Neural Machine Translation

Hengyu Tang
New York University
ht1162@nyu.edu

Liangzhi Li
New York University
ll3399@nyu.edu

Yida Zhou
New York University
yz4499@nyu.edu

Abstract

With several innovative natural language processing techniques introduced in recent years, especially with the deployment of recurrent neural network to process variant length of input sequences, it enables us to build recurrent neural network based machine translation system using encoder-decoder structures. Moreover, with the aid of attention mechanism, and its extension self attention, we are able to better our model performances by paying attention to compatibility of specific previous contexts and translated information to generate the next word more 'wisely'. Under this model schemes, we attempt to tackle the challenge of Chinese-to-English and Vietnamese-to-English translations with qualitatively analysis on contributions of adopted techniques to improve the model performances.

1 Introduction

The goal of this project is to build a sequence to sequence neural machine translation system for two language pairs, Vietnamese(Vi) to English(En) and Chinese(Zh) to English(En) based on recurrent neural network (RNN). We adopted the standard Encoder-Decoder structure for neural machine translation (Sutskever et al., 2014). Two RNNs are trained together to capture distinctive features of both source and target languages. The encoder will compress a source sentence into a single vector representation, while the decoder will decode such representation one token at a time to generate translation based on learned features. Moreover, we are particularly interested in exploring the importance of contexts in translation tasks, hence we introduced attention mechanism into our decoders to integrate both source sentences and translated target tokens for better translations. Besides improving the decoder by spotting which source token needs to be translated most immediately, we would also like the source sentence to learn from itself within its local regions, by introducing self attention mechanism in encoding phases.

Therefore, We implemented three models, a RNN encoder decoder without attention, a RNN encoder and attention deployed decoder, and a self-attention deployed RNN encoder and attention deployed decoder. We challenged the models on both language pairs. A pre-trained embeddings were adopted for all three languages to speed up training process while providing better representations of input sequences. Teacher forcing was also used during training process to make our network converge faster. Additionally, we adopted corpus BLEU(Matt and Rico, 2018) score as the evaluation metric and picked out the best model for both languages based on their performances validation set to carry out further hyper parameter tunings. The final BLEU score over test dataset is reported under the optimal model configurations, followed by a detailed analysis on model performances under different language settings.

2 Related Work

Neural Machine Translation(NMT) is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems(Koehn et al., 2003) Unlike the traditional statistical approach to machine translation, NMT is mostly based on a encoder-decoder based architecture(Sutskever et al., 2014) for each language.

Attention mechanisms have also become an integral part of translation models in various tasks, allowing modeling of dependencies without to their distance in the input or output sequences. (Wu et al., 2016) Transformer is a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. (Cho, 2018)

3 Dataset

Both Zh-En and Vi-En datasets are pre-tokenized with different tokenization schemes. Chinese dataset is tokenized into mono, bi and tri-grams by character counts, where English and Vietnamese are tokenized to only monograms. However, the definition of a token for Chinese is rather ambiguous, as a word could contain two characters for a single meaning. Both datasets are derived from International Workshop on Spoken Language Translation (IWSLT, 2018) automatic transcription and translation of TED and TEDx talks. We are going to use IWSLT Vi-En and Zh-En datasets in this paper. The 2013 IWSLT set is used for development and 2012 IWSLT set for testing.

3.1 Preprocessing

After implementing the baseline model using pytorch Embeddings, our models trained very slow and scored poorly. Due to the fact that Chinese dataset is rife with bi and trigrams, it hence led us to refer to a better vector representations of the tokens. We first adopted wiki pre-trained vector representations(Mikolov et al., 2018). However, since our source data were derived from speeches, the incompatibility with wiki drove us to switch to fastText(Grave et al., 2018) pre-trained word vectors for three languages, English, Chinese and Vietnamese to build the vocabulary and create the embeddings for the following neural networks. Due to the constraints in computational power, we limit our vocab size to 50000 and the maximum length of the sentences set at 80. Such decision is supported by the distribution of the sentence length being extremely left skewed as at length 80, Chinese data preserve 99.44% of original training data samples, while Vietnamese preserve 99.60% of original training data samples.

