jargon and everyday language, enabling a wider audience to comprehend specialized content. This capability can be especially valuable in fields like science, law, and medicine, where dense and complex text can be challenging for the average reader.

In recent years, the use of deep learning models, particularly transformer-based architectures, has gained traction in various NLP tasks, including text generation and summarization. Text simplification presents a unique challenge because it requires not only preserving meaning but also producing grammatically correct and contextually appropriate simpler sentences. Traditional rule-based systems often fall short in handling the nuances of language, prompting the exploration of more advanced AI techniques.

This project leverages a pre-trained transformer model, fine-tuned on a text simplification task. The goal is to develop an automatic system capable of simplifying complex sentences, with the ultimate aim of improving text accessibility. The performance of the system is measured using BLEU scores, comparing the model's output with reference sentences from a test dataset.

Methodology

The methodology for this text simplification project encompasses several crucial stages, each contributing to the overall performance and effectiveness of the model:

Dataset Preparation:

The dataset used for this task consists of pairs of complex and simplified sentences. Each pair includes a complex sentence as input and its corresponding simplified version as the target output.

The dataset is split into training, validation, and test sets to ensure that the model is trained on one subset of the data and evaluated on unseen examples from the test set.

Preprocessing steps include tokenization, where sentences are broken down into individual words or subwords, and padding, ensuring that all sequences are of uniform length for batch processing.

Model Selection:

The model employed in this project is the T5 (Text-To-Text Transfer Transformer) model, a transformer-based architecture known for its versatility across various NLP tasks. Developed by Google, T5 converts all NLP tasks into a text-to-text format, making it particularly well-suited for text generation tasks such as text simplification. The T5 model is implemented using the Hugging Face Transformers library, which provides access to state-of-the-art pre-trained models and easy fine-tuning capabilities.

Transformer models, including BERT, GPT, and T5, have revolutionized NLP by effectively capturing complex language patterns and dependencies over long sequences of text. T5 is especially powerful because of its unified framework, where both input and output are treated as text, allowing it to handle a wide range of tasks within the same model architecture.

In this project, we fine-tuned the T5 model specifically for the task of text simplification. By leveraging the large-scale language understanding gained during pre-training, the model adapts to the nuances of simplifying complex sentences while maintaining their original meaning. Fine-tuning the T5 model on a specific dataset of complex and simplified sentence pairs enables it to learn the intricacies of generating more accessible text.

Fine-tuning Process:

Fine-tuning involves training the model on the task-specific dataset, adjusting the weights of the pre-trained model to improve its performance on the text simplification task.

Key hyperparameters, such as learning rate, batch size, and the number of epochs, are carefully tuned to optimize the model's performance.

During the fine-tuning process, the model learns to map complex sentences to their simplified versions, adjusting its internal representations to produce more accessible text.

Evaluation Metrics:

The primary evaluation metric used in this project is the BLEU (Bilingual Evaluation Understudy) score. BLEU is a precision-based metric that compares the n-grams (sequences of n words) in the generated sentence with those in the reference sentence. It calculates how many n-grams from the candidate sentence appear in the reference sentence, providing an indication of the model's accuracy in generating similar text.

The project computes BLEU scores at both the individual sentence level and across the entire test set. The average BLEU score provides a summary of the model's overall performance.

Although BLEU is a widely used metric in text generation tasks, it has limitations, such as its inability to capture the fluency or grammatical correctness of the generated text. Therefore, further analysis of the model's output, including qualitative assessments, is necessary to fully understand its effectiveness.

Post-Processing:

After generating simplified sentences, post-processing steps may be applied to refine the model's output. These steps could include removing unnecessary punctuation, fixing grammatical errors, or applying additional heuristics to improve the readability of the simplified text.

Results and Analysis

The notebook demonstrates the process of training and evaluating the text simplification model. After training, the model is evaluated on the test set, and BLEU scores are computed to measure the quality of the simplified sentences.

Sentence-level BLEU Scores: The model generates simplified sentences for individual test cases, and BLEU scores are computed for each sentence. This provides a granular view of the model's performance, highlighting specific examples where the model succeeds or fails in generating accurate simplifications.

Average BLEU Score: The project calculates an average BLEU score of approximately 0.17 across the test set. This score indicates that while the model is capable of producing simplifications, the accuracy is moderate, and there is room for improvement. The relatively low BLEU score suggests

that the model may struggle with certain complex sentences, either by oversimplifying and losing important details or by failing to adequately simplify the text.

Challenges Identified: Some of the key challenges identified during the analysis include:

Preserving Meaning: Ensuring that the simplified text retains the meaning of the original complex sentence is a major challenge. The model occasionally produces sentences that, while simpler, may miss critical details.

Fluency and Readability: Although the BLEU score measures overlap with reference sentences, it does not assess the fluency and grammatical correctness of the output. Qualitative analysis shows that some generated sentences may be structurally correct but lack natural readability.

Diversity of Input: The complexity of the input sentences can vary widely, from slightly complex to very intricate. The model's performance may degrade when dealing with particularly convoluted or abstract sentences.

Conclusion

This project demonstrates the potential of AI-driven models for text simplification, utilizing a transformer-based approach to simplify complex sentences. The model achieves moderate success, with an average BLEU score of 0.17, indicating that the generated sentences have a reasonable overlap with the reference simplified sentences. However, the BLEU score also reveals that the model has limitations, particularly in handling diverse and complex sentence structures.

To further enhance the model's performance, future work could explore several avenues:

Model Optimization: Experimenting with different model architectures, such as larger transformer models or models specifically designed for text generation tasks, could improve the simplification quality.

Data Augmentation: Expanding the dataset with more examples of complex and simplified sentence pairs, or utilizing data augmentation techniques, could help the model generalize better.

Alternative Metrics: In addition to BLEU, incorporating other evaluation metrics, such as ROUGE or METEOR, or conducting human evaluations, could provide a more comprehensive assessment of the model's output quality.

Post-Processing Enhancements: Implementing more sophisticated post-processing techniques could improve the readability and fluency of the simplified sentences.

In conclusion, while this AI-driven text simplification model shows promise, it requires further refinement to achieve high-quality simplification across a broad range of sentence complexities. With continued development, such models have the potential to significantly enhance the accessibility of written content, making information more understandable and widely accessible.