Machine Translation

Seetu Agarwal

University of Washington, Bothell
seetu5@uw.edu

Abstract

Natural language processing is gaining wide popularity in improving human-computer interaction.One of the prominent application of natural language processing is Machine Translation.It is the ability of machine to translate text from one language to another.Google translate is one of the popular tool widely used to convert text from one language to another. Currently it supports 108 languages. The aim of this research work is to convert the text in English to french. Various deep learning models like Recurrent Neural Network (RNN) , RNN with embedding, Bidirectional RNN and various encoder-decoder models will be explored on the WMT-2014 dataset.The models will be compared based on their accuracy in converting the text to target language.Bi-Lingual Evaluation under Study (BLEU) score will be compared which is the standard used for evaluation of machine translation systems.

Keywords: Natural Language Processing (NLP), Machine Translation (MT), Recurrent neural network (RNN), Encoder-Decoder models

1 Introduction

Machine Translation is the ability of machines to transform text from one language to another. It is the most prominent field in natural language processing which aims to improve human-computer interaction. At times when we have international meetings and conferences or when we visit different countries, a tool that can convert the audio into the text or audio that we can easily understand is really very helpful.Often we came across a video which is in language which we can't understand. At this time, subtitles in a language which can be easily understood is necessary in understanding the content of the video.The machine translation tools will remove the language barrier and makes life easy in other countries.Section 2 describes all the related and existing work in the field of machine translation as well as on the dataset WMT-2014. Section 3 describes the architecture of this research work and focuses on the main steps of the pipeline. Section 4 describes about the BLEU score which is the standard used for evaluating machine translation systems.Section 5 states the dataset used for this work. Section 6 discusses some important data cleaning and pre-processing steps.Section 7,8,9,10,11 constructs various Recurrent Neural Network models and their variants with the model evaluation and translation result.Web Application is later developed and Conclusion and future work is noted down.

Table 1. Categories of RBMT Systems

Type	**Description**
Direct Systems [1]	Input sentences are mapped directly to the output sentences
Transfer RBMT Systems [1]	These use morphological and syntactic analysis to translate sentences
Interlingual RBMT Systems [1]	Input sentence is transformed to an abstract representation

2 Exiting Related Works

In 1970's Rule Based Machine Translation (RBMT) was the major focus for the researchers. [1] In this method source text is converted into language-free conceptual representation. Information that was implicit in the source text was augmented to this representation.Finally the augmented representation is converted into target language. For the languages which have different structures (English-Japan),this technique is complicated.Table 1 describes various categories of RBMT Systems.

Statistical Machine Translation (SMT) is introduced later. Suppose S is a sentence in the source language and T is the translation in target language. This system assigns each (S,T) sentence pair the probability P(T—S) which is the probability that sentence T is the translated equivalent in the target language of the sentence S in the source language.These systems are categorized into two categories Word-based SMT [1] and phrase-based SMT [1]. In the word-based SMT, the source text is partitioned into a set of fixed locations. Glossary and contextual information is used to select the corresponding set of fixed locations in the target language.The words of the target fixed location are arranged into a sequence that forms the target language. In phrase-based SMT translation system, phase translation probability is defined to map phrases in source language to phrases in target languages.

Neural machine translation which uses artificial neural networks forms the backbone of the most popular tool in machine translation that is google translate. Google translate is the most popular tool in machine translation which currently supports 108 languages. It internally uses multiple encoder and decoder layers in Long Short Term Memory (LSTM) network and have been able to achieve BLEU score of 39. [2] For WMT-2014

Table 2. Methods on WMT-2014 English-French dataset

Model	Description	BLEU score
Transformer [3]	Very deep transformers for neural machine translation - 60 encoder layers and 12 decoder layers	43.8
ConvS2S [3]	Convolutional sequence to sequence learning	41.3
CSLM + RNN + WP [3]	Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation	34.54
LSTM [3]	Sequence-to-Sequence Learning	34.8

English-French dataset, Table 2 describes some existing researches:

3 Architecture

Below is the initial design of this research work:

