Natural Language Translations Using Deep Learning

Dennis Dang
Yan Chen
CS271

December 3, 2019

1 Introduction

In this project, we aim to produce a machine learning model capable of producing translations from language to another. Specifically, we will be producing a model that takes as an input phrases in French and translates these phrases into English. We perform multiple experiments by adjusting different model parameters to determine which sets of parameters play a dominant role in producing accurate translations. We compare the results of each of our models using a BLEU score that quantifies the accuracy of translations. We also directly the word by word output with Google translate.

2 Background

2.1 Ruled Vs. Statistical Based Machine Translation

Rule based machine translating heavily relies on human input for specifying how a certain language is constructed [1]. For this, reliance upon individuals such as linguists with deep knowledge of how a certain language is constructed are often employed. In this rule based system, morphology (how words are formed), syntax and semantics are used to generate a translation from the target to a source language. At the start of translation, analysis on the input sentence is done to extract morphology, parts of speech syntax and semantics. In the next step, word translation is done where each input word is replaced by the corresponding word in the target language using a dictionary. Finally in the last step, the words of the target language are arranged together using the syntactic and grammar rules from both the source to construct an accurate translation. While this model has been shown to work, heavy reliance on human input to come up with every single linguistic rule in every source to target language is incredibly cumbersome.

Statistical machine translating, on the other-hand, is a more natural approach in a machine learning sense. This approach involves using a body of text called a *corpora* containing two languages that are parallel aligned where every phrase in one language has an equivalent translation in the other language. The problem of a translation then becomes a mathematical one:

$$P(e|f) = \frac{P(f|e)P(e)}{P(f)} \tag{1}$$

Where $P(e|f)$ is the probability that an English sentence is the correct translation from the source French sentence, $P(f|e)$ is the probability that the French and English text from the corpus are direct translations of each other and $P(e)$ and $P(f)$ are the probabilities of finding the input French sentence to be translated and the output translated English sentence within the corpus.

2.2 Sequence to Sequence Learning

Sequence to sequence learning is a paradigm in deep learning that sims to map a fixed-length input to a fixed-length output where the length of the input and output may differ [2]. This framework of deep

learning is absolutely essential in machine translating since the lengths of the source and target languages will generally differ. To accomplish this, sequence to sequence learning relies on the encoder-decoder model. An encoder-decoder model consists of using 2 recurrent neural networks (RNN) [3]. The encoding layer takes the input sequence and generates an output vector called a "thought vector" that captures the meaning of the input sentence and expresses it as a vector of floating point values. These values are learned through training. The output of the encoding layer then gets fed into the decoder layer which generates an output sequence while reading from the thought vector for every step in the output time steps. During training, the encoder learn how to better capture the meaning of the input sequence while the decoder learns how to make predictions using the output sequence of the decoder.Figure 1 taken from [4] illustrates the architecture of the encoder-decoder framework.

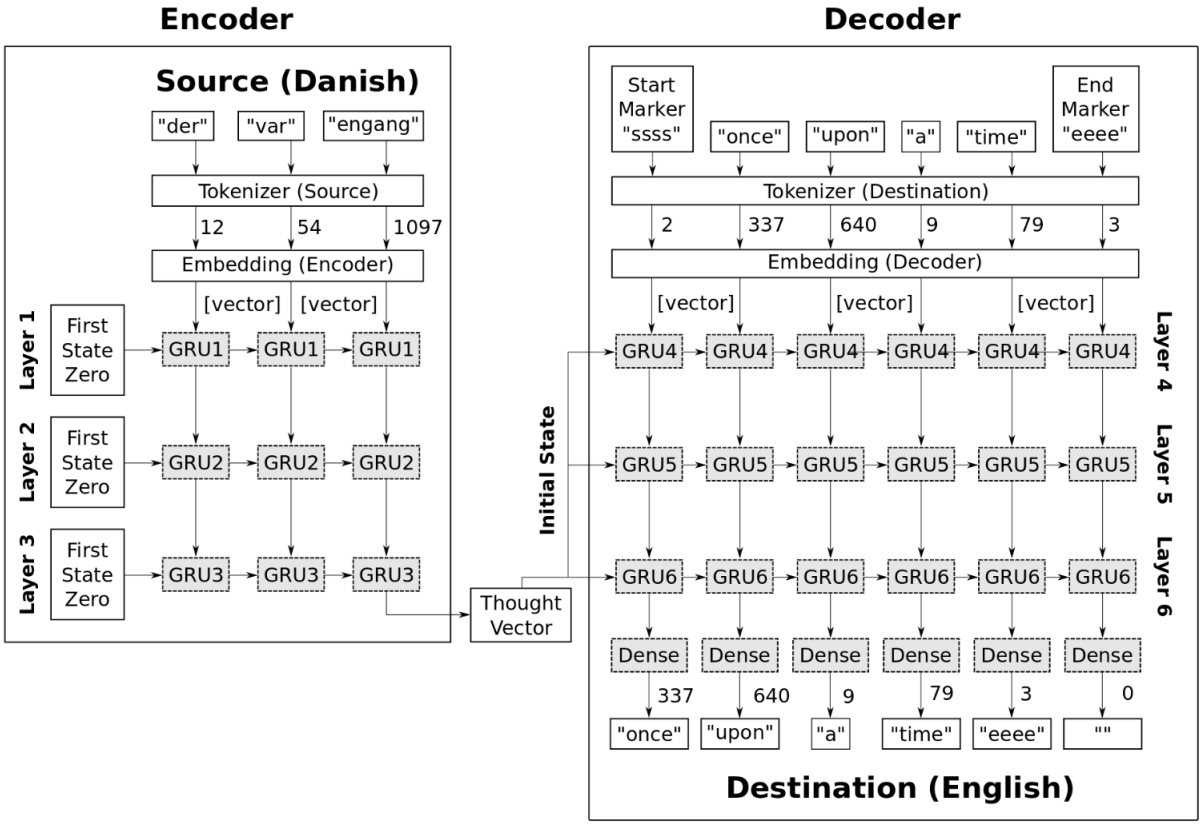

Figure 1: An illustration of an encoder-decoder framework architecture.