We further cleaned the data by removing blank pairs, transforming to lower cases, and trimmed away punctuation except period, question mark and explanation mark. We noticed the existence of special characters such as *apos, quot*generated from tokenizing process. We removed those symbols due to its high occurrence. The final touch was to add the *eos* to indicate the end of a sentence, which also enabled the machine to learn when to stop generating tokens.
We also post-processed the data by cleaning out the *pad* token from the generated sentences, which was used to pad the sentence into fixed length. It was removed since it may jeopardize the BLEU score performance by introducing extra counts of overlapped tokens, while preventing short sentences being penalized by brevity penalty(reduction in BLEU scores on shorter sentences due to the tendency of machine generating shorter sentences). (Papineni et al., 2002)

4 Methods

4.1 RNN based encoder-decoder without attention

We adopt recurrent neural network as the basis for our encoder decoder machines. The basic structure of a RNN is given(Bahdanau et al., 2014):

$$h_t = f(h_{t-1}, e_t) \tag{1}$$

where h_t refers to the hidden state at time step t, while e_t represents the vector representation of next input. Hence by adopting RNN structures, we are able summarize the entire input sequence into a single vector. Moreover, previous language models all focused heavily on assumption of autoregressive models, where:

$$p(x_1, x_2, x_3, x_4 ... x_T) = \prod_{t=1}^{T} p(x_t | x_{<t}) \tag{2}$$

Such model assumes that the distribution of the t^{th} token only depends on tokens before it. Such assumption is rather counter-intuitive, and jeopardizes the performance of the model by ignoring the information from the tokens appearing after the t^{th} token. Therefore we adopted a bidirectional Gated Recurrent Unit(GRU) (Cho et al., 2014) to utilize information from both directions. The GRU unit is adopted to introduce shortcuts in neural networks to avoid gradient vanishing, which prevents influence of further away tokens from fading during back propagation stage. An update gate and a reset gate are trained over the training data to allow the models remembering short and long term dependencies.

4.2 RNN based encoder-decoder with attention

It is intuitively clear that when translating a sentence from one language to the other, parts of the sentences usually correspond well with each other both in positions and meanings. For example:

在 恐惧中 角色 就是 我们自己
in fears the characters are us.

The underlined tokens in the same position of two sentences aligns well in translation meanings. Therefore, rather than encoding an entire input sentence into a single context vector, we would like to encoder the source sentence into context dependent representation set. Moreover, the already translated words also provide important insight to generate the next token. Such effect is specially enhanced when we compare each vector representation of source sentences with the last hidden state of the decoder, within which contains the information of previously translated tokens. A score will be generated to rate the compatibility of each of the encoder outputs and the current hidden state of the decoder as the attention weights, to indicate which hidden representation of the input sentences needs immediate translations.

In this part, we referred to the Lab 8 material (Sean, 2017). The scoring function we adopted is to (Bahdanau et al., 2014) concatenate embedded and last hidden state that represents the translated information is fed through a neural network. The result was then compared with all encoder outputs to generate a score indicating the importance of each source token in deciding the next output. The most important token is soft selected in the form of a weighted sum to indicate which token in the source sentence needs immediate translation. The weighted sum is then concatenate again with embedded previous translated token and fed through the recurrent neural network to generate conditional probability of next input over the entire vocabulary.

4.3 Self-attention based encoder

After deploying attention mechanism onto our decoder machine, it is rather reasonable to consider applying similar mechanism onto the encoder machine. Due to lacking of additional information from source sentences during the encoding phases, we will let the source sentence to study from its nearby local regions.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstract summarization, textual entailment and learning task-independent sentence representations. End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks. However, The Transformer is the first translation model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution. (Vaswani et al., 2017)

In this part, we adopted the code from Harvard The Annotated Transformer (Rush, 2018) to implement an encoder which contains self-attention layers that values and queries come from the output of the previous layer in the encoder. In order for the sentence to study its own local regions, we split a hidden layer into pre-specified n intervals. We chose n=6. Within each interval, we adopted self attention mechanism by letting query, key and value to be the same. Each position in the encoder can attend to all positions in the previous layer of the encoder. Besides, we used the same decoder with attention from previous section to implement this encoder-decoder architecture.

4.4 Teacher Forcing

Teacher Forcing, or maximum likelihood sampling, means using the real target outputs as each next input when training. The alternative is using the decoder's own guess as the next input. We randomly chose to use teacher forcing with an if statement while training and sometimes we'll feed use real target as the input (ignoring the decoder's output), sometimes we'll use the decoder's output(McLeskey and Billingsley, 2008). In this project, we set the teacher forcing ratio to 0.5 for training. That means half of the input will be fed by the real target.

5 Evaluation

5.1 BLEU score

One of the most widely used automatic evaluation metric for assessing the quality of translations is BLEU(Matt and Rico, 2018). BLEU computes the geometric mean of the modified n-gram precision

scores multiplied by brevity penalty.

In this paper, we are going to use sacreBLEU(Matt and Rico, 2018) to calculate the BLEU@4 scores as the evaluation and report the BLEU@4 scores on the test dataset in the results section.

5.2 Greedy Search

Machine translation system is just like conditional language model. Given a foreign language $x = (x_1, x_2, ..., x_{T'})$, the translation function should return the conditional probability of $p(Y|X) = \prod_{t=1}^{T} P(y_t|y_1, ..., y_{t-1}, X)$, where $Y = (y_1, ..., y_T)$ is the target sentence.