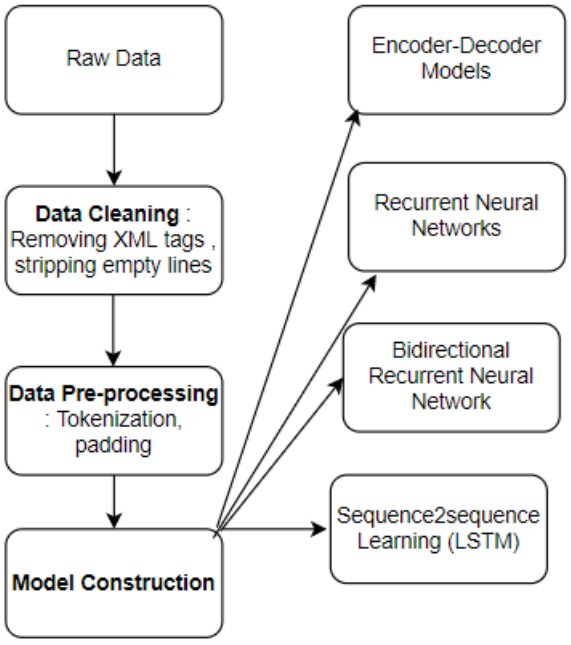


Figure 1 : System Architecture

Dataset is taken from WMT-2014 [4]. Initial phase is data cleaning and data pre-processing in the design. In the cleaning phase all the lines starting with xml tags are removed and empty lines are stripped. Pre-processing mainly includes tokenization and padding. In the model construction phase various models will be explored like encoder-decoder models, recurrent neural networks (RNN) and its variants and sequence to sequence learning models. [5]For evaluating BLEU score is considered for all the explored models.BLEU scores aboye 30 represents understandable translations and above 50 represents fluent translations.

4 BLEU Score

Bi-Lingual Evaluation under Study (BLEU) is the standard used for evaluation of machine translation systems.The correlation between a machine's output and that of a human is referred to as quality.[6]The primary premise behind BLEU is that "the closer a machine translation is to a competent human translation, the better it is."BLEU was one of the first metrics to claim a strong correlate with human quality judgments, and it is still one of the most used automated and low-cost measures today.Individual translated segments—generally sentences—are scored by comparing them to a set of high-quality reference translations.These scores are then averaged over the entire corpus to determine the overall quality of the translation.Intelligibility and grammatical accuracy are not considered.

Corpus blue Api of Natural langugae toolkit is used to calculate the BLEU Score. For the dataset , firt model.predict funciton is called to find out the predicted translations. Predicted as well as actual translations are given as input to this API. BLEU scores above 30 respresents understandable translations and above 50 represents fluent translations.

1-gram Blue uses same weights, that is (1,0,00).The 1-2-gram weights assign a 0.5 to each of 1-gram and 2-gram that is (0.5,0.5,0,0). The 1-3-gram weights are 0.33 for each of the 1, 2 and 3-gram score , that is (0.3,0.3,0.3.0).

5 Dataset Description

Dataset is taken from WMT 2014 [4].It contains collection of datasets used in the workshop for on statistical machine translation for various tasks medical text translation task, news translation task etc. It contains textual data as well as audio data . From WMT 2014 , Europarl dataset is considered which contains the preceedings of europian parliament.The textual data is of 2.5 GB and in various pair of languages like German-English, Danish-English, Dutch-English. The dataset in consideration is French-English which is approximately 617 MB in size. The dataset contains 50263459 English words and 52562231 French words. Samples from the dataset :

Text in English:

new jersey is sometimes quiet during autumn , and it is snowy in april .

Text in French:

new jersey est parfois calme pendant l' automne , et il est neigeux en avril .

6 Data Preprocesssing

Empty lines and their correspondences are stripped from the dataset.The lines that start with XML-Tags(starting with ”¡”) are removed.

6.1 Tokenization

is done next.Neural Networks did not understand ASCII character encodings. Therefore text like “Apple” , “Cat” etc needs to be converted into numbers as Neural Network is just a sequence of addition and multiplication operations. Either each character can be assigned a number or each word can be assigned a number .

```
{'the': 1, 'quick': 2, 'a': 3, 'brown': 4, 'fox': 5, 'jumps': 6, 'over': 7,
'lazy': 8, 'dog': 9, 'by': 10, 'jove': 11, 'my': 12, 'study': 13, 'of': 14,
'lexicography': 15, 'won': 16, 'prize': 17, 'this': 18, 'is': 19, 'short': 20,
'sentence': 21}

Sequence 1 in x
  Input:  The quick brown fox jumps over the lazy dog .
  Output: [1, 2, 4, 5, 6, 7, 1, 8, 9]
Sequence 2 in x
  Input:  By Jove , my quick study of lexicography won a prize .
  Output: [10, 11, 12, 2, 13, 14, 15, 16, 3, 17]
Sequence 3 in x
  Input:  This is a short sentence .
  Output: [18, 19, 3, 20, 21]
```

Figure 2 : Tokenization Output

6.2 Padding

Tokenization is followed by padding.Sentences are dynamic in length , when aggregating word id’s together padding can be added to make the sentences of equal length

```
Sequence 1 in x
  Input:  [1 2 4 5 6 7 1 8 9]
  Output: [1 2 4 5 6 7 1 8 9 0]
Sequence 2 in x
  Input:  [10 11 12  2 13 14 15 16  3 17]
  Output: [10 11 12  2 13 14 15 16  3 17]
Sequence 3 in x
  Input:  [18 19  3 20 21]
  Output: [18 19  3 20 21  0  0  0  0  0]
```