2.3 BLEU Scores

Evaluating how well a translation is given an input sentence can be a subjective matter. With the hundreds of thousands of words that comprise any single language, there can be many ways of expressing the same meaning. Take, for example, the sentence: *"he ate an apple because he was hungry"*. The same sentence can be rewritten as *"he ate an apple because he was famished"* or *"his ravenous appetite compelled him to consume an apple for nourishment"*. All of these sentences capture the same meaning. This makes the job of finding the translation of a sentence difficult because with so many ways of expressing sentences, how do we know which sentence is the most accurate translation? The sentence *"you should sit down"* can be translated to *"tu devrais t'asseoir"* or *"vous devriez vous asseoir"* in French. While it maybe simple for a French speaker to determine that these two translations are accurate, a machine cannot. Nonetheless, we still need a way of quantifying the accuracy of translations in order to improve the model. Bilingual Evaluation Understudy (BLEU) scores achieve this goal [5]. BLEU scores are calculated using an algorithm where consecutive words of the generated translation is compared consecutive corresponding words of the reference translation. The

number of matches of words are then counted in a weighted fashion. The higher count, the higher the degree of similarity between the generated and reference translations. In calculating BLEU scores, intelligibility and grammatical correctness of the generated translation are not taken into account. These reasons along with others such as not considering meaning or sentence structure makes BLEU an imperfect way of evaluating translations. However, because it is still a popular metric, we will be using it to evaluate translations for this project.

3 Method

To train the model, a parallel corpus as obtained. We used a parallel corpus of the European Parliament Proceedings Parallel Corpus from 1996-2011 [6]. This database of files contains parallel translations from English to many European languages such as French, German, Spanish and Italian. For this project, we used the English to French parallel corpus containing over 2 million sentences.

Before any of the data can be fed into the model, preprocessing is necessary. Preprocessing the data involves cleaning the data whereby word with special characters such as accents are removed. After preprocessing, the words in each sentence of the dataset must be converted to a numerical representation for the model to work with. As such, every unique word that has not yet been encountered is mapped to an integer to create a sequence of integers for every sentence, as shown in the example below.

$$\texttt{"Mary had a little lamb"} \rightarrow [5, 31, 57, 6, 40]$$

This process is called tokenizing the input. Each sequence is padded with zeroes to a maximum length (and truncated if necessary) in order to have the same length. The next step is to reverse the sequence of tokens. Research from [2] suggests that reversing the tokens improves performance because the last words seen by the encoder more closely match the first words produced by the decoder improving the short-term memory.

The last step before the sequences can be fed into the model is embedding. Embedding is necessary for two reasons: Firstly, neural networks cannot work on just integers. These values must be converted to floating-point values. Secondly, embedding can significantly improve the performance of the model by reducing the dimensionality of the data and allowing similar words to be represented with similar values. If our source language has a vocabulary of 10 words, we can use embedding to represent each word with using a sequence of, say, 4 numbers. For example, the word `"Mary"` corresponding to integer 5 from the above example can be represented as [0.12, -0.56, -0.67, 1.19]. Continuing from this example:

$$\begin{aligned}\texttt{"Mary had a little lamb"} \rightarrow [5, 31, 57, 6, 40] \rightarrow &[[0.12, -0.56, -0.67, 1.19],\\ &[0.23, 1.55, 1.58, -0.48],\\ &[-1.64, 1.02, -0.51, 0.80],\\ &[1.66, -0.20, -0.07, 0.51],\\ &[-1.42, 0.71, 1.51, -1.26]]\end{aligned}$$

In addition to reducing the dimensionality of the input data, the embedding layer can also learn the semantics and closeness of words. For example, words such as `"Toyota"` and `"Honda"` hardly seem close in letter space but we know that each word is related to cars. Through embedding, the similarity of both words can be represented with vectors whose values are close or similar to each other which facilitates the process of finding accurate translations.

After embedding, these vectors can then be fed into the model. The translations that are obtained are then evaluated using the BLEU algorithm to obtain a numerical score. For this project, we used a NVidia 1070TI GPU to decrease training times.

4 Results

For this project, we performed six different experiments with the goal of finding which of these parameters provided the best overall translation based on the BLEU scores. Table 1 shows the parameters used for each experiment and Table 2 describes the meaning for each parameter. We scored each model by translating short, medium and long length sentences from French to English. Short sentences were chosen to be 10 words or less, medium legnth sentences were between 11-25 words while long sentences contained 26 or more words. We used sentences from the training model and from [7] which contain proceedings from the Canadian Parliament which are parallel aligned from English to French. We used 10 short, medium and long sentences from bot the training set and from [7] for a total of 60 sentences. For the sake of brevity, we only show two sentences for each length from both sets in the sample translation results sections below.

Experiment Number	`num_epochs`	`vocab_size`	`state_size`	`embedding_size`	`num_layer`	`reverse_seq`
1	10	10,000	512	128	3	True
2	10	10,000	512	128	3	**False**
3	10	10,000	512	128	**4**	True
4	10	10,000	**256**	128	3	True
5	10	**30,000**	512	128	3	True
6	**20**	10,000	512	128	3	True

Table 1: 5 models were trained using the following parameters.

Parameter Name	**Parameter Meaning**
`num_epochs`	Number of times to train the model on the same dataset
`vocab_size`	Number of words that the source and target languages will contain
`state_size`	The number of hidden layers in RNN cell
`embedding_size`	Size of that the vector that the embedding layer outputs
`num_layer`	The number of RNN layers for both the encoder and decoder, the depth of these layers
`reverse_seq`	Boolean used to determine if the input source language sequence should be reversed during tokenization

Table 2: The parameters used to train the model and their meanings.

