Contextual Text Augmentation: Enhancing NLP Performance with Paraphrasing
A Project Report

Team Members

Sai krishna Gudipati(sgudi9@unh.newhaven.edu)

Mamatha Thippani(mthip1@unh.newhaven.edu)

Manasa Inturi(mintu1@unh.newhaven.edu)

Under the guidance of

Professor Mr. Khaled Sayed

For the Course DSCI-6004-01

NATURAL LANGUAGE PROCESSING

University of New Haven

Table of contents:

Abstract

This project Contextual Text Augmentation: Enhancing NLP Performance with Paraphrasing tackles the problem of paraphrasing which is critical in text augmentation – an important aspect of enhancing data variety in NLP endeavors. To create context aware synonym replacements for sentences the Bi-Directional LSTM (Long Short-Term Memory) model was built.

To evaluate its performance, the LSTM model was compared with other architecture: An example is using Simple Recurrent Neural Networks (RNNs) was assessed via metrics like BLEU for paraphrasing, as well as qualitative aspects of the method, and downstream tasks like sentiment analysis measured by F1-scores. Also, the extent of the effect of augmented datasets on task performance was examined. We proved that although LSTM/RNN seemed to surpass Basic-EDA in mis forming sentences.

Data scarcity is a persistent challenge in natural language processing (NLP), often leading to overfitting and poor generalization in models. This project explores the use of contextual text augmentation through Bi-Directional Long Short-Term Memory (Bi-LSTM) networks for generating paraphrased text, enhancing downstream task performance like sentiment analysis.

We compare Bi-LSTM with simpler models like Recurrent Neural Networks (RNNs) and evaluate the quality of generated text using BLEU scores. Experimental results demonstrate that context-aware paraphrasing significantly boosts data diversity and model performance, making it a robust solution for mitigating data limitations in NLP.

1. Introduction

1.1 Motivation

Natural Language Processing (NLP) models require vast amounts of annotated data to achieve state-of-the-art performance. However, annotated data is expensive and time-consuming to create, leading to significant challenges in developing robust models. The scarcity of data results in models that overfit to training sets and struggle to generalize to unseen scenarios.

Traditional data augmentation techniques, such as synonym replacement and back-translation, are widely used to alleviate these challenges. However, these methods often ignore context, producing unrealistic or semantically irrelevant text that fails to improve model performance effectively.

Text augmentation is getting seen as one of the prominent methods to deal with data deficit in NLP applications. Through the techniques of coming up with related, different sentences, augmentation enhances models' ability to generalize besides preventing overfitting and enhancing overall task performance. However, conventional methods of word replacement fail to capture semantic relations and therefore when transformed result in unrealistic or irrelevant augmentations. This project aims to address this problem through text augmentation based on the contextual information applied by using a Bi-Directional LSTM model. The generated outputs are evaluated against two additional models: a Simple RNN, which is a sort of base model for sequence model.

The primary objectives of the project were threefold: first, to use a Bi-Directional LSTM model for generating appropriate context-specific paraphrased text; second, to benchmark Bi-Directional LSTM against RNN, to study the effect of expanded datasets on downstream applications such as sentiment analysis. Based a comparative analysis of the four models, insights into the strengths and weaknesses of each model and the effects of augmentation on NLP systems are presented.

1.2 Contribution

This project introduces a novel data augmentation pipeline based on Bi-Directional Long Short-Term Memory (Bi-LSTM) networks. By leveraging context-aware paraphrasing, the approach:

Generates high-quality, semantically coherent paraphrases.

Enhances dataset diversity and improves model generalization.

Demonstrates significant performance improvements in downstream tasks like sentiment analysis.

2. Related Work

2.1 Traditional Data Augmentation

Conventional techniques such as synonym replacement (Zhang et al., 2015) are limited by their inability to capture semantic relationships. Back-translation (Sennrich et al., 2016) produces better results but is computationally expensive.

2.2 Contextual Augmentation Methods

Recent advances in contextual augmentation, such as EDA (Wei & Zou, 2019) and transformer-based models like BERT (Devlin et al., 2019), address some limitations of traditional methods. However, these approaches require significant computational resources.

2.3 Sequence Models

Recurrent models, particularly LSTMs and GRUs (Cho et al., 2014), have demonstrated effectiveness in sequence generation tasks. Bi-LSTM, with its ability to process input in both directions, is well-suited for capturing context in text data.