Figure 3 : Padding Output

7 Recurrent Neural Networks (RNN)

In Recurrent Neural Networks, the output depends on previous inputs and computations.They are recurrent because of their repetitive nature on the every element of the input sequence which is a variable length source sequence. Output unit is a variable length target sequence.It consists of hidden state h which is updated at each time t. Recurrent neural networks learns the probability distribution over a sequence. The output distribution (Softmax layer) size is equal to the size of the vocabulary V at every unit.Input layer is used where input is reshaped to work with basic RNN.Gated Recurrent Unit (GRU) is used.Time Distributed layer is used in which a layer is applied to each temporal slice of an input using this wrapper. Softmax Activation function is used. Model is compiled with Adam optimizer and loss as sparse categorical crossentropy. Fit function is called to train the model and the final output for the validation sample is fed to logits to text function to convert the output from numbers to words.

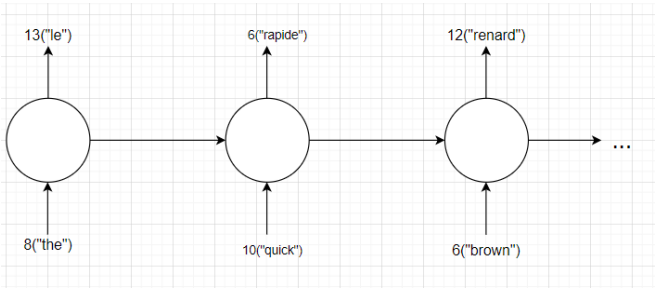


Figure 3 :System Architecture:RNN

7.1 Translation Result

```
# Print prediction(s)
print("Prediction:")
print(logits_to_text(simple_rnn_model.predict(tmp_x[:1])[0], french_tokenizer))

print("\nCorrect Translation:")
print(french_sentences[:1])

print("\nOriginal text:")
print(english_sentences[:1])
```

```
Prediction:
new jersey est parfois calme en mois et il et il en agréable <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>

Correct Translation:
["new jersey est parfois calme pendant l' automne , et il est neigeux en avril ."]

Original text:
['new jersey is sometimes quiet during autumn , and it is snowy in april .']
```

Figure 4: Translation Result with RNN

For the sample text, translation is correct up to the ”quiet” keyword, after that some words are incorrectly translated.

7.2 Evaluation

BLEU Score :The 1-gram BLEU score for validation dataset for RNN is 0.20. The 1-2 gram , 1-3 gram and 1-4 gram BLEU score is calculated and the results are plotted in Figure 5.Model is trained with varying number of epochs from 10 to 50, validation loss and accuracy are recorded in Table 3. Graphs are plotted for visualization.

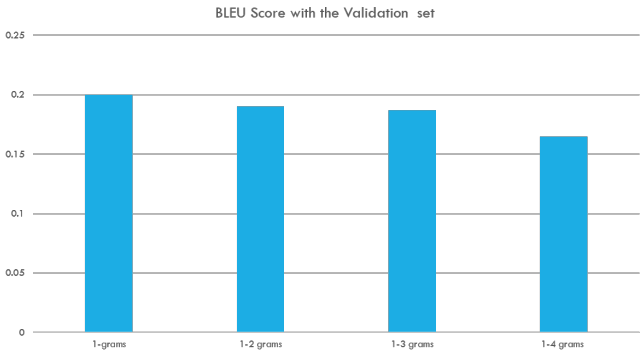


Figure 5 : BLEU Score with RNN

Table 3. Variation in loss and accuracy with number of epochs

Number Of Epochs	Validation Loss	Validation Accuracy
10	1.4477	0.6184
20	1.1885	0.6539
30	1.0956	0.6667
40	1.0217	0.6854
50	0.9957	0.6940

