English to Indonesian Neural Machine Translation

ANURAG CHATTERJEE (A0178373U), BHUJBAL VAIBHAV SHIVAJI(A0178321H),
GOH CHUNG TAT KENRICK(A0080891Y), LIM PIER (A0178254X),
LIU THEODORUS DAVID LEONARDI (A0178263X)

Abstract

Bahasa Indonesia, or Bahasa is the main language of Indonesia which is the largest economy in South East Asia. Hence, there is a demand for translation from English to Indonesian for various business needs. In this work we evaluate two different neural machine translation models on a simple English to Indonesian text corpus. We compare our approaches using both the Bilingual Evaluation Understudy Score (BLEU score) and a Turing test. We see that the attention-based sequence to sequence (Seq2Seq) translation model performs better than the vanilla Seq2Seq model on both the metrics.

1 Introduction

Indonesia is the largest market in South East Asia (SEA) with an estimated population greater than 250 million in 2017 [1, 2]. With an economic boom in the recent years, the middle-class group in Indonesia is growing with higher purchasing power. As such, there is an increasing appetite on good and services in the largest SEA market in the world. Based on the McKinsey report, 78% of the internet users in Indonesia made online purchases. Thus, there is an untapped potential in the e-commerce market considering the growth of Internet penetration and mobile users in Indonesia [3].

The official language of Indonesia is Bahasa Indonesia. As such, it is pertinent to cater to the domestic market by promoting one's product and services in their local language. Thus, there is a demand for machine translation of product details and features to appeal to the domestic market. In addition, one can also use machine translation to facilitate the translation of business documents. Last but not the least, it prevents the breakdown in communication when tourists were to travel in Indonesia for a holiday.

The earliest Machine language translation was developed during World War II where it was used to decode the German's military communication. Ever since then, the scope has largely shifted from the military side to the commercial area [4].

In recent years, with the huge increase in computational power, there is an increasing interest in Machine Translation. New techniques are constantly developed, bringing Machine Translation to a new level. In this report, we focus on two techniques, a vanilla Seq2Seq model and a Seq2Seq model with attention module for translation from English to Bahasa Indonesia.

With a native speaker within the team on this project, it grants us the ability to quickly verify the authenticity of the generated sentences for the modelling stage of this project.

Here are some examples of the quirks of English to Indonesian:

English	**Indonesian**
These are my boys	Mereka ini adalah anak-anak laki-lakiku
Breakdowns	
boy	anak laki-laki (anak: kid, laki-laki: man)
Plural form • kids/children • boys	***Repeat the words*** • anak-anak • anak-anak laki-laki
Possessive form • my [noun] – 1st person ownership • E.g. my son / my boy • his/her [noun] – 3rd person ownership • E.g. his pencil	**Possessive Form** • [noun] - *ku* • anak laki-laki*ku* • [noun] – *nya* • pensil*nya*

2 Data Sources and References

Through our research, we have found 3 different sources of English-Indonesian Corpus. There are namely:

Figure 3 English to Indonesian samples

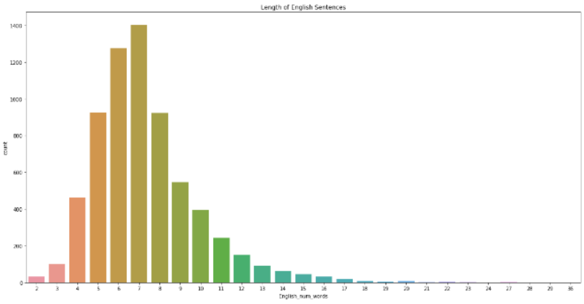
No	Name of Corpus	Sourced From	Number of Pairs
1	BBC-468.en	BBC	468
2	SMERU_en	SMERU Research Institute	26,966
3	Manything_anki	Tatoeba	6,752

Some examples of the English to Indonesian translation are shown below:

Corpus	English	Indonesian
SMERU	We are truly grateful to the Family Court of Australia, especially Leisha Lister, the executive adviser	Kami sangat berterima kasih kepada Family Court of Australia, terutama Leisha Lister, selaku penasihat eksekutif
Tatoeba	May I use your phone?	Apakah saya boleh menggunakan ponsel Anda?
BBC	Two French Muslim women have become the first to be convicted of covering their faces with veils in public, in defiance of a new law.	Dua Muslimah Prancis menjadi orang-orang pertama yang dinyatakan bersalah mengenakan burka di tempat umum, bertentangan dengan undang-undang baru.

From the examples above, we can see that the Tatoeba corpus has simple sentences compared to SMERU and BBC. SMERU and BBC sources were not considered further, and the simpler dataset sourced from Tatoeba was chosen.

The sentences from the Tatoeba were simple enough and an analogy of teaching a child Bahasa from English could be made while training the Neural network. The “Manything_anki” corpus from Tataoeba consists of 6752 English Bahasa pairs[1].



2.1 Data Exploration

We did initial data exploration of the length of the sentences in the combined corpus. The following graph shows the distribution of the number of words in a sentence across the corpus for English and Bahasa.