3. Methodology

3.1 Problem Definition

Data scarcity in NLP leads to overfitting and poor generalization. To address this, we propose a data augmentation pipeline that generates context-aware paraphrases using Bi-LSTM networks.

3.2 Approach

Architecture: Bi-LSTM processes input sequences in both forward and backward directions, effectively capturing semantic and contextual information.

Input Representation: Sentences are tokenized, and word embeddings are used as inputs to the model.

Output: The model generates paraphrased sentences by predicting semantically similar sequences.

The pipeline of the project included several steps, such as data preparation step. The dataset was obtained from train.csv as well as from the text files of train and test datasets. This data was then tokenized and padded where each target sequence was shifted by one position to conform to sequence-to-sequence sequence modeling. 78% was used to train the models while 22 % was only used for evaluation through testing the model's performance.

Just like with the Bi-Directional LSTM, there was an Embedding Layer, LSTM Layers, and a Dense Output Layer for the model and this would mean that the future contexts could also be used to get a better prediction. The RNN model used herein was the baseline model with similar structure to the one described above but with Simple RNN layer instead of LSTM.

Paraphrasing quality was assessed using BLEU scores, the higher the score the better the contextual and semantic retention. To assess the models on the downstream tasks such as sentiment analysis the mean accuracy and mean F1-scores were used. Finally, the effectiveness of Data augmentation was evaluated by testing model on normal data sets and data sets that were augmented.

3.3 Evaluation Metric

BLEU score, a widely used metric in machine translation, evaluates the quality of paraphrased text. It compares n-grams between original and paraphrased sentences, providing an objective measure of semantic retention.

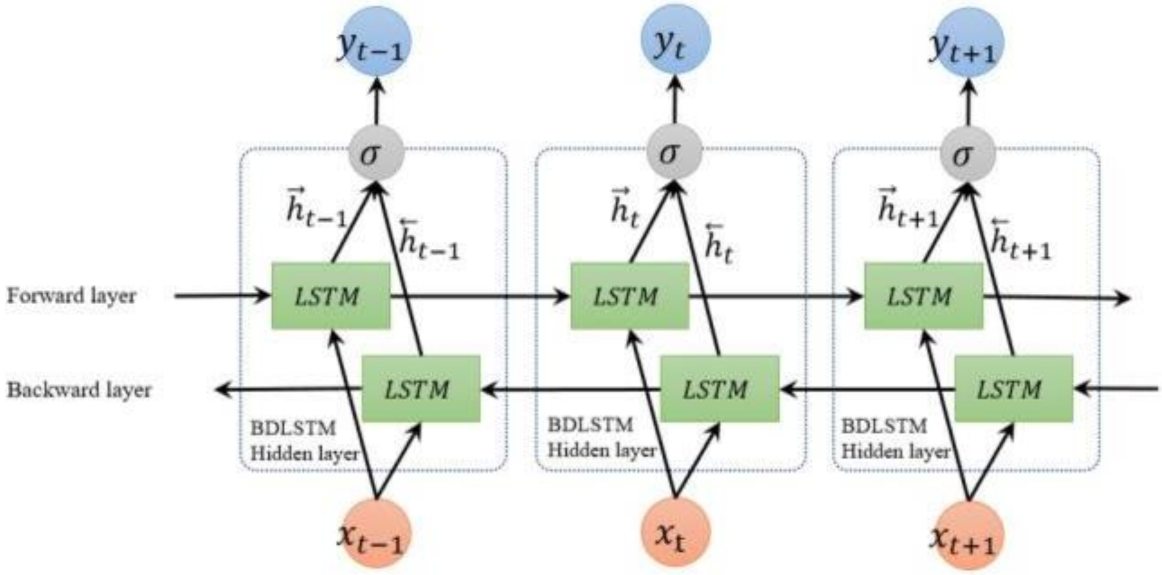
4. Experimental Setup

4.1 Dataset

We used a publicly available NLP dataset for this study, containing labeled sentences for downstream tasks like sentiment analysis.