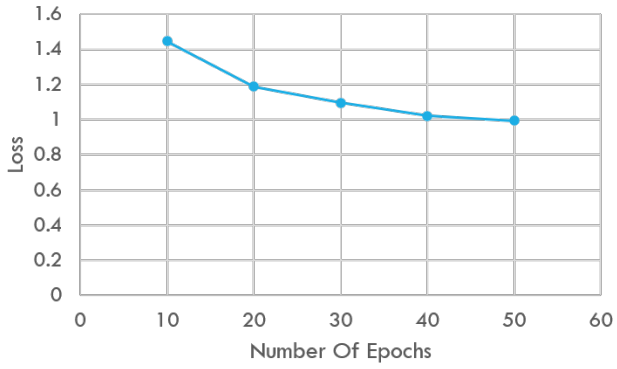


Figure 6: Variation in validation loss with increase in number of epochs

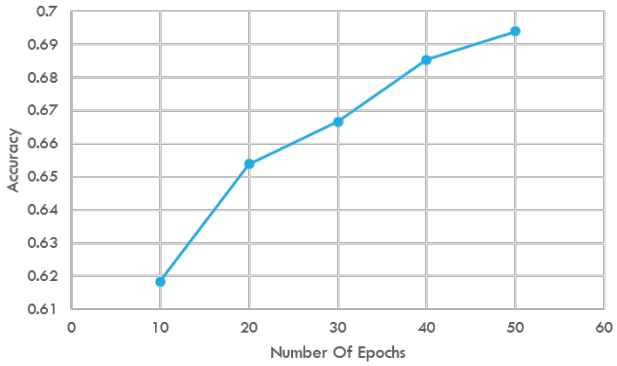


Figure 7: Variation in validation accuracy with increase in number of epochs

8 Recurrent Neural Networks (RNN) with Word Embedding

Individual words are represented as real-valued vectors in a predetermined vector space in a process known as word embedding. Words with the same meaning are represented similarly here.Gated Recurring Unit (GRU) is used.Embedding layer of keras is added. It converts positive integers (indexes) into fixed-size dense vectors.Time-Distributed layer is added .This layer can be applied to each temporal slice of an input using this wrapper

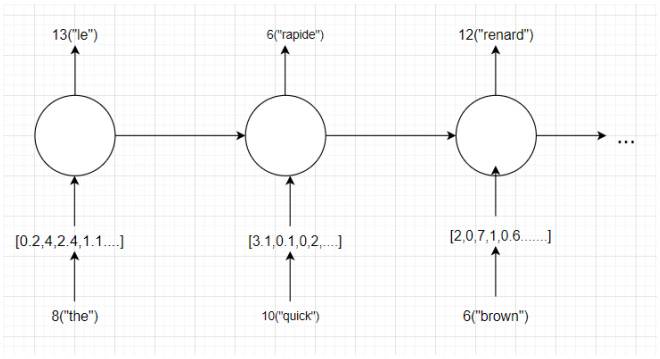
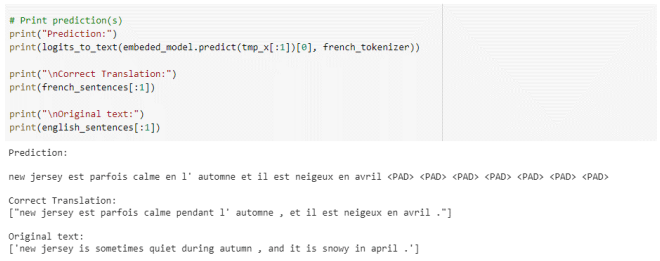


Figure 8:System Architecture:RNN with word embedding

8.1 Translation Result



```
# Print prediction(s)
print("Prediction:")
print(logits_to_text(embeded_model.predict(tmp_x[:1])[0], french_tokenizer))

print("\nCorrect Translation:")
print(french_sentences[:1])

print("\nOriginal text:")
print(english_sentences[:1])
```

```
Prediction:

new jersey est parfois calme en l' automne et il est neigeux en avril <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>

Correct Translation:
["new jersey est parfois calme pendant l' automne , et il est neigeux en avril ."]

Original text:
['new jersey is sometimes quiet during autumn , and it is snowy in april .']
```

Figure 9: Translation Result : RNN with word embedding

Translations are correct after word embedding.

8.2 Evaluation

BLEU Score :The 1-gram BLEU score for validation dataset for RNN with word embedding is 0.27.The 1-2 gram , 1-3 gram and 1-4 gram BLEU score is calculated and the results are plotted in Figure 10.Model is trained with varying number of epochs from 10 to 50, validation loss and accuracy are recorded in Table 4. Graphs are plotted for visualization.

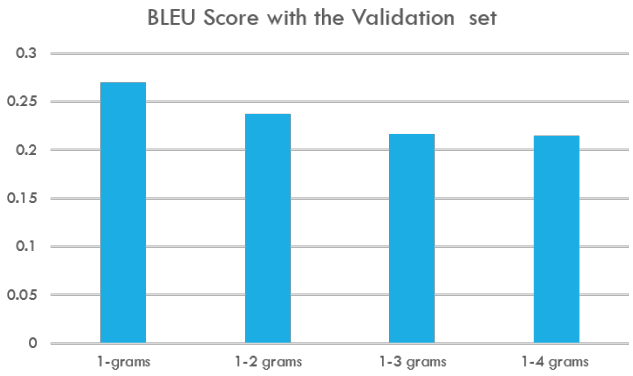


Figure 10: BLEU Score :RNN with word embedding

Table 4. RNN with Word Embedding : Variation in loss and accuracy with number of epochs

