NANYANG TECHNOLOGICAL UNIVERSITY
AI6127-DEEP NEURAL NETWORKS FOR NATURAL LANGUAGE PROCESSING
Assignment 2 Individual
Topic: Sequence-to-Sequence (Seq2Seq) Model For Machine Translation
NTAMBARA Etienne G2304253K
April 30, 2024

NTAMBARA Etienne G2304253K [1 2 3] April 30, 2024

1. Introduction

In this assignment, we explore the implementation and analysis of a machine translation model using the sequence-to-sequence (seq2seq) architecture (French-to-English), as pioneered by (Sutskever et al., 2014). Our experiments began with a GRU-based example, extending our exploration to substitutions with LSTM and bi-LSTM units, integrating attention mechanisms as discussed by (Bahdanau et al., 2015), and further experimenting with a Transformer Encoder inspired by (Vaswani et al., 2017). This all-encompassing method sought to determine how these different configurations affected translation quality, which was measured quantitatively using ROUGE scores. Standardized learning rate of 0.01 was used in the experimental setting, and SGD was used as the optimizer for both the encoder and the decoder throughout a total of five epochs. This framework was used to identify the most effective model adjustments to improve translation performance in a regulated yet diverse computing environment.

2. Experimentation Results From Different Tasks

Performance Metrics:
To compute the following metrics used in the evaluations:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Remember, both high precision and recall indicating that our model is good. Again a high value of F_1-measure/score ensures that precision and recall are reasonably high (Zhang, 2024).

Task Two: GRU

Table 1. ROUGE Scores for Training

Metric	F-measure	Precision	Recall
Rouge1	0.79736894	0.74745107	0.86145955
Rouge2	0.67777973	0.6236413	0.7515094

The ROUGE-1 score from Table 1 demonstrates the model's high degree of lexical similarity, with an F-measure of 0.79736894. The precision of 0.74745107 indicates a robust match of unigrams between the generated translations and the reference translations. Additionally, the recall of 0.86145955 signifies that the model effectively captures nearly all unigrams present in the reference translations, indicating comprehensive coverage.

The ROUGE-2 score, as seen in Table 1, reflects the model's proficiency in handling bigram sequences, essential for maintaining a logical and coherent structure in the generated translations. The F-measure of 0.67777973, combined with a recall of 0.7515094, suggests effective incorporation of bigram sequences from the reference translations. The precision of 0.6236413 further confirms that a significant portion of these sequences in the model's output matches those in the references.

Table 2. ROUGE Scores for Testing

Metric	F-measure	Precision	Recall
Rouge1	0.66455585	0.62737286	0.7160275
Rouge2	0.49348292	0.45716476	0.5456641

In Table 2, the ROUGE-1 score for testing displays an F-measure of 0.66455585, indicating a reduction in unigram

overlap compared to the training results. The precision of 0.62737286 shows a decrease, suggesting that fewer unigrams in the model's output match those in the reference translations during testing. The recall of 0.7160275 indicates a diminished capacity to capture the reference unigrams compared to training.

The ROUGE-2 score in Table 2 shows an F-measure of 0.49348292, which represents a further decline in the model's ability to predict bigrams accurately when compared to training scores. The lower precision of 0.45716476 and recall of 0.5456641 further highlight the model's reduced effectiveness in capturing consecutive word pairs in the generated content during testing.

Comparison Analysis

When compared to the training performance given in (Table 1), the F-measure of the ROUGE-1 Score in (Table 2) is 0.66455585, indicating a decrease in unigram overlap. Additionally, it appears that the model's output during the testing phase is less consistent with the reference translations due to the lower precision and recall, which are 0.62737286 and 0.7160275, respectively. In contrast, the ROUGE-2 Score indicates a decline in the model's accuracy in bigram prediction, with all metrics showing declining values from the training to the testing phases. In particular, the precision dropped to 0.45716476, the recall to 0.5456641, and the ROUGE-2 F-measure to 0.49348292.

When compared to the pre-modification performance shown in (Table 1), the performance metrics in (Table 2) generally indicate a decrease in the model's capacity to capture both unigram and bigram relationships in the reference translations. This suggests that post-training adjustments to the model have adversely affected its efficacy, resulting in reduced alignment with the reference translations.

Task Three: LSTM

Table 3. ROUGE Scores for Training

Metric	F-measure	Precision	Recall
Rouge1	0.7412647	0.69928163	0.7959698
Rouge2	0.5989766	0.5556374	0.6587637

ROUGE-1 Score derived from (Table 3): The F-measure of 0.7412647 indicates a strong match of unigrams to the reference translations, demonstrating good quality translation. The precision of 0.69928163 shows that most unigrams produced by the model are found in the reference translations, while the recall of 0.7959698 indicates comprehensive coverage of reference unigrams.