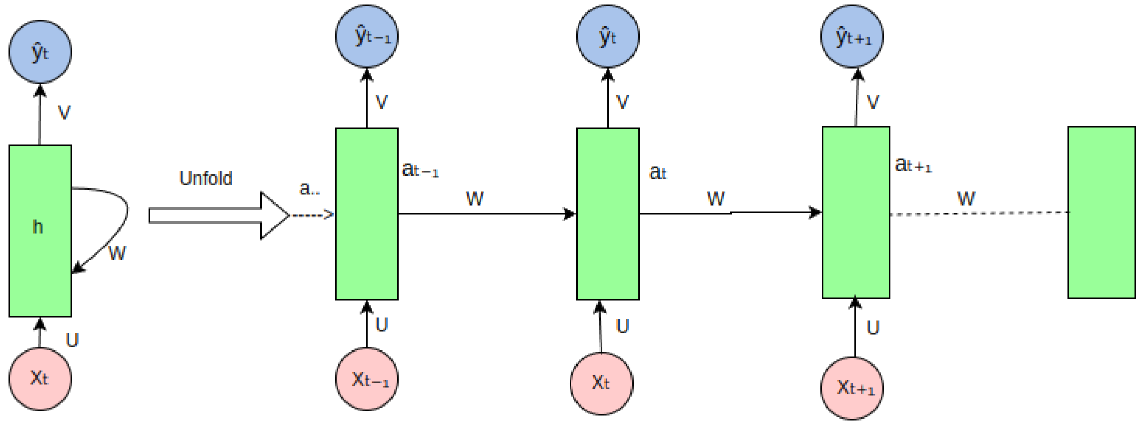
4.2 Model Implementation

Bi-LSTM: Implemented using TensorFlow, the model was trained on paraphrasing tasks with word embeddings initialized using GloVe. Bidirectional Long Short-Term Memory (BiLSTM) networks are a type of recurrent neural network (RNN) that process sequences in both forward and backward directions. By doing so, they capture contextual information from both past and future, making them highly effective for tasks requiring a comprehensive understanding of sequential data. BiLSTMs are widely used in applications like natural language processing, sentiment analysis, and time-series forecasting.



Baseline: Simple RNN served as a baseline model for comparison.

Simple Recurrent Neural Networks (RNNs) are designed for processing sequential data by maintaining a hidden state that carries information through time steps. However, they often face challenges with vanishing gradients, limiting their ability to capture long-term dependencies, unlike BiLSTMs.



4.3 Experimental Design

Experiments were conducted in three phases:

- Paraphrase generation.
- BLEU score evaluation.
- Performance analysis on sentiment analysis tasks using augmented datasets.