Number Of Epochs	Validation Loss	Validation Accuracy
10	0.5514	0.7927
20	0.3526	0.8164
30	0.2770	0.8168
40	0.2656	0.8279
50	0.2597	0.8300

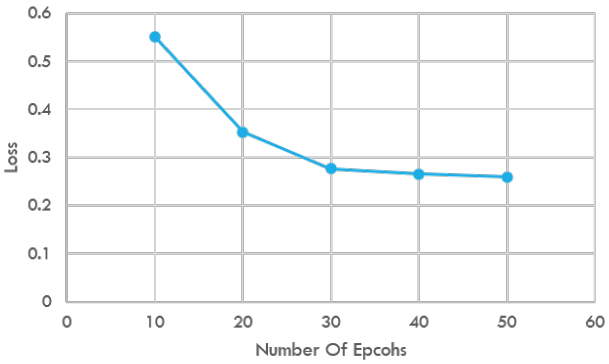


Figure 11: Variation in loss with increase in number of epochs

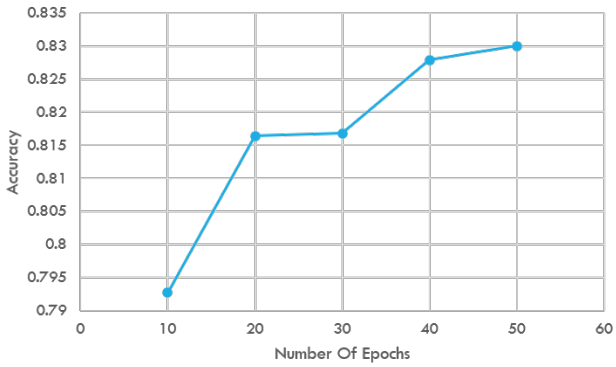


Figure 12: Variation in Accuracy with increase in number of epochs

There is a slight increase in Accuracy from 20 to 30 epochs.There is a significant decrease in loss from 10 to 50 epochs.

9 Bidirectional Recurrent Neural Networks

This model has a look-ahead feature. Instead of only beginning an RNN from the first token and running it forward, we may start another one from the final token and run it backwards.It add a hidden layer that passes information in a backward direction to process information more flexibly.

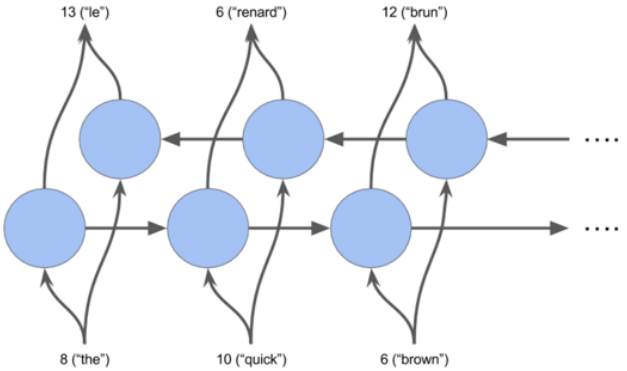


Figure 13: Bidirectional RNN System Architecture

9.1 Translation Result

```
# Print prediction(s)
print("Prediction:")
print(logits_to_text(bidi_model.predict(tmp_x[:1])[0], french_tokenizer))

print("\nCorrect Translation:")
print(french_sentences[:1])

print("\nOriginal text:")
print(english_sentences[:1])
```

```
Prediction:
new jersey est parfois occupé en printemps mais il gèle agréable l' mai <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>

Correct Translation:
["new jersey est parfois calme pendant l' automne , et il est neigeux en avril ."]

Original text:
['new jersey is sometimes quiet during autumn , and it is snowy in april .']
```

Figure 14: Translation Result : Bidirectional RNN

Translation results are little bit inaccurate with bidirectional RNN.

9.2 Evaluation

BLEU Score :The 1-gram BLEU score for validation dataset for Bidirectional RNN is 0.29.The 1-2 gram , 1-3 gram and 1-4 gram BLEU score is calculated and the results are plotted in Figure 15.Model is trained with varying number of epochs from 10 to 50, validation loss and accuracy are recorded in Table 5. Graphs are plotted for visualization.