ROUGE-2 Score derived from (Table 3): The F-measure of 0.5989766, while lower than the ROUGE-1 score, still suggests a competent capability of the model to capture bigram sequences. The precision of 0.5556374 and recall of 0.6587637 show that the model performs reasonably well in producing two-word sequences that appear in the reference translations.

Table 4. ROUGE Scores for Testing

Metric	F-measure	Precision	Recall
Rouge1	0.6423244	0.60943043	0.6873895
Rouge2	0.46831056	0.43705827	0.5132266

ROUGE-1 Score from (Table 4): The F-measure of 0.6423244 reflects a good match of unigrams with the reference translations, though with a decline compared to training. The precision of 0.60943043 and the recall of 0.6873895 show a slight decrease in both the accuracy and coverage of unigrams compared to the training phase.

ROUGE-2 Score from (Table 4): The F-measure of 0.46831056, although lower than ROUGE-1, indicates a moderate capability of the model to predict bigram sequences. The precision of 0.43705827 and recall of 0.5132266 suggest that the model has a respectable capacity for capturing sequential bigrams, albeit with a reduction from training results.

Comparison Analysis: The analysis reveals a decline in performance between training (Table 3) and testing (Table 4), highlighting areas where the model may require further improvements to maintain effectiveness across different datasets. This comparison underscores the challenges in sustaining model performance when transitioning from a controlled training environment to a more variable testing context.

Task Four: bi-LSTM

Table 5. ROUGE Scores for Training

Metric	F-measure	Precision	Recall
Rouge1	0.7740094	0.7274317	0.8346181
Rouge2	0.6426103	0.59303427	0.71076614

(Table 5), presents the ROUGE-1 Score with an F-measure of 0.7740094, indicating very good lexical matching and a significant improvement in unigram matching. The precision of 0.7274317 shows that a substantial majority of the generated words are relevant, while the recall of 0.8346181 signifies that the model covers most of the information contained in the reference texts comprehensively.

ROUGE-2 Score in (Table 5): The F-measure of 0.6426103 demonstrates the model's enhanced capability to identify bigrams or two-word sequences. The precision of 0.59303427 and the recall of 0.71076614 reflect considerable improvement in managing more complex text structures and accu-

rately capturing relevant bi-gram content from the references.

Table 6. ROUGE Scores for Testing

Metric	F-measure	Precision	Recall
Rouge1	0.6502421	0.61461663	0.6998735
Rouge2	0.4736586	0.4394902	0.52330333

(Table 6) shows the ROUGE-1 Score with an F-measure of 0.6502421, reflecting competent unigram matching with the reference translations, though with a reduction from training results. The precision of 0.61461663 and the recall of 0.6998735 demonstrate good, yet slightly diminished accuracy and coverage compared to the training phase.

ROUGE-2 Score in (Table 6): The F-measure of 0.4736586, though lower than ROUGE-1, indicates moderate capability of the model to capture bi-gram sequences. The precision of 0.4394902 and recall of 0.52330333 suggest progress in the model's ability to predict and capture complex bi-gram sequences but also highlight areas for potential improvement in sequence accuracy and completeness.

Comparison Analysis: When comparing Table 5 and Table 6, it is evident that there is a decline in performance from training to testing. The decrease from an F-measure of 0.7740094 to 0.6502421 in ROUGE-1 and from 0.6426103 to 0.4736586 in ROUGE-2 highlights this drop. Both precision and recall values exhibit a reduction, pointing to a decrease in both the accuracy and coverage of unigrams and bi-grams. This reduction in performance may be attributed to model adjustments, variability in test data, or other external factors impacting model efficacy in different testing conditions.

Task Five: GRU Attention Mechanism Between Encoder and Decoder

Table 7. ROUGE Scores for Training

Metric	F-measure	Precision	Recall
Rouge1	0.76076420	0.70994465	0.82803904
Rouge2	0.61894433	0.56646474	0.69266288

ROUGE-1 Results obtained from (Table 7): The F-measure of 0.76076420 demonstrates excellent lexical matching, indicating a high quality of translation. With a precision of 0.70994465, most unigrams produced by the model are relevant, while the recall of 0.82803904 shows comprehensive coverage of reference material.

ROUGE-2 Results from (Table 7): The F-measure of 0.61894433 reflects effective bi-gram matching, showcasing improved capacity to handle complex text structures. Precision at 0.56646474 and recall at 0.69266288 indicate that the model performs well in matching two-word sequences and enhancing sequential prediction.

