It's Simple!: QA-based Lexical Simplification and Controlled Text Generation

Deepti Saravanan
New York University / New York, NY
ds6812@nyu.edu

Yuxuan Xia
New York University / New York, NY
yx2432@nyu.edu

Abstract

Text Simplification is the process of reducing the complexity of a text to make it easier to comprehend, while also preserving the semantics of the sentence. In this paper, we propose an end-to-end framework for domain-independent sentence-level text simplification, by combining text pre-processing techniques for lexical simplification and controlled text generation techniques for structural simplification. The lexical simplification components employ various natural language processing techniques, such as POS tagging, question-answering, and masked token prediction using Large Language Models. The proposed framework performed almost on par with the baseline model with a SARI score difference of 0.6%. The project can have several applications, such as improving accessibility for people with cognitive disabilities and non-native speakers of the language.

1 Introduction

Text simplification is a crucial task in natural language processing (NLP) that aims to transform complex and difficult-to-read texts into simpler ones while preserving their meaning and essential information. It involves a variety of linguistic and computational challenges, such as identifying difficult words and phrases, rephrasing or substituting them with simpler alternatives, and adapting the syntactic and semantic structures of the text to make it more accessible to the target audience.

1.1 Significance

Simplifying text make it more accessible to a wider audience, including people with cognitive disabilities and non-native speakers. It also helps readers better understand complex ideas and information.

1.2 Our Contribution

In this paper, we propose an end-to-end framework for domain-independent sentence-level text simplification. The following points summarize the main contributions of our work:

- (1) We propose a refined architecture for the task of text simplification, building on top of the ACCESS model.
- (2) We propose a question-answering-based lexical simplification as a viable alternative to complex word identification and replacement.
- (3) We add phrase-level and POS tag-level feature characteristic input for controlled text generation to create a modified ACCESS model.

2 Methodology

Text Simplification generally involves either lexical or structural simplification, or both. Lexical Simplification helps in replacing hard-to-understand words with their simpler synonyms while preserving the overall semantics of the sentence. Structural Simplification, on the other hand, focuses on text characteristics such as the length of the sentence, the relative placement of POS tags, and the depth of the dependency tree. By combining information from both these aspects, it is possible to gain a more complete understanding of the richness of a sentence.

2.1 Problem Formulation

Given a complex text T_c, the objective of text simplification is to generate a simplified text T_s that retains the core semantics of T_c while making it easier to comprehend. The simplified text T_s should be grammatically correct and semantically coherent.

In order to achieve this, the text simplification task involves several subtasks, such as identifying complex words and replacing them with simpler alternatives, rephrasing or splitting long and convoluted sentences, and ensuring the overall coherence and readability of the simplified text. The quality

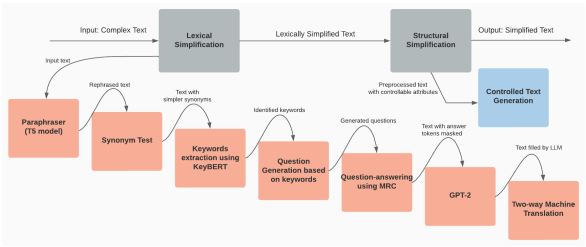

Figure 1: Proposed Framework for Text Simplification

of the simplified text T_s can be evaluated from multiple persepctives using various metrics, such as readability scores and semantic similarity.

2.2 Proposed Framework

Our proposed framework comprises two main components, as can be seen in Figure 1. The Lexical Simplification Component acts as a text preprocessing step where we propose a new approach not tried in past literature in this context. In short, our technique involves question-answering-based important tokens identification where questions are framed on identified key Nouns from different perspectives covering the who, when, and how aspects. This is a possible feasible alternative for complex word identification, which is a challenging task in itself given the degree of complexity varies with readers.