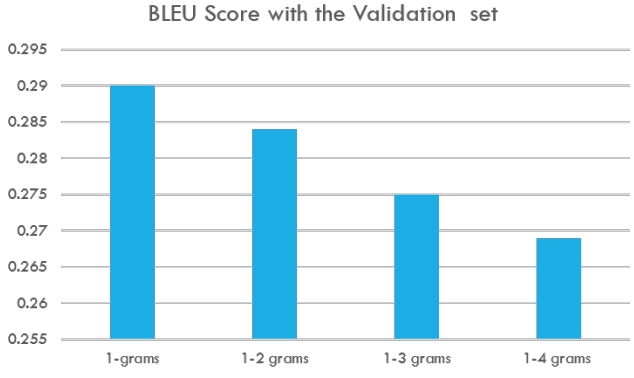


Figure 15: BLEU Score : Bidirectional RNN

Table 5. Bidirectional RNN : Variation in loss and accuracy with number of epochs

Number Of Epochs	Validation Loss	Validation Accuracy
10	1.1006	0.6720
20	0.8737	0.7105
30	0.7411	0.7557
40	0.6202	0.8071
50	0.5555	0.8266

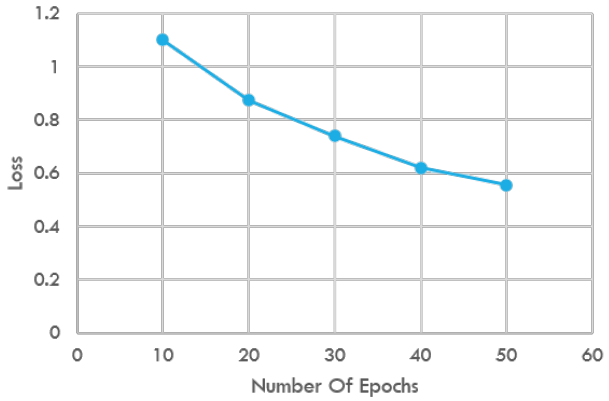


Figure 16: Variation in loss with increase in number of epochs

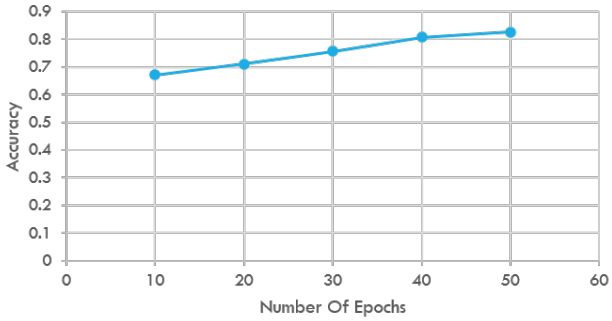


Figure 17: Variation in Accuracy with increase in number of epochs

10 Bidirectional Recurrent Neural Networks with Word Embeddings

This model combines RNN with word embedding to Bidirectional RNN.

Table 6. Bidirectional RNN with word embedding : Variation in loss and accuracy with number of epochs

Number Of Epochs	Validation Loss	Validation Accuracy
10	0.912	0.701
20	0.836	0.732
30	0.789	0.768
40	0.623	0.820
50	0.563	0.846

10.1 Translation Result: Bidirectional RNN with word embedding

```
# Print prediction(s)
print("Prediction:")
print(logits_to_text(bidi_model.predict(tmp_x[:1])[0], french_tokenizer))

print("\nCorrect Translation:")
print(french_sentences[:1])

print("\nOriginal text:")
print(english_sentences[:1])
```

```
Prediction:
new jersey est parfois calme pendant l' automne et il est neigeux en avril <PAD>

Correct Translation:
["new jersey est parfois calme pendant l' automne , et il est neigeux en avril ."]

Original text:
['new jersey is sometimes quiet during autumn , and it is snowy in april .']
```

Figure 18: Translation Result : Bidirectional RNN with word embedding

For the sample text , translation is completely accurate.

10.2 Evaluation

BLEU Score :The 1-gram BLEU score for validation dataset for Bidirectional RNN with word embedding is 0.32.The 1-2 gram , 1-3 gram and 1-4 gram BLEU score is calculated and the results are plotted in Figure 19.Model is trained with varying number of epochs from 10 to 50, validation loss and accuracy are recorded in Table 6. Graphs are plotted for visualization.