The rationale behind each experiment are as follows:

- **Experiment 1:** This is the baseline model. Other models are trained based on minor parameter adjustments from this model.
- **Experiment 2:** It is claimed from [2] that reversing the input sequence allows for better translation. We wanted to test if this is really the case.
- **Experiment 3:** By increasing the depths of the encoder and decoder networks, we hope to allow the network to capture more information from the input and allow for better and more complicated translations.
- **Experiment 4:** We could not increase the state size to a value any larger than 512 since training the model would consume too much GPU memory. Thus, by halving this value, we anticipate seeing a loss in translation quality as the complexity of the networks decrease.

- **Experiment 5:** Expanding the vocabulary size allows the model to work with more letters and thus generate more complex translations.
- **Experiment 6:** Increasing the number of epochs gives the model more opportunities to learn the semantics of the languages from the dataset.

4.1 Using Sentences From the Training Data

Input Sentence	Êtes-vous satisfait du rythme de ces progrès
Actual Translation	Are you satisfied with this rate of progress
Experiment 1: Default	Are you satisfied with the progress made
Experiment 2: Reverse Input Off	Are you satisfied with the progress made
Experiment 3: 4 Layers	Are you satisfied with the progress made
Experiment 4: 256 State Size	Are you satisfied with the progress made in this regard
Experiment 5: 30,000 Vocab Size	Are you satisfied with the progress that has been made
Experiment 6: 20 Epochs	Are you satisfied with the pace of progress

Table 3: Test from training data, short length sentence 1.

Input Sentence	Sans approvisionnement, leur santé et leurs vies sont en danger
Actual Translation	Without supplies, their health and their lives can be at risk
Experiment 1: Default	Without their health and safety they are dangerous
Experiment 2: Reverse Input Off	Without their health and safety they are dangerous
Experiment 3: 4 Layers	Without their health and their lives they are threatened
Experiment 4: 256 State Size	Without health the health and their lives will be threatened
Experiment 5: 30,000 Vocab Size	Without health and their lives there is a risk
Experiment 6: 20 Epochs	Without their health and their lives are at risk

Table 4: Test from training data, short length sentence 2.

Input Sentence	Mais il s'agit effectivement d'un problème auquel nombre de citoyens européens sont confrontés chaque jour
Actual Translation	But it is a problem which is affecting many European citizens on a daily basis
Experiment 1: Default	But this is a problem that affects many european citizens every day
Experiment 2: Reverse Input Off	But this is a problem that affects many european citizens every day
Experiment 3: 4 Layers	But it is a problem which affects many european citizens
Experiment 4: 256 State Size	But this is indeed a problem that affects many european citizens
Experiment 5: 30,000 Vocab Size	But it is indeed a problem of many european citizens
Experiment 6: 20 Epochs	But this is a problem that many european citizens are facing every day

Table 5: Test from training data, medium length sentence 1.

Input Sentence	Monsieur le Président, la concurrence est l'âme et le moteur de la politique européenne en matière de marché intérieur
Actual Translation	Mr President, competition is at the heart of the European internal market policy and is also its driving force
Experiment 1: Default	Mr president i should like to begin by saying that we are not always aware of the events and the events
Experiment 2: Reverse Input Off	Mr president competition is a driving force and the european driving force for the internal market
Experiment 3: 4 Layers	Mr president competition is the european internal market and the construction of the european internal market
Experiment 4: 256 State Size	Mr president competition is the driving force behind the european monetary policy
Experiment 5: 30,000 Vocab Size	Mr president competition is the cornerstone of the internal market and the european policy on the internal market
Experiment 6: 20 Epochs	Mr president competition is and the european driving force behind the internal market

Table 6: Test from training data, medium length sentence 2.

Input Sentence	Je suppose que la question a trait à la question de savoir où la compétence de cette autorité commence et prend fin et où la compétence et l'autorité des agences de sécurité alimentaire au sein des États membres commence et prend fin
Actual Translation	I suspect that the question is focused on the issue of where the competence of the authority begins and ends and where the competence and authority of food safety agencies in Member States begin and end
Experiment 1: Default	I ask the authority to decide where authority authority is concerned and when the competence of the authority and the competence of the member states and the agencies authority is ultimately at stake
Experiment 2: Reverse Input Off	I ask the authority to decide where authority authority is concerned and when the competence of the authority and the competence of the member states and the agencies authority is ultimately at stake
Experiment 3: 4 Layers	I assume that the question of the authority which is being asked is the authority and the authority of the member states and hence the authority of the authority of the authority and agencies
Experiment 4: 256 State Size	I assume that the question of how this is to be taken is to be monitored by the authority and authority and when the authority and authority of member states are finally taken into account
Experiment 5: 30,000 Vocab Size	I would like to know that the issue is the competence of the authority where the competence and authority of the member states is beginning to be integrated into the food and health bodies
Experiment 6: 20 Epochs	I assume that the question arises as to the competence of the authority and the question of whether the authority and the food authority in the member states are now starting to be and

Table 7: Test from training data, long length sentence 1.

Input Sentence	Les dommages sociaux et économiques, dont on a déjà parlé ici aujourd'hui, en termes tant de perte d'emplois que de ressources marines et touristiques, sont d'une ampleur telle qu'ils justifient amplement une action décidée et marquante de la part des institutions communautaires
Actual Translation	The economic and social damage, which we have spoken about today, in terms of the loss of jobs and fishing and tourist resources, is so great that they fully justify decisive and thorough action on the part of the Community institutions
Experiment 1: Default	The economic and social spheres of the european institutions today has already spoken about a massive scale of employment and employment which are a very critical and demanding action and that they are both justified and justified in the european institution
Experiment 2: Reverse Input Off	The economic and social spheres of the european institutions today has already spoken about a massive scale of employment and employment which are a very critical and demanding action and that they are both justified and justified in the european institutions
Experiment 3: 4 Layers	The economic and social consequences that we have already achieved here today are being used as a result of the relocation of companies and tourism and tourism and the action plan and the commission and the european union
Experiment 4: 256 State Size	The economic and social consequences of which have already been mentioned today have been lost in a number of areas such as employment and tourism are often too high to justify their actions and to the community institutions
Experiment 5: 30,000 Vocab Size	The economic and social problems that have already been mentioned today are of the opinion that this is a major concern for the transport of resources and of the resources that are being paid out and which are very important and that the european institutions are making a few comments
Experiment 6: 20 Epochs	The economic and social damage that has been done here today is already a result of the loss of jobs and marine resources which is why it is so serious that the actions taken by the community and the institutions are justified

Table 8: Test from training data, long length sentence 2.

4.2 Using Never Before Seen Sentences

Input Sentence	Les problèmes sont toujours là
Actual Translation	The problems have not gone away
Experiment 1: Default	The problems are always being solved
Experiment 2: Reverse Input Off	The problems are always being solved
Experiment 3: 4 Layers	The problems are still unresolved
Experiment 4: 256 State Size	The problems are still unresolved
Experiment 5: 30,000 Vocab Size	The problems are still there
Experiment 6: 20 Epochs	The problems are still

Table 9: Test from new data, short length sentence 1.

Input Sentence	Nous souffrons des compressions budgétaires aux aéroports
Actual Translation	We suffer from the cuts to airports
Experiment 1: Default	We are spending budgetary airports
Experiment 2: Reverse Input Off	We are spending budgetary airports
Experiment 3: 4 Layers	We will be paying the budget for airports
Experiment 4: 256 State Size	We are asking for budgetary constraints to be made
Experiment 5: 30,000 Vocab Size	We are dealing with the budgetary burden on airports
Experiment 6: 20 Epochs	We are budget for airports

Table 10: Test from new data, short length sentence 2.

Input Sentence	Il devrait aider les entreprises à prendre notre ressource et en faire autre chose pour les gens de chez nous
Actual Translation	It should help businesses to take our resource and develop other products from it for the good of our own people
Experiment 1: Default	It should help our businesses to invest and make our own people more attractive
Experiment 2: Reverse Input Off	It should help our businesses to invest and make our own people more attractive
Experiment 3: 4 Layers	It should help businesses to invest in and to help people in our regions
Experiment 4: 256 State Size	It should help to help companies and to do so in other words to people in our own people
Experiment 5: 30,000 Vocab Size	It should help to make our own research and we do so in order to help people in our own interest
Experiment 6: 20 Epochs	It should help businesses to do something and to take them to other people

Table 11: Test from new data, medium length sentence 1.

Input Sentence	Le devoir de satisfaire à leurs besoins est d'une importance fondamentale pour les personnes handicapées, de même que pour des groupes comme les minorités religieuses
Actual Translation	The duty to accommodate is of vital importance to persons with disabilities as well as to groups such as religious minorities
Experiment 1: Default	It is important to have a fundamental understanding for people with disabilities and for religious groups as well as religious minorities
Experiment 2: Reverse Input Off	It is important to have a fundamental understanding for people with disabilities and for religious groups as well as religious minorities
Experiment 3: 4 Layers	The need for such a crucial body is essential for people with disabilities as well as for the freedom of movement
Experiment 4: 256 State Size	The obligation to respect the needs of the people is a fundamental issue for disabled people and as such as groups of minorities
Experiment 5: 30,000 Vocab Size	The need to meet their needs is of fundamental importance for their people as well as for their religious groups
Experiment 6: 20 Epochs	The duty to meet their needs for fundamental groups for people and for religious groups as well as religious minorities

Table 12: Test from new data, medium length sentence 2.

Input Sentence	Le député va dans la bonne direction, c'est-à-dire qu'il a le coeur accroché à la bonne place, mais je crois que nous pourrions mettre l'argent dans un régime de pension plus progressif
Actual Translation	The member is going in the right direction in terms of his heart and is being very thoughtful, but again I think we could put the money into a more progressive pension system
Experiment 1: Default	The honourable member is moving at the right direction but I think that the pension system could be extended to us in a more
Experiment 2: Reverse Input Off	The honourable member is moving at the right direction but I think that the pension system could be extended to us in a more
Experiment 3: 4 Layers	The point is that it is in the right direction but I think that we would be very much in favour of the creation of a fund
Experiment 4: 256 State Size	The member is quite right in saying that it is right that the money should be used but I think we can even be able to introduce a more favourable pension system
Experiment 5: 30,000 Vocab Size	The honourable member is in the right but I believe that we have the right to go further than we can in a pension system
Experiment 6: 20 Epochs	The honourable member is right in his right sense of the word but I think we could make a pension in a lot of money

Table 13: Test from new data, long length sentence 1.

Input Sentence	Franchement, que cela nous plaise ou non, n'eut été des mesures prises par le gouvernement, ce serait bien étonnant si nous avions une entente internationale sur les stocks de poisson
Actual Translation	Frankly, whether we like it or not, were it not for the actions of this government I would be surprised if we had an international agreement that deals with fish stocks
Experiment 1: Default	Frankly we would be wrong if we were to take action by the government if it were not a international fish system then we would have a credible effect on the stocks of fish
Experiment 2: Reverse Input Off	Frankly we would be wrong if we were to take action by the government if it were not a international fish system then we would have a credible effect on the stocks of fish
Experiment 3: 4 Layers	We would be very grateful if we were to be able to act on the other side of the atlantic if we were to fish down the international public
Experiment 4: 256 State Size	Frankly we do not think that this would be a matter of or if we were to act as a response to the international community we are now being told that there is a risk of being in the future
Experiment 5: 30,000 Vocab Size	Frankly if we were not to be the measures we had or have been told by the government we would have been well aware of the international fishing stocks
Experiment 6: 20 Epochs	Frankly we were told that it was not so much a question of whether the measures taken by the government were in fact known to us if we had international fish stocks

Table 14: Test from new data, long length sentence 2.