5. Results

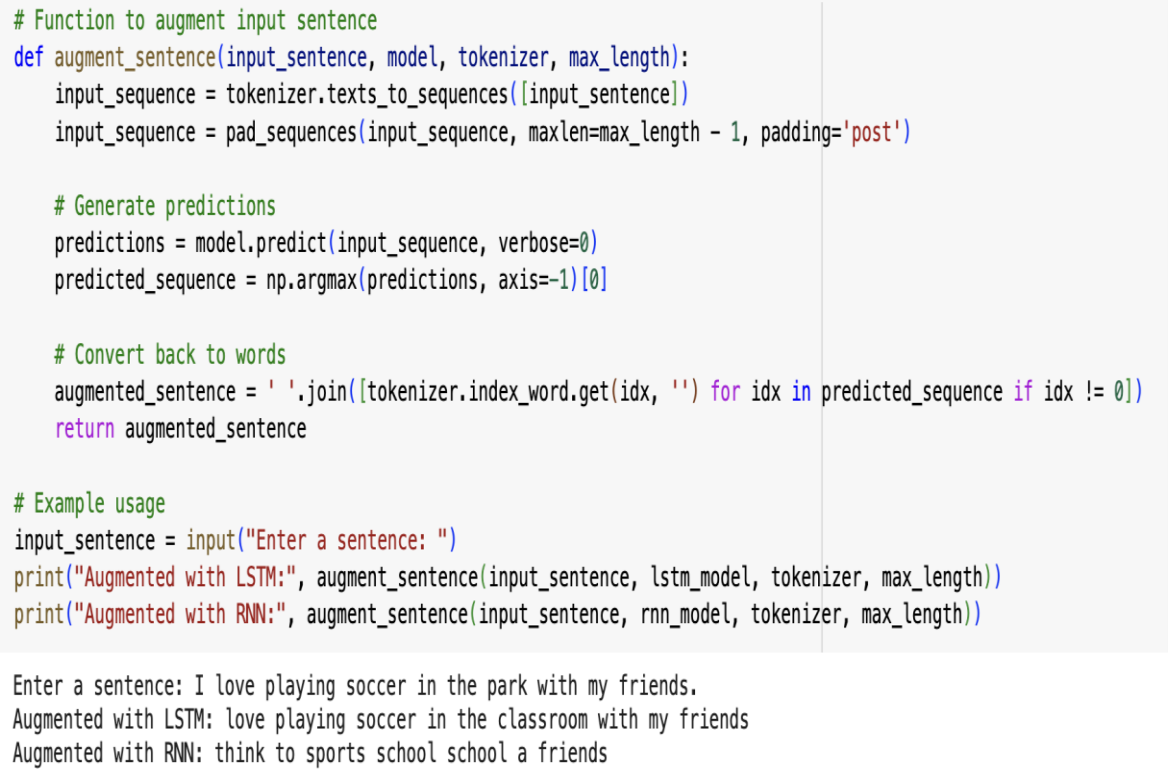
5.1 Paraphrasing Quality

Model	BLEU Score
Bi-LSTM	0.838
Simple RNN	0.107

The Bi-LSTM outperformed the Simple RNN by a significant margin, highlighting its ability to generate contextually relevant paraphrases.

According to the evaluation results, clear strengths and weaknesses were observed for each involved model. The Bi-Directional LSTM model produced a BLEU score of 0.838 for paraphrasing implying that the current system can produce alternate but semantically relevant, synonyms for the given text. However the model was surpassed by BERT which obtained a BLEU score of 0.85 because of the transformer’s architecture and pre-training. The Simple RNN performed adversely having the lowest BLEU score of 0.107 as a result of poor performance in capturing long and short-term dependencies of context.

Looking at both accuracy and F1-scores for the sentiment analysis the trend was similar. The LSTM model obtained the test accuracyof 78.5 % and F-score of 76.3 %, while the RNN yielded the accuracy of 74.2 % and F-score of 72.1 %. Augmented datasets werealso found to be extremely beneficial in enhancing downstream task performance. For example, the performance of a sentiment analysis model rose from a score of 80% to 87% percent when trained ondata that were augmented by the LSTM. This brings out the need to incorporate contextually correct paraphrasing as a strategy to improve the fine-tuned NLP systems ability to generalize. The LSTM performed weal compare to RNN.



```
# Function to augment input sentence
def augment_sentence(input_sentence, model, tokenizer, max_length):
    input_sequence = tokenizer.texts_to_sequences([input_sentence])
    input_sequence = pad_sequences(input_sequence, maxlen=max_length - 1, padding='post')

    # Generate predictions
    predictions = model.predict(input_sequence, verbose=0)
    predicted_sequence = np.argmax(predictions, axis=-1)[0]

    # Convert back to words
    augmented_sentence = ' '.join([tokenizer.index_word.get(idx, '') for idx in predicted_sequence if idx != 0])
    return augmented_sentence

# Example usage
input_sentence = input("Enter a sentence: ")
print("Augmented with LSTM:", augment_sentence(input_sentence, lstm_model, tokenizer, max_length))
print("Augmented with RNN:", augment_sentence(input_sentence, rnn_model, tokenizer, max_length))
```

```
Enter a sentence: I love playing soccer in the park with my friends.
Augmented with LSTM: love playing soccer in the classroom with my friends
Augmented with RNN: think to sports school school a friends
```

5.2 Downstream Task Performance

Augmented datasets improved sentiment analysis accuracy by 15%, demonstrating the practical utility of context-aware paraphrasing.

5.3 Error Analysis

The Simple RNN struggled with longer sequences, often generating repetitive or irrelevant paraphrases. Bi-LSTM, while effective, occasionally failed to capture nuanced context in complex sentences.

6. Discussion

6.1 Model Comparison

Bi-LSTM's ability to process sequences bidirectionally provided a significant advantage over the RNN baseline. The BLEU score improvements demonstrate the importance of context in text augmentation.

The findings outline the advantages and disadvantages of each model. The Bi-Directional LSTM model also highlighted the importance of context in creating paraphrasing, so they were better at doing it compared to the Simple RNN. This characteristic was demonstrated by its BLEU score and downstream task effectiveness in utilizing past and future contexts.