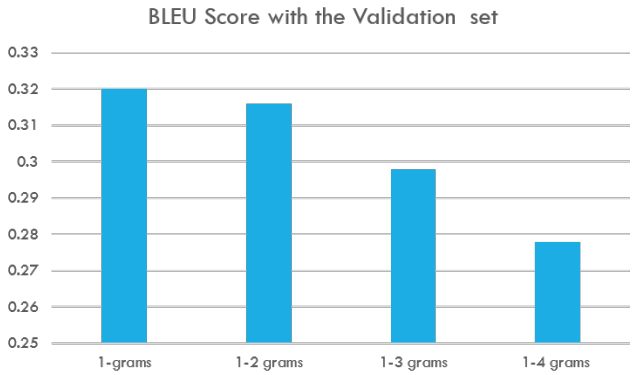


Figure 19: BLEU Score : Bidirectional RNN with word embedding

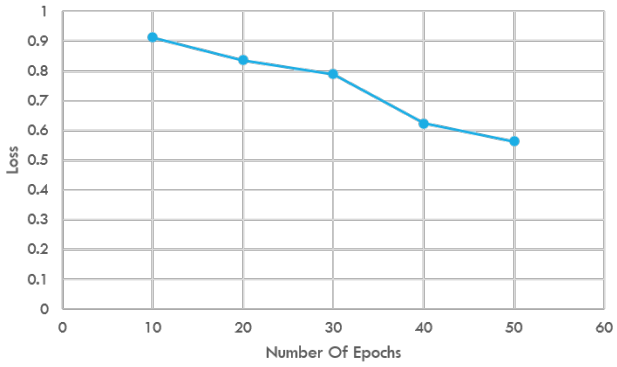


Figure 20: Variation in validation loss with increase in number of epochs

With increase in number of epochs , validation loss decreases.

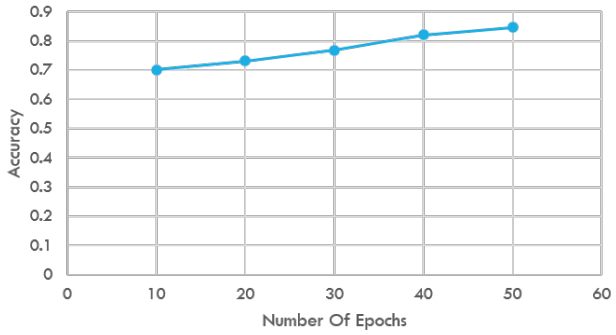


Figure 21: Variation in validation accuracy with increase in number of epochs

With increase in epochs, validation accuracy increases.

11 Encoder – Decoder Model

There are two recurrent neural networks in the Encoder-Decoder model: Encoder and Decoder. The encoder compiles the data into a state variable, also known as a context variable. The output sequence is then created when the context has been deciphered.

Because they are both recurrent, both the encoder and the decoder include loops that process each portion of the sequence at different time steps.At each time step, the encoder reads the input word and transforms the hidden state. Following that, the secret state is passed on to the next time step. Hidden state and word from sequence are the two inputs for each time sequence. The following word in the input sequence is what the encoder looks for. The preceding word in the output sequence is used by the decoder.

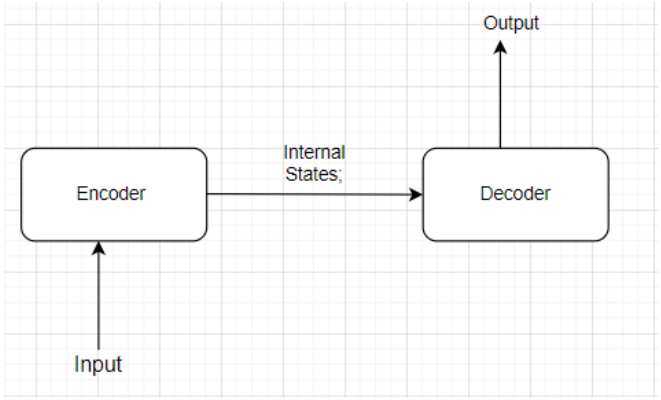


Figure 22: System Architecture : Encoder- Decoder Model

11.1 Translation Result

```
# Print prediction(s)
print("Prediction:")
print(logits_to_text(encodeco_model.predict(tmp_x[:1])[0], french_tokenizer))

print("\nCorrect Translation:")
print(french_sentences[:1])

print("\nOriginal text:")
print(english_sentences[:1])
```

```
Prediction:

new jersey est généralement agréable en mois et il est est en en <PAD> <PAD> <

Correct Translation:
["new jersey est parfois calme pendant l' automne , et il est neigeux en avril ."]

Original text:
['new jersey is sometimes quiet during autumn , and it is snowy in april .']
```

Figure 23: Translation Result : Encoder- Decoder Model
For the sample text , translation is completely accurate.

11.2 Evaluation

BLEU Score :The 1-gram BLEU score for validation dataset for encoder decoder model is 0.29.The 1-2 gram , 1-3 gram and 1-4 gram BLEU score is calculated and the results are plotted in Figure 24.Model is trained with varying number of epochs from 10 to 50, validation loss and accuracy are recorded in Table 7. Graphs are plotted for visualization.