4.3 Analysis

Overall, each model that we trained tries to understand the meaning of the entire sentence before trying to translate to English using their limited capabilities, which is consistent with the theory behind sequence to sequence learning using encoder-decoder networks discussed earlier. This is most apparent in shorted sentences. Compared to the actual translation, these translations use less complex vocabulary while still expressing the same meaning. In Table 12, instead of using the more accurate phrase *"the duty to accommodate"*, some of the models use simpler phrasing, such as *"the need to meet"* in Experiment 5 and *"the duty to meet their needs"* in Experiment 6. In Table 5, the last two words of the actual translation *"daily basis"* get translated to *"every day"* for Experiments 1, 2 and 6. In longer sentences, the translations become more nonsensical. We also see a tendency for each model to try to translate each French word or find a simpler but equivalent English word to use in the translation even if the resulting translation does not make any sense. This is most apparent in long sentences such as in Table 8. We see the same results in Table 12 with some translated sentences losing the meaning of the original sentence completely.

In terms of sentence length, it seems that shorter sentences generally translate better than longer sentences even when using test sentences from the training set. We also see that there is a tendency for translated sentences to be truncated without any attempt to express the meaning of the input sentence. We can see this in Tables 9 and 10. In Table 9, Experiment 6 outputs *"the problems are still"* ending abruptly without saying that the problems have gone away as expressed in the actual translation. In Table 10, Experiment 4 does not mention the world *"airports"* and Experiment 6 makes no attempt at trying to express budgetary cuts.

Firstly, comparing the results of Experiment 1 with Experiment 2 where in Experiment 2 the parameters are the same as Experiment 1 except with the input sentence reversed, it can be seen that translations are generally identical, with some translations in Experiment 2. For example, in Table 6, we can see that Experiment 2 is able to capture a little bit more of the meaning than in Experiment 1 by using the words *"internal market"*. We find that by using 4 layers, the model tends to use vocabulary that is more complex and consist with the actual translations. Using a smaller state size in Experiment 4, we get mixed results. In some cases, the translations are better than the default model while in other cases,the translations are worse. For example, Table 9 shows an instance where the translated sentence from Experiment 4 is better than Experiment 1. However, in Table 7, Experiment 4 outputs a sentence that excessively uses the words *"authority"* making the sentence more nonsensical than the generated sentence in Experiment 1. If we had more time, we would investigate closer the relationship between the state size and translation quality. Using a 30,000 vocabulary size tends to make the model better. This is because the model is less constrained in the words that it can use and therefore have access to words that better fit the translation. For example in Table 10, the generated sentence from Experiment 5 is the only sentence to use the word *"burden"* which better encapsulates the meaning of the original sentence. Finally, training the model using 20 epochs instead of 10 generally leads to better translations as well. This is apparent in Table 6 with Experiment 6.

4.4 BLEU Scores

For each plot in Figures 2, 3 and 4, scores from the training data are generally higher which makes sense because the models should be able to translate sentences from which it was trained on better than sentences that the model has not seen. We also see that for shorter sentences, the difference between the average BLEU scores are relatively small compared to longer sentences. What is most surprising, however, is that by halving the state size from 512 to 256 alone, the model drastically in terms of BLEU score. This suggests that perhaps reducing the complexity of the both neural networks is beneficial. For shorter sentences, it seems that the difference between the models are more minimal, however for larger sentences, it seems that decreasing the state size, using a larger vocabulary is more beneficial. While the case with decreasing the state size requires further investigation, using a larger vocabulary size makes sense since larger sentences will likely contain words that may not be found with a limited 10,000 vocabulary size. While translating, if a model encounters a word that is not part of the vocabulary, the model will ignore it. With a larger vocabulary size, the model is more likely to find a suitable word for translation. Overall in Figure 5, it seems that using a smaller state size, more vocabulary and training with more epochs will yield better translations. Figures 6 and 7 summarizes the results of the project for short, medium and long sentences for each experiment using

sentences from the training data and data from the Canadian Parliament proceedings that the models have never seen before.

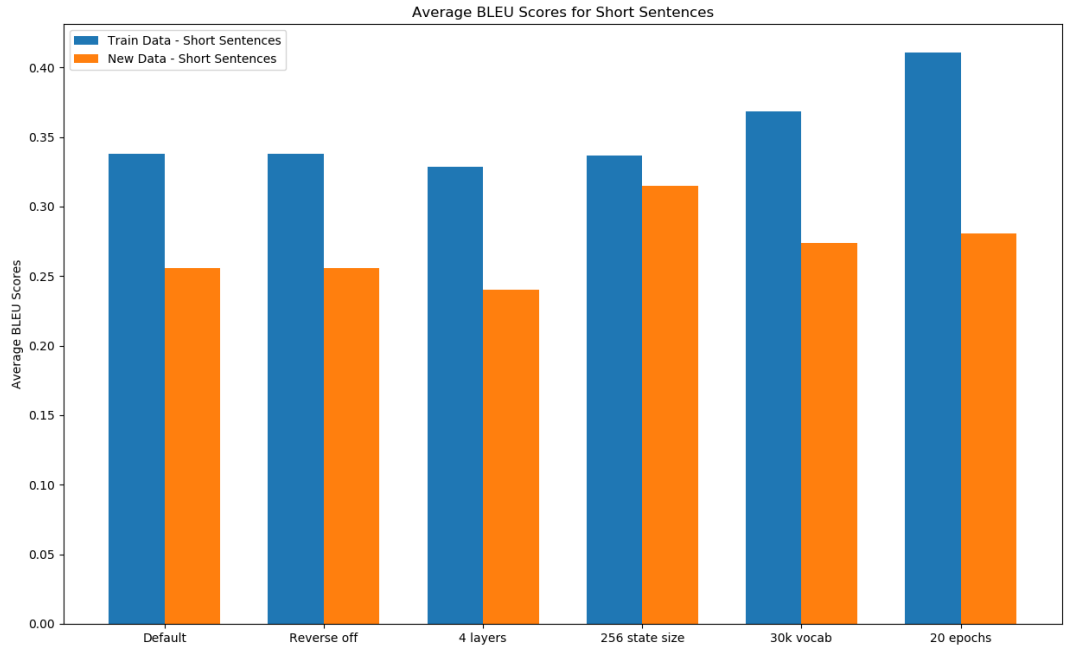

Figure 2: Average BLEU scores for short length sentences.