Greedy search is an algorithm that helps to find the most likely English sentence given Chinese or Vietnamese sentence. 'Greedy' here means that it always choose the one that seems to be most likely at the current moment. In our encoder-decoder model, it first pick pick the mostly likely first word, according to the conditional language model, the pick the most likely second word, then pick the most likely third word, etc.

6 Results

In this part, we are going to show the results after training with three methods each on two language pairs. The results include train loss plots and the BLEU scores on the validation language pairs dataset.

6.1 Machine Translation for VI-EN

For the VI-EN language pairs, we set the same hyperparameters to the three methods to make easier to compare their performances. As mentioned in preprocessing section, we set 80 as the maximum length of Vietnamese sentences and English sentences, with the embedding created from the vocabulary of a vocab size 50000 generated by fasttext pre-trained word vectors. We set the embedding size to 300 and the hidden size to 300 used for the hidden layers in the neural network based encoder and decoder. Adam is used for the encoder and decoder optimizer, with learning rate equal to 0.001, and use log likelihood loss as the loss function. Besides, the teacher forcing ratio is set to 0.5 as mentioned above.

After 5 epochs for each method, we got the training loss curve by time in Figure 1, other two methods have similar curves, and the BLEU scores on the validation dataset in Table 1. From the results, we can observe that each model has achieved an converge on the loss, and all BLEU scores above 12 from which the Encoder-Decoder with attention model performs best.

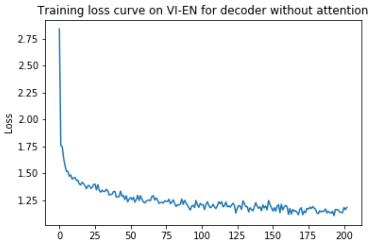

Figure 1: Training loss curve on VI-EN for the basic encoder-decoder architecture

6.2 Machine Translation for ZH-EN

The strategy to train the machine translation system for the ZH-EN language pairs is almost the same as the one for VI-EN. The only differences are that we set the maximum sentence length to 50 and used the different pre-trained vectors for Chinese sentences. All the other hyperparameters are the same as VI-EN.

After 5 epochs, we got the best model self-attention encoder-decoder with attention. We omit the three loss curves here because they are highly close to those from VI-EN above.

6.3 BLEU scores

The results of the three proposed models on Zh-En and Vi-En are shown in table 1. From the table, it is clear that the RNN based encoder-decoder with attention model outperforms the other two models on both language pairs.

Pairs	w/o attn	attn	self-attn
VI-EN	12.82	**22.42**	15.06
ZH-EN	12.79	14.50	**16.57**

Table 1: BLEU@4 scores on two language pairs by three methods (validation)

6.4 Hyperparameter Tuning

We have selected the best machine translation models for the two language pairs (VI-EN, ZH-EN). Now we can perform hyperparameter tuning on the best model we selected. In this part, we are going to try different hidden sizes for the encoder

Hidden Size	BLEU
100	20.12
300	**22.42**
500	21.51

Table 2: BLEU scores after hyperparameter tuning on the encoder-decoder with attention on VI-EN

and decoder with attention model on the VI-EN pairs.

Figure 2: Train loss curves by different hidden sizes for the encoder-decoder with attention model on VI-EN

6.5 Visualizing Attention

A useful property of the attention mechanism is its highly interpretable outputs. Because it is used to weight specific encoder outputs of the input sequence, we can imagine looking where the network is focused most at each time step.

Figure 3 is attention plots from the attention model of Vi-En language pairs. The x axis and y axis are tokens of source sentence (Vi) and translated sentence in target language (En) respectively. The pixels in the plot refer to the attention weights of each source token on each target token. It is clear from these plots which tokens in the source sentence are important (need more attention) when generating each word in target language. Usually in a well trained attention model, the target token should attend to its corresponding source word and its close context. Thus we would expect a diagonal pattern in the attention plot like figure 3.

6.6 Results on Test Data

After training, for both language pairs, we choose the model with the best performance on the validation dataset to test on the test set. The model we chose for VI-EN is encoder-decoder with attention model and the other one we chose for ZH-EN is self-attention encoder-decoder with attention model. The results showed that BLEU score is **23.98** on the VI-EN test pairs, and it's **18.59** on the ZH-EN test pairs.

7 Analysis and Discussion

It has been shown that bleu score will typically get decreased as the length of sentence increase. This is because it is difficult for a neural network to memorize a very long sequence. On the other hand, attention mechanism makes the machine behave more like a human that it allows the machine to pay attention to critical parts of the input sequence by assigning the attention weights. Therefore, the performance of attention model will be less affected by increase of sentence length.

One observation we encountered was that the cc-wiki pre-trained embeddings does not include numeric number in its vocab. The situation is particularly outstanding in English language, where most numbers are marked as UNK token. The fact that both Chinese and Vietnamese languages tend to use characters rather than digital to represent numbers compared with English, when decoding the numbers into English, it is hence challenging for decoder to understand numbers. One example is shown as below. The machine is unable to generate seven as the corresponding translation to Chinese where the ground truth marked as "UNK".

>那 是 早上 七点 我 还 穿着 .
= he is painting a picture? .
<that s my morning i was morning .

Another observation we found interesting is that most sentences generated during the inference session, all translates well at the beginning, but deviates when time steps increases. One example is given below:

> 这 不是 个 隐私 反对 安全 的 问题 .

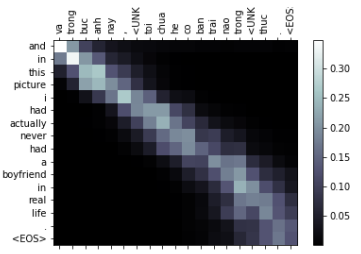

Figure 3: Attention visualization example on VI-EN

= this is not a question between privacy against security.
< this is not a issue ...
The machine translates well at the beginning of the target sentence. Also, we noticed that the machine generates a synonym, issue, to the original word question. It from another perspective verifies a good hidden representation of each words, as similar words are close to each other in continuous vector space. Another common phenomenon was that the greedy search only performs well at the beginning of each sentence translation. When the length increases, due to the limitation of greedy search which optimizes looking at the current time step, greedy search performs poorly. On the other hand, greedy search tend to perform well when the to-be-translated sentence is short, an example of such is provided below:
> hay gio no len .
= hold it up.
< please raise it.
The predicted sentence is very similar in meanings compared with the target sentence. However, as a human it is intuitive that the prediction and the ground truth are interchangeable, i.e. greedy search scored high in inferring shorter sentences. However it may not be the case for BLEU score. Although the performance of the model is measured on corpus level, the focus on counting of overlapping n-grams presents its limitations in capturing similarities in meaning.
The fact that self attention machine underperformed simple RNN encoder may be because of the complex structures of multi-headed attention, making the model possibly harder to train under same number of iterations of training. Based on domain knowledge of Chinese, there are abundant of characters with abstract meanings, such as grammar tense indications(Ren, 1968). These words usually follow closely to common verbs but get tokenized into monograms. Hence, when deploying self attention onto Chinese source sentences, it usually provides extra information in local regions. Such leak of information, usually help perfects the grammar tense of the translation, hence improving ZH-EN translation better than VI-EN translation.

8 Conclusion

In this project, we showed that the encoder-decoder RNN model can be used to build a neural machine translation system. Attention mechanism helps the model to focus on the most important parts on source text, and we see attention decoder outperformed the basic RNN encoder-decoder model for both language pairs and self-attention encoder outperformed simple RNN on Vi-En. Therefore, we conclude that attention model can help to improve the performance of machine translation system.

It is also interesting to note that different source language pairs may have different optimal model for the translation task, maybe due to different innate structure of languages. This is actually a case-by-case problem, so we should always try different models before actual deployment.

Future work on neural machine translation task can be focused on improve the search algorithm. We used greedy search in this project, but there are more advanced search algorithm such as beam search that can be applied to the model and may achieve higher performance.

9 Contribution

Hengyu Tang implemented models, pretrain embeddings, data cleaning, greedy search and miscellaneous, and Liangzhi Li worked on the encoder and decoder, test set and run the models, and Yida Zhou focused on training models, generating and plotting results. All of the team members worked on this report. The code for this project is stored in our Github repository (Hengyu et al., 2018).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Kyunghyun Cho. 2018. Natural language understanding with distributed representation. https://github.com/nyu-dl/NLP_DL_Lecture_Note/.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tang Hengyu, Liangzhi Li, and Yida Zhou. 2018. Neural machine translation. https://github.com/SuperLatte/Neural-Machine-Translation.

IWSLT. 2018. Iwslt dataset. https://workshop2018.iwslt.org/.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 48–54.

Post Matt and Sennrich Rico. 2018. sacrebleu package. https://github.com/mjpost/sacreBLEU.

James McLeskey and Bonnie S Billingsley. 2008. How does the quality and stability of the teaching force influence the research-to-practice gap? a perspective on the teacher shortage in special education. *Remedial and Special Education* 29(5):293–305.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Chao Yuen Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press, Berkeley, US.

Alexander Rush. 2018. The annotated transformer. http://nlp.seas.harvard.edu/2018/04/03/attention.html.

Robertson Sean. 2017. Translation with a sequence to sequence network and attention. https://pytorch.org/tutorials/intermediate/seq2seq_translation.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* .