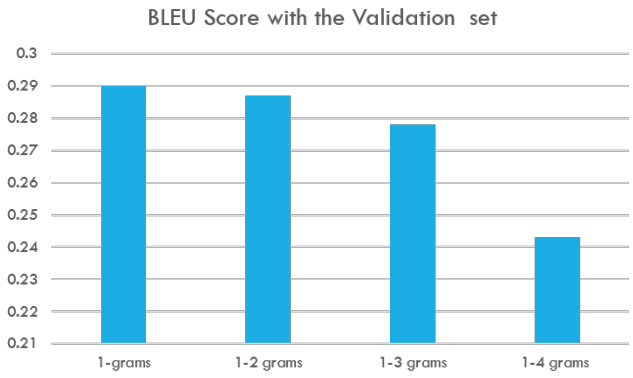


Figure 24: BLEU Score : Encoder- Decoder Model

Table 7. Encoder-Decoder : Variation in loss and accuracy with number of epochs

Number Of Epochs	Validation Loss	Validation Accuracy
10	0.996	0.613
20	0.882	0.714
30	0.689	0.759
40	0.567	0.798
50	0.4173	0.8310

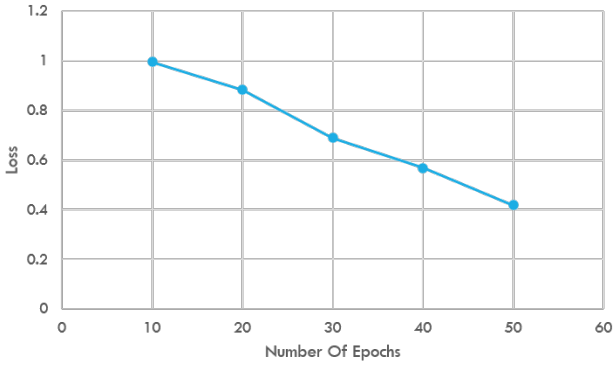


Figure 25: Variation in validation loss with increase in number of epochs
With increase in number of epochs , validation loss decreases.

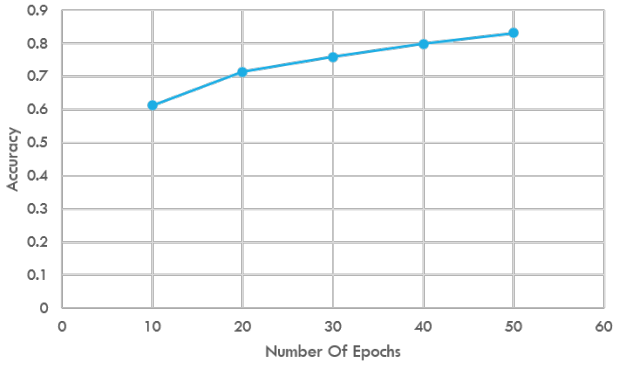


Figure 26: Variation in validation accuracy with increase in number of epochs
With increase in epochs, validation accuracy increases.

12 Web Application

A web application is developed in which user text input in English language is required.When user clicks on translate button, internally translate function is called which converts the text in English to text in French with the help of the trained model. The front-end is developed with the help of HTML, CSS and JavaScript web technologies. Flask web framework is used in the backend for the development of the application.

Figure 27 :Snapshot of the Web Application

The users inserts the text ”thank you for your time”. The equivalent french translation is diplayed on clicking the translate button.

Figure 28: Translated output displayed in the Web Application

13 Conclusion

Of all the explored models, RNN (word embedding + bidirectional) has highest translation accuracy of 0.84 and BLEU score of 0.32 With increase in number of epochs , slight increase in accuracy and decrease in loss was observed. RNN models seems to be faster in training than encoder-decoder model. Word embedding is the better representation of the word. In comparison with Basic RNN, RNN with word embedding performs better both in terms of accuracy and BLEU Score. Bidirectional RNN performs similar to Encoder-Decoder models in terms of validation accuracy and BLEU Score.Bidirectional RNN with word embedding performs slightly poor than baseline methods , but was the best among all the tried combination of models.Thus it was selected to use in the web application.

14 Future Work

- Currently Web application supports only the conversion from English to French language.Support for other languages like German , Spanish , Hindi can be provided to increase the functionality of the application.
- Various other deep learning models like transformers , LSTM , sequence to sequence learning models can be explored.Phrase representation of the word can be considered.
- Application can be integrated with video or image captioning or generating subtitles in video in target language.
- Application can further be enhanced to convert text obtained in different languages to speech so that along with

subtitles user can also listen the video in different language. Text-to-Speech models can be explored.

References

[1] A. Garg and M. Agarwal, "Machine translation: A literature review," 2018.

[2] K. Stevens, G. Kurian, N. Patil, Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[3] "Machine translation on wmt-2014 english-french available on https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-french."

[4] "Wmt-2014 dataset availaible on http://www.statmt.org."

[5] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015.

[6] T. W. Kishore Papineni, Salim Roukos and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation."