Analyzing the Performance of Plain RNNs and LSTM-based RNNs in Language Translation

Nelson Dufitimana, Kelvin Darfour, Bereket Nigussie, Wilbert Fundira

November 12, 2023

1 Introduction

When humans engage with written text, their thoughts persist through the reading. This means that as they transition from one sentence or paragraph to another, their comprehension builds upon prior content rather than starting from scratch. Despite the apparent simplicity of this cognitive task, traditional neural networks lack this capability. The introduction of Recurrent Neural Networks (RNNs) addresses this limitation, incorporating "loops" to facilitate information persistence. However, as the temporal gap widens between the current sentence and the context to be retained, RNNs encounter challenges with remembering this context. Long Short-Term Memory networks (LSTMs) offer a solution. LSTMs are a specialized type of RNN capable of maintaining long-term dependencies. They have proven successful in tasks that demand the retention of substantial contextual information. This project seeks to evaluate the effectiveness of LSTMs compared to plain RNNs. The following sections will introduce our primary hypothesis, the chosen model, the experiments to be conducted, and the methods for validation as well as the data we intend to use for our work.

2 Central Hypothesis

Does the utilization of LSTM-based RNNs in language translation tasks improve result translation accuracy and fluency, especially when applied to languages with complex grammatical structures and syntax, compared to plain RNNs?

3 Algorithm\Model

For the translation tasks, we'll rely on Tensorflow and Keras to put together an Encoder-Decoder, Sequence 2 Sequence[**?**] Language translation model first using Recurrent Neural Network and then improving the plain RNN model with Long Short Term Memory for comparison.

4 Data

For development purposes, we'll rely on a toy French-to-English dataset that is small enough to allow us to have a working model that can translate from French to English. This would be a proof of concept. With the proof of concept, we'll go ahead and use a larger dataset from OPUS . . . the open parallel corpus[1], and to be specific we'll use their Many-to-English Translation data[2], which contains a translation of over a hundred languages to English. We'll attempt to translate an indigenous language, Bambara. We'll not be creating a new dataset for real-world problems since we are relying on translated text from an existing corpus. This data is a standard repository data set and is managed using PIP.

5 Experiments

For experiments, we'll rely on cross-validation with k-fold and repeated hold out on our dataset. The analysis will rely on the accuracy and BLEU[**?**] Scores to gauge accuracy as we change the number of folds. To see changes in our accuracy and BLEU scores, we'll be tuning the number of units in our layers, varying both the loss and optimizer functions, and adding an embedding layer. We'll also use different tokenizers to see the impacts of tokenization methods on the results. With a certain number of results from multiple experiments, we'll also calculate our results' statistical validity(p-values) to comment on the differences between the data.

6 Impacts

The global linguistic landscape is rich and diverse, with many languages, particularly indigenous ones, remaining underrepresented on the global stage. For this project, we hope to bring an indigenous language like Bambara to the forefront, facilitating its accessibility to a broader global audience through English translation. This project stands to benefit various groups, including language speakers seeking connection through their language, linguistics scholars engaged in the study of indigenous languages, language instructors, and numerous other communities.

While acknowledging the potential positive impacts of this endeavor, it is essential to recognize potential challenges. Adverse effects, though not anticipated to be significant, may include inaccuracies in translation leading to miscommunication and general language confusion. We will work to mitigate such issues to ensure the overall success and positive impact of this project.

[1] `https://opus.nlpl.eu/`
[2] `http://rtg.isi.edu/many-eng/`

References

[1] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[2] https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-ke
html

[3] https://machinelearningmastery.com/calculate-bleu-score-for-text-python/