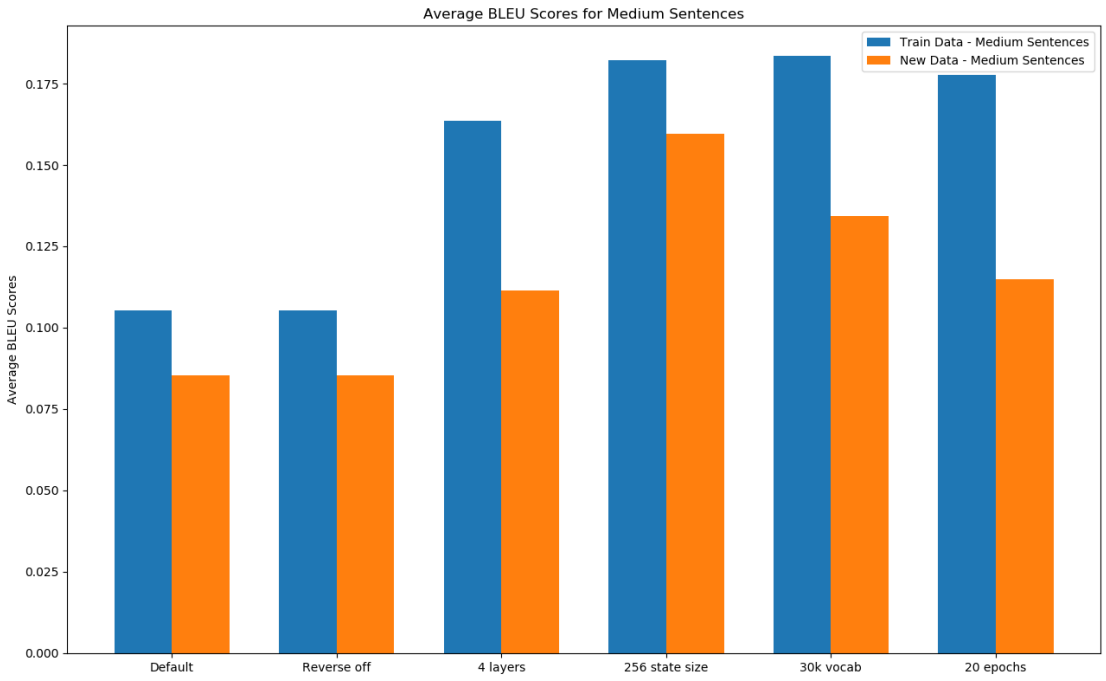

Figure 3: Average BLEU scores for medium length sentences.

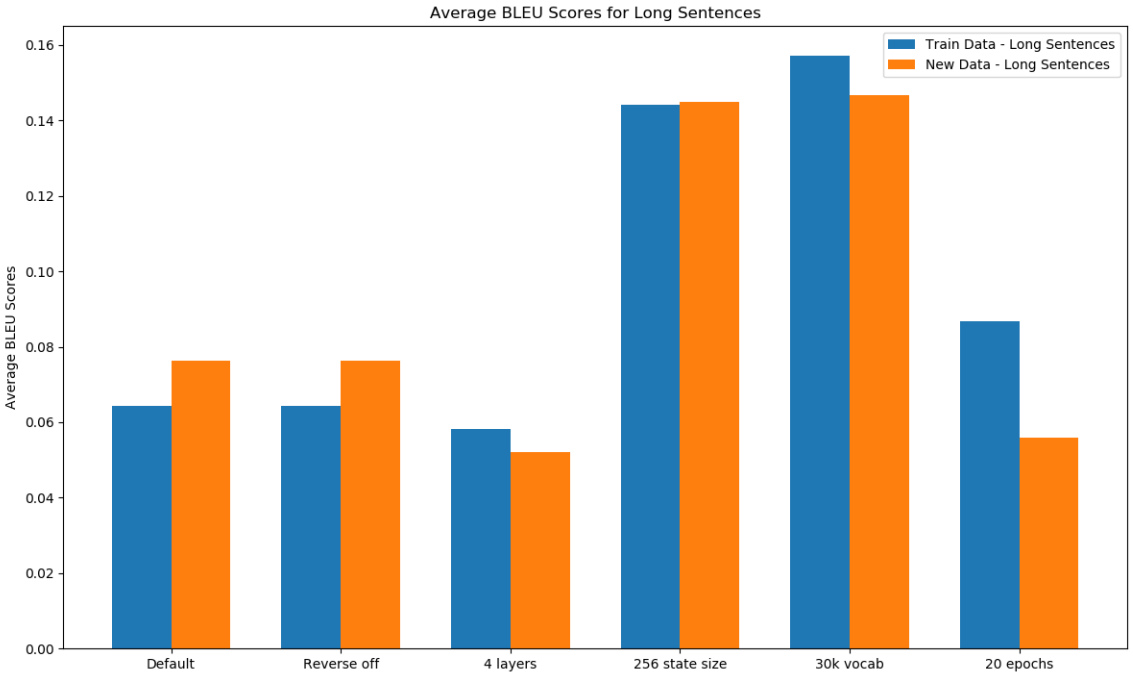

Figure 4: Average BLEU scores for long length sentences.