Figure 1 further illustrates the subcomponents of the Lexical Simplification Component which are executed sequentially. A pre-trained paraphraser T5 model was employed to paraphrase the input text. Then, the verbs in the paraphrased text are identified and replaced if there exists a simpler synonym. Following this, a pretrained KeyBERT was used to identify Noun keywords in the text. Using these keywords, questions from the aspects of what it is, what it does, and how it does it are generated. A pretrained BERT-based Machine Reading Comprehension (MRC) model originally trained on the SQUAD dataset was used to extract answers for the generated questions. These answer tokens, excluding the identified Named Entities, are the potential candidates for replacement, similar to the idea of complex word replacement. This set of tokens is masked and fed into a pretrained Generative Pre-trained Transformer 2 (GPT-2) model for filling in using the surrounding context. GPT-2 is a state-of-the-art Large Language Model (LLM) that is trained on a massive amount of text data using unsupervised learning techniques. Finally, a two-way translation technique was applied where the English text was converted to Chinese and then translated back to English. This serves as the data augmentation process.

Table 1 illustrates an example of the intermediate results of the input text as it passes through the different subcomponents. The output at this point is lexically simplified and fed into the next component. The Structural Simplification Component employs the Controlled Text Generation Technique, inspired by the paper on the ACCESS model.

The AudienCe-CEntric Sentence Simplification (ACCESS) model (Martin et al., 2020) used explicit control tokens such as Dependency Tree Depth to control the process of Sentence Simplification. To achieve this, a transformer model was trained on WikiLarge dataset using the FairSeq toolkit.

The following structure-based control attributes are primarily employed in the original model:

Compression Ratio is calculated as the ratio of the length of a simple sentence to the length of a complex sentence at the character level. Complex sentences are expected to be longer in general.

$$CompressionRatio = \frac{len(target)}{len(input)} \quad (1)$$

Dependency Tree Depth Ratio gives the depth of the dependency tree that captures the long-distance relationship of the tokens in the input sentence. Complex sentences are expected to show deeper dependency between the tokens.

$$DepTreeDepthRatio = \frac{DepTreeDepth(target)}{DepTreeDepth(input)} \quad (2)$$

While the compression ratio captures the sentence length, the dependency tree depth ratio captures the token-level structural dependency. This fails to capture the contextual placement of these tokens as they are just considered individually.

We propose an enhanced version of the ACCESS model by incorporating two additional structural features of the input text:

(1) **Dependent Clause Ratio** captures the structural and dependency characteristics of the input text at the phrase level. Complex sentences are expected to have more number of dependent clauses, contributing to more complex phrasal structure.

$$DepClauseRatio = \frac{NumDepClause(target)}{NumDepClause(input)} \quad (3)$$

Subcomponent	Intermediate Output
Paraphraser	genes have been **extended** to **allow** breeders **to create** desired germlines for new crops
Synonym Test	gene have been **widen** to **let** breeder to **make** desired germlines for new crop
Keyword Extraction	desired germlines, gene, new crop
Question Generation	**What is** desired germlines? **What does** desired germlines **do**? **How does** desired germlines **work**?
Question Answering	for new crop to **let** breeder to **make** desired germlines gene have been **widen**
Masked Prediction	gene have been **widen** to **let** breeder to **make** desired germlines for new crop
Machine Translation	The gene has been widen to **leave** the breeder to make desired germ lines for a new **culture**

Table 1: Lexical Simplification Subcomponents and Intermediate Results

(2) **Number of distinct POS tags Ratio** indicates the variety of syntactic categories used to label the words in a sentence. This goes one level deeper than the dependency tree depth ratio by analyzing the actual POS tags (nodes) instead of just the edges of the tree (relationship between tags). Complex sentences are expected to have more number of different POS tags due to the possible presence of redundant and unnecessary information.

$$POSRatio = \frac{NumPOS(target)}{NumPOS(input)} \quad (4)$$

The proposed text preprocessing component along with the modified model was evaluated and compared against the performance of the original model.

3 Experiments

3.1 Dataset

To accomplish the task of sentence simplification in English, the Turk dataset was used where each complex sentence has 8 human simplifications created by Amazon Mechanical Turk workers. It is a multi-reference dataset containing 2,359 sentences from the Parallel Wikipedia Simplification (PWKP) corpus.

3.2 Model Hyperparameters

The base model architecture was adapted from the original ACCESS model(Martin et al., 2020) where the backbone is a sequence-to-sequence neural network architecture with content-based attention mechanism. The original training protocol was followed with a maximum of 100 epochs and an early stopping method when the validation score did not improve over 10 epochs. Adam optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of $lr = 0.00011$. For the generation part, beam search was used with a beam size of 8.