Table 8. ROUGE Scores for Testing

Metric	F-measure	Precision	Recall
Rouge1	0.63591086	0.59678572	0.69113595
Rouge2	0.44621293	0.41107261	0.49830843

ROUGE-1 Results from (Table 8): The F-measure of 0.63591086 indicates competent unigram matching, though there is a reduction compared to training. Precision of 0.59678572 and recall of 0.69113595 reflect respectable accuracy and coverage, suggesting areas for improvement.

ROUGE-2 Results from (Table 8): The F-measure of 0.44621293 shows moderate success in bi-gram matching. Precision at 0.41107261 and recall at 0.49830843 highlight the challenges in producing accurate and complete bi-gram sequences.

Comparison Analysis of ROUGE Scores for both Table 7 and 8: A notable decline in performance from training to testing is shown by the comparison. In particular, the F-measure for ROUGE-1 shrank in testing from 0.63591086 to 0.76076420 in training, suggesting less efficient performance with unknown data. In a similar vein, the ROUGE-2 F-measure decreased from 0.61894433 to 0.44621293, highlighting the model's limited ability to generalize predictions for bi-gram sequences.

These variations point to either the need for better generalization techniques or a possibility that overfitting of the model during training is affecting performance consistency between datasets.

Task Six: Transformer Encoder

Table 9. ROUGE Scores For Training

Metric	F-measure	Precision	Recall
Rouge1	0.19680132	0.29680830	0.15325800
Rouge2	0.11689621	0.20234284	0.08700258

The model's capacity to catch unigram overlaps with reference translations is rather low, as indicated by the F-measure of 0.19680132 in the ROUGE-1 Score from Table 9. While the recall of 0.15325800 reveals that the model covers just a small percentage of the pertinent content in the reference texts, the precision of 0.29680830 suggests that approximately 30% of the generated words are relevant.

An F-measure of 0.11689621 is displayed in the ROUGE-2 Score from Table 9, indicating a moderate improvement in the model's capacity to recognize precise bigram sequences. While the precision of 0.20234284 is a promising indication of the model's growing ability to properly predict sequences, the recall of 0.08700258 is poor, indicating limited coverage.

Table 10. ROUGE Scores For Testing

Metric	F-measure	Precision	Recall
Rouge1	0.20051461	0.30030555	0.15630783
Rouge2	0.11948121	0.20536883	0.08882900

An F-measure of 0.20051461 is shown by the ROUGE-1 Score from (Table 10), indicating a marginal improvement in the model's efficacy in unigram matching when compared to training. The recall of 0.15630783 and precision of 0.30030555, however, still show difficulties in reaching greater coverage and accuracy of pertinent unigram material.

With an F-measure of 0.11948121, the ROUGE-2 Score from Table 10 indicates a slight improvement in the model's capacity to match bi-grams. In spite of this, the precision of 0.20536883 and recall of 0.08882900 demonstrate the persistent difficulties in generating and accurately identifying word pairs that follow one another.

Comparison Analysis: A comparison of Training (Table 9) and Testing (Table 10) shows that throughout testing, there were only slight gains in ROUGE-1 and ROUGE-2 scores. The ROUGE-1 and ROUGE-2 F-measures, respectively, increased from 0.19680132 to 0.20051461, and from 0.11689621 to 0.11948121. These little improvements indicate that the model's ability to translate bigrams and unigrams may have improved. However, the overall performance is still low, emphasizing the necessity of further improvements in model construction and training to more accurately capture and recreate the target language's complexity.

3. Conclusion

The experiments conducted with the seq2seq model, using a consistent learning rate of 0.01 and SGD as the optimizer across 5 epochs, reveal significant insights into the model's performance and areas for improvement. While the model demonstrated some capacity to handle unigrams and bigrams, the overall ROUGE scores remained low, indicating room for substantial optimization. The slight improvements in testing scores compared to training suggest that the model can marginally adapt to new data, yet the overall effectiveness in capturing the complexity of language translation is limited. These outcomes underscore the need for further refinement of the model parameters and training process to enhance translation accuracy and reliability. Continued adjustments and more extended training may help improve the model's understanding of language nuances and its ability to generalize from training to real-world applications.

[1]Nanyang Technological University (NTU), Singapore [2]School of Computer Science and Engineering (SCSE) [3]Master of Science in Artificial Intelligence. Correspondence to: Ntambara Etienne <ntam0001@e.ntu.edu.sg>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Zhang, H. AI6102: Machine Learning Methodologies & Applications - L9: Evaluation & Density Estimation. Lecture presented in AI6102 course at Nanyang Technological University, 4 2024. Available online: https://mreallab.github.io/.