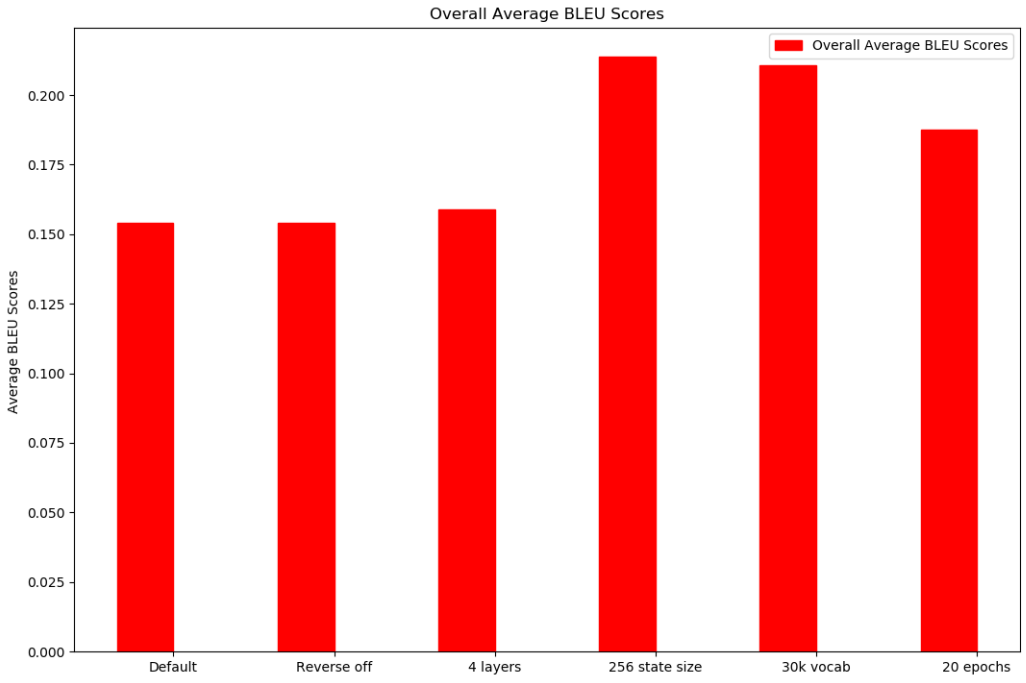

Figure 5: Overall average BLEU scores for every sentence length.

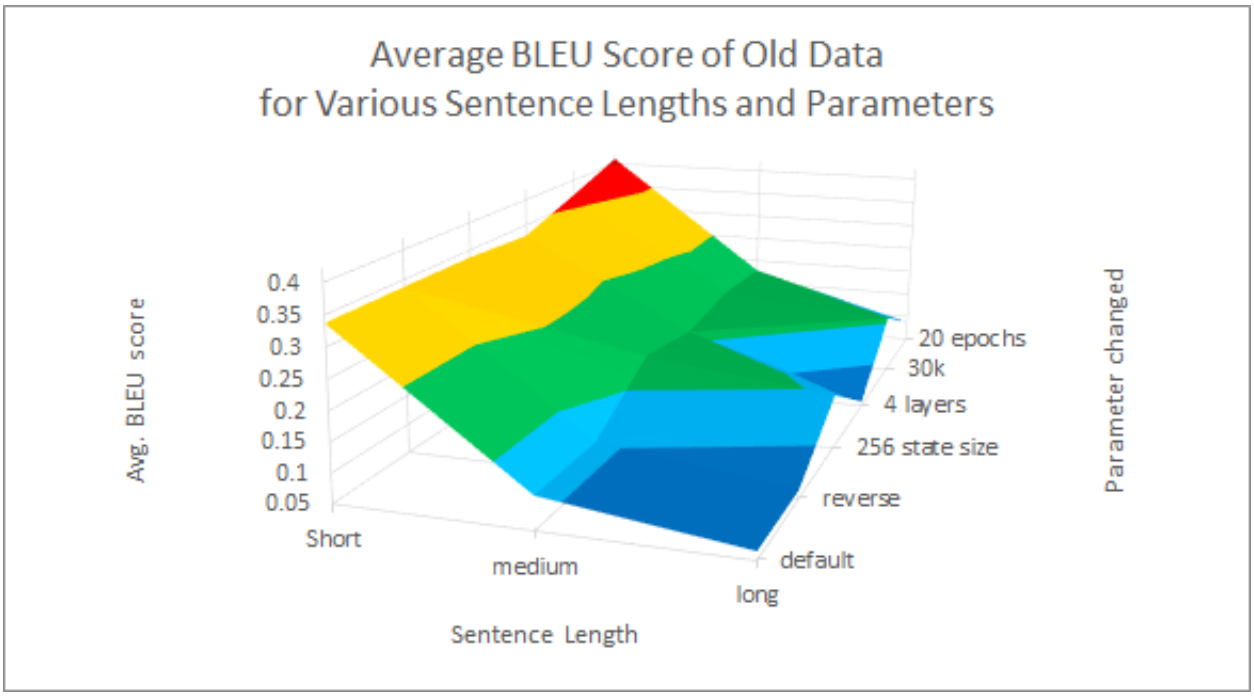

Figure 6: Average BLEU scores for every sentence length for every experiment using training data.