As for the lexical simplification component, various natural language processing toolkits were employed such as NLTK, transformers for GPT-2, and TextBlob for the Translation task.

3.3 Evaluation Metrics

The simplicity and readability of the generated sentence was evaluated using SARI (Xu et al., 2016). SARI compared the predicted simplification by averaging F1 scores for three n-gram operations: additions, keeps and deletions.

3.4 Result and Ablation Studies

The importance of the two components were analyzed by removing them and evaluating the model performance. Table 2 shows the performance comparison of original ACCESS model without our preprocessing component, modified ACCESS model without our preprocessing component, original ACCESS model with our preprocessing component and modified ACCESS model with our preprocessing component.

As can be inferred, the proposed model performs slightly worse than the original model. But if just one of the components of the framework is retained, then the model does not perform well. Hence, the components work better together. The proposed model *might* perform better in case of domain-specific data due to the possible presence of more dependency clauses.

SARI Score	Additional Processors	No Additional Processors
Include Preprocessing	36.07 (Proposed Model)	34.17
No Preprocessing	35.63	**36.3** (Original ACCESS)

Table 2: Evaluation Results

4 Related Work

Over the years, text simplification has gained increasing attention from both academia and industry, leading to the development of various algorithms and models for automatic text simplification.

(Vo et al., 2022) employed a decoder for structure simplification and a back-translation for lexical simplification. Decoder generates a shorter sentence that has a similar semantic meaning to the input by pretrained SBERT. Back-translation translates the decoder output to another language and then translates it back. (Mehta et al., 2022) showed how context simplification could improve the performance of MRC-based event extraction by more than 5% for actor extraction and more than 10% for target extraction. Text simplification suffers from scarcity of data. (Sun et al., 2023) proposed that summarization is sometimes a kind of simplification and uses summarization dataset to create simplification data pairs and expand the training dataset. (Martin et al., 2021) proposed a new approach to sentence simplification that can overcome the unavailability of labeled data by training models using sentence-level paraphrase data. (Lu et al., 2021) leveraged the technique of translation to Germany and back to English for lexical simplification. (Sun et al., 2021) successfully performed document-level simplification, built the corresponding dataset and introduced modified evaluation metrics. Domain-specific simplification was accomplished by (Cemri et al., 2022) using an unsupervised approach that circumvents the labeled data unavailability.

5 Conclusion

In this study, we proposed an end-to-end framework for domain-independent text simplification task. Two main components were involved. The first component focused on the lexical simplification of the input text by employing a series of techniques. It employed question-answering based extraction of tokens of high interest as an alternative to complex words identification. These tokens were then masked for GPT-2 to predict using surrounding context. Finally, a two-way translation was employed for data augmentation. The lexically simplified output was then fed into the structural simplification component that employed controlled text generation. In addition to the control attributes employed by the original model, we introduced a phrase-level and the POS tag-level features.

Evaluation result showed that the modified model along with the preprocessing component performed almost on par with the original model with a SARI score of 36.07. Ablation studies concluded that the two proposed components work better together.

Potential future work would be to create domain-specific text simplification data with the help of domain experts. This enables us to use the proposed model for transfer learning by freezing the lower layers and finetuning the rest with the domain-specific dataset created. This inturn would evaluate the adaptability of the model.

Limitations

The main limitation faced by the study is the unavailability of labeled domain-specific data such as legal and medical texts. This blocked the possibility of testing the adaptability of the model to niche domains via transfer learning technique. Labeling of such domain-specific texts scraped from the internet was also not feasible due to the unavailability of domain experts. In addition, human evaluation was not performed due to the same issue.

References

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022. Unsupervised simplification of legal texts.

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2021. Muss: Multilingual unsupervised sentence simplification by mining paraphrases.

Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. 2020. Controllable sentence simplification.

Sneha Mehta, Huzefa Rangwala, and Naren Ramakrishnan. 2022. Improving zero-shot event extraction via sentence simplification.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Renliang Sun, Zhixian Yang, and Xiaojun Wan. 2023. Exploiting summarization data to help text simplification.

Vy Vo, Weiqing Wang, and Wray Buntine. 2022. Unsupervised sentence simplification via dependency parsing.

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.