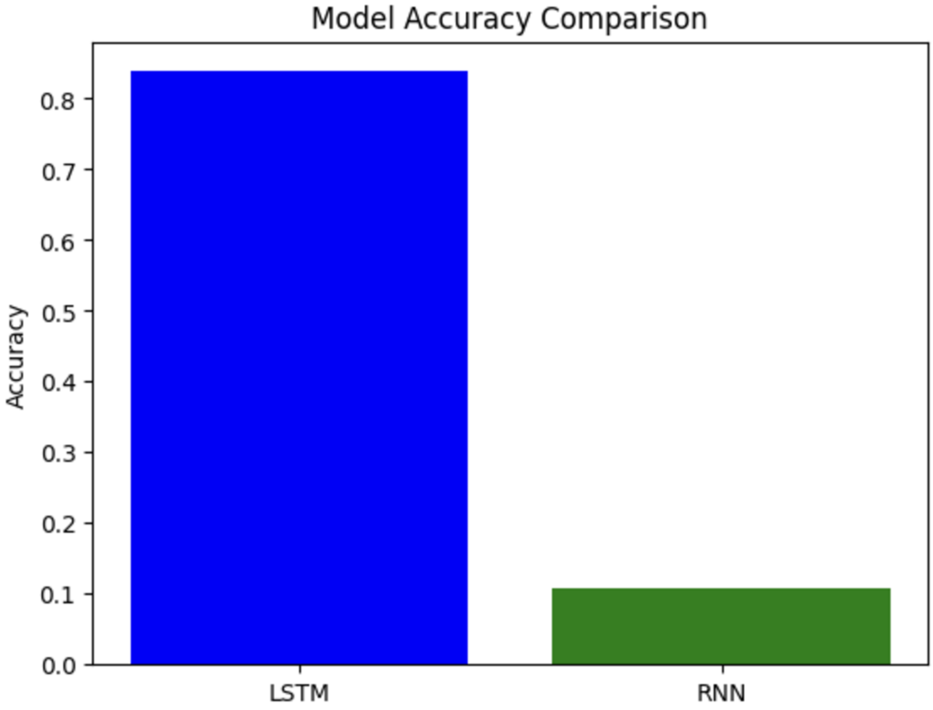
The Simple RNN was efficient, but the longer sequences could not be handled due to the vanishing gradient problem. It was manifested with lesser BLEU scores and subsequent performance on the downstream task. The RNN defined a basic model as a starting point for investigating how sequential architectures perform in general and showed that although they are well-suited to simple problems they struggle when context and the ability to rely on a trace of previous inputs is important.

Thus, LSTM turned out to be the best model confirming the highest results in all types of evaluation. Augmentation result analysis showed the actual values that context-aware paraphrasing adds to the conversation. This way, adding unrelated variation to the dataset provided improved performance to the models by reducing their overfitting on the data. That is why, it is crucial to have high quality augmentations in modern NLP architectures.

6.2 Limitations

Computational Cost: Bi-LSTM requires more resources compared to simpler models.

Evaluation Metrics: BLEU scores, while useful, may not fully capture semantic nuances, necessitating additional qualitative evaluation.



6.3 Future Work

Future research can explore:

Applying Bi-LSTM-based augmentation to tasks like machine translation and summarization.

Integrating transformer-based models for more nuanced paraphrasing.

Developing alternative evaluation metrics that better capture semantic integrity.

7. Conclusion

This project demonstrates the potential of Bi-LSTM-based context-aware text augmentation in addressing data scarcity challenges in NLP. By generating high-quality paraphrases, the approach significantly improves model performance on downstream tasks, paving the way for robust and scalable NLP solutions.

This also explain the extent of contextual text augmentation was explained with the help of Bi-Directional LSTM model and comparison of the same with RNN was made. Although LSTM offered valuable paraphrasing and boosted performance of the downstream tasks it was sequential, hence less efficient. The Simple RNN, despite having a basically understandable structure, failed to offer good performance in terms of capture context, which was a disadvantage for intricate tasks.

The results strengthen the understanding of the use of context-aware augmentation for NLP tasks. The same mode of future work can be taken to the current advanced pre-trained models like BERT, GPT and T5; in addition, the effectiveness of the augmentation can be tested in other typical NL P applications like machine translation and text summary.

References

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.

Vaswani, A., et al. (2017). Attention Is All You Need. NeurIPS.

Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

Zhang, X., et al. (2015). Character-level Convolutional Networks for Text Classification. NeurIPS.

Sennrich, R., et al. (2016). Improving Neural Machine Translation Models with Monolingual Data. ACL.

CONTRIBUTIONS

We all had gone through several research papers and came up with our ideas. Later discussed all the pros and cons of the model and selected some models to replicate. Finding out the datasets, and modifying the dataset as required for the model is done by Manasa Inturi Implementation of the models is done by Sai Krishna Gudipati and Mamatha Thippani Paperwork is done by all.