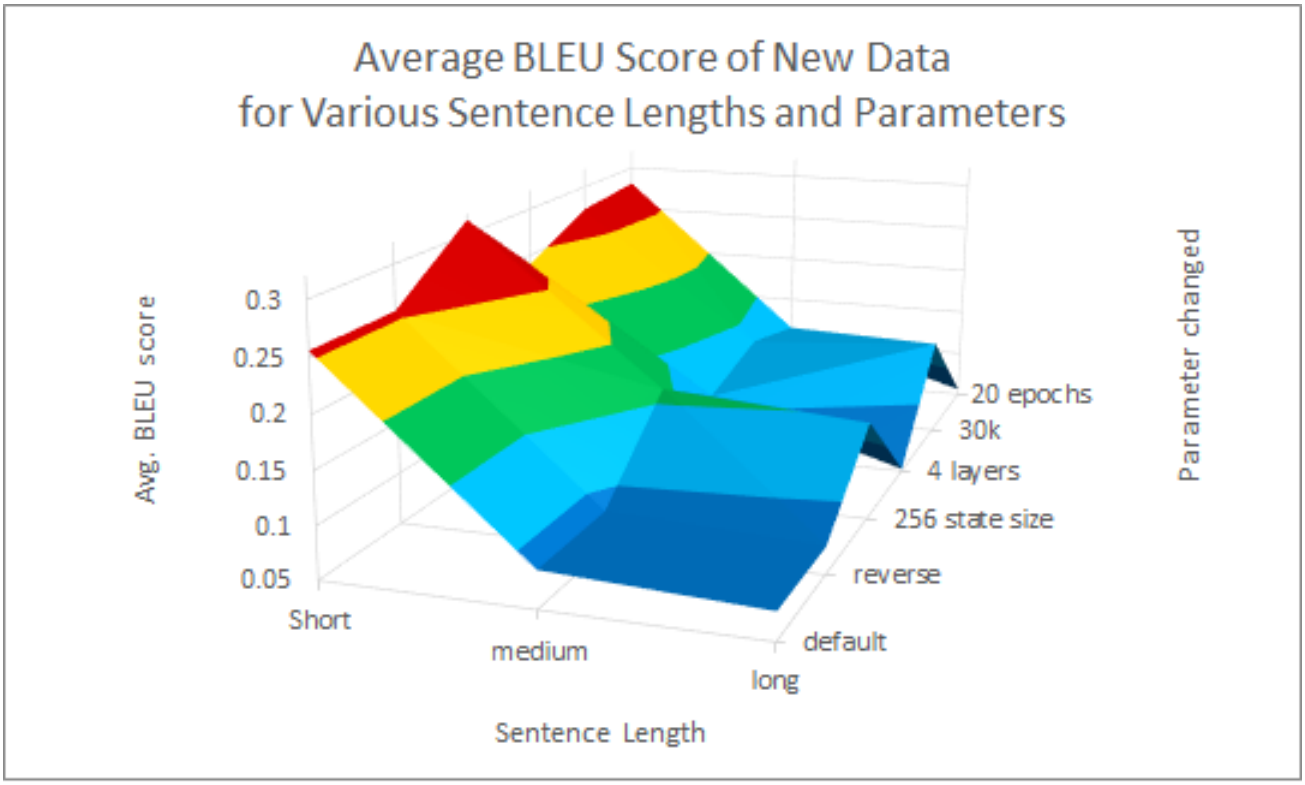

Figure 7: Average BLEU scores for every sentence length for every experiment using data from the Canadian Parliament proceedings.

5 Conclusion

In this project, we trained various models to perform natural language machine translation using sequence to sequence learning with encoder-decoder framework. We found that the models tended to translate shorter sentences better than longer sentences and that models tended to use simpler and less complex vocabulary during translations. Overall, we found that using a smaller state size, using a larger vocabulary size and training with more epochs led to better models.

5.1 Ways to Improve the Model

The model that we have used in this project so far is the most basic form of encoder-decoder sequence-to-sequence training. There are various way of improving the model including:

1. **Attention mechanism**: This would be employed in the decoding layer. During translation, attention allows the decoder to focus only the most relevant parts of the input sentence [8]. The most relevant parts of the sentence will tend to be the words surrounding the current word to be translated, since these words provide context. The signals of these words are then boosted in value drowning out the other words in sentence.

2. **Bidirectional RNNs**: In a vanilla RNN, the output at t only depends on the outputs from previous timesteps. This is because vanilla RNNs only use the information the past or its "memory" to predict the output at timestep t. In a bidirectional RNN, however, the RNN can use information from future timesteps in addition to information from past timesteps to generate a prediction [9]. This has the potential to improve the quality of the translation by allowing the model to understand the context of the entire sentence while generating predictions for the translated word.

3. **Teacher forcing**: This technique can be used to quickly and efficiently train a RNN. This technique works by using the actual output from the training dataset at the current time t as input for the next timeset $t+1$ rather than using the predicted output result generated by the network [10]. For example, say we wanted to train a model to generate the phrase *"Mary had a little lamb"*. During training, if the first word that the model generates is *"little"*, teaching forcing stipulates that instead of using this word as the input for the next timestep, we instead use the word *"Mary"*. Doing this can ensure that the model converges and learns quickly.

4. **Beam searching**: Beam searching is a technique used to predict an translation. Rather than using a greedy algorithm to find the next best possible translation at time t, beam search involves looking at and keeping track of the best n translations at any time step [11]. This allows the model to work with more than 1 translation at a time give it more options to choose from. The value n is called the beam width.

References

[1] S Sreelekha. Statistical vs rule based machine translation; a case study on indian language perspective. *arXiv preprint arXiv*, 1708, 2017.

[2] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. *Advances in NIPS*, 2014.

[3] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[4] Magnus Erik Hvass Pedersen. Tensorflow-tutorials. https://github.com/Hvass-Labs/TensorFlow-Tutorials, 2018.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[6] P. Koehn. Europarl a multilingual corpus for evaluation of machine learning. 2002.

[7] Aligned hansards of the 36th parliament of canada. https://www.isi.edu/natural-language/download/hansard/, 2001.

[8] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[9] Mathias Berglund, Tapani Raiko, Mikko Honkala, Leo Kärkkäinen, Akos Vetek, and Juha T Karhunen. Bidirectional recurrent neural networks as generative models. In *Advances in Neural Information Processing Systems*, pages 856–864, 2015.

[10] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[11] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *CoRR*, abs/1702.01806, 2017.