Figure 1 English word counts

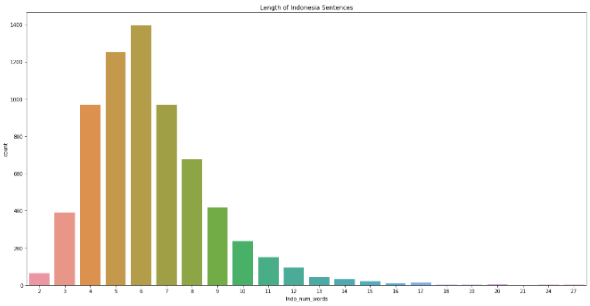
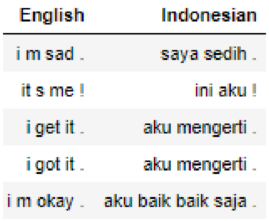


Figure 2 Indonesian word counts

We see that the median number of words per sentence is 7 for both English and Bahasa. The English vocabulary consists of 3500 unique words while the Bahasa vocabulary consists of 4500 unique words. The below shows a few samples from the corpus:



English	Indonesian
i m sad .	saya sedih .
it s me !	ini aku !
i get it .	aku mengerti .
i got it .	aku mengerti .
i m okay .	aku baik baik saja .

[1] The corpus used can be downloaded from this link: https://www.manythings.org/anki/ind-eng.zip.

The two images below show the Word Cloud of both English and Indonesian Corpus.

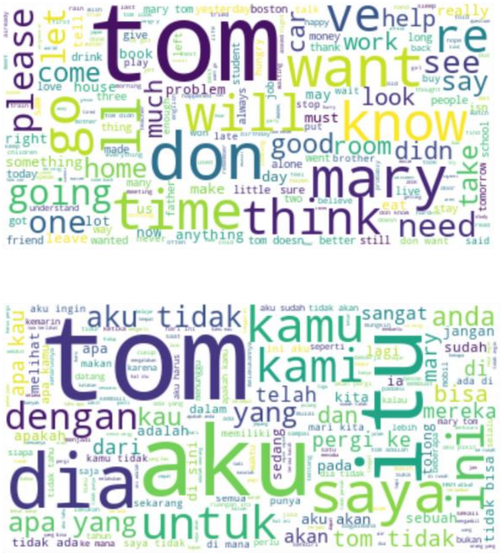


Figure 4 Word clouds of the training vocabs

Initial inspection of the corpus through visual inspection shows that there are a huge number of subjects in the datasets. Tom and Mary are the subjects of the sentence. This leads to the hypothesis that having such many of these subjects will affect the translation accuracy.

Therefore, our team wished to delve deeper into this problem by updating the corpus. We increase the size of the corpus by creating more sentence pairs with updated subject in both English and Bahasa.

Proposed updates to the Corpus

Before		After	
English	Bahasa	English	Bahasa
Tom hates me.	Tom membenciku.	Tom hates me.	Tom membenciku.
Tom is tense.	Tom tidak tenang.	Tom is tense.	Tom tidak tenang.
Tom needs it.	Tom membutuhkannya.	Tom needs it.	Tom membutuhkannya.
		John hates me.	John membenciku.
		John is tense.	John tidak tenang.
		John needs it.	John membutuhkannya.

We hypothesized that with the increase of the corpus as well as decrease in the importance of the subjects in the corpus, we are able to get better results in Machine Translation from English to Bahasa. This is be explained further in the Preprocessing portion.

2.2 Preprocessing

The English and the Bahasa words are first converted to lower case and then tokenized using NLTK library's "word tokenize". All sentence pairs where number of words in English and Bahasa are less than 25 and more than 2 are kept. However, since a sentence would have a start '<s>' and stop '</s>' symbol when input to the model, hence only sentences with length less than 23 are kept. The start of sentence and end of sentence marker (<S> and </S>) are added to each sentence in the pair. Gensim dictionary structure was used to contain the mapping between the words and their integer IDs for both the English and the Bahasa words. A tensor is created to contain the ID mappings for the English words for each input English sentence.

The dataset is then split into 85% training set and 15% validation set.

2.3 Augmentation of the dataset.

Based on the initial corpus observation, we realized that there were two subjects that stood out, “tom” and “mary”. Thus, we bootstrapped the dataset by replacing “tom” and “mary” with another common name like “john”, “sarah”, “lim”, etc randomly and repeated this process six times for each sentence.

	Before Augmentation
English	
Bahasa Indonesia	

	After Augmentation
English	
Bahasa Indonesia	

3 Software and Libraries Used

We used Python 3.6 as the main language for this task, and PyTorch 1.0.0.post2 for defining and training the deep learning models. Some of the other libraries used are as follows:

Library	Version
PyTorch	1.01post2
Gensim	3.7.1
NLTK	3.4
Matplotlib	3.0.2

The main reason for choosing PyTorch is the dynamic computational graphs, which allows modification during runtime. Like Tensorflow, it also has high adoption rates and is much favored by the research community. Gensim was used mainly for the availability of the Dictionary utility, which allows us to map easily from word to the integer IDs. NLTK was used mainly for its word_tokenize functionality. Matplotlib and seaborn are used for visualization purposes.

4 Model

4.1 Vanilla Seq2Seq Encoder-Decoder Model

The key benefit of a Seq2Seq model is that is allows end to end training on a single model directly on the source and target sentences and can handle variable length input and output sequences of text. The Seq2Seq model is used as the basic architecture for most translation models.

The baseline model that we have created uses an encoder and a decoder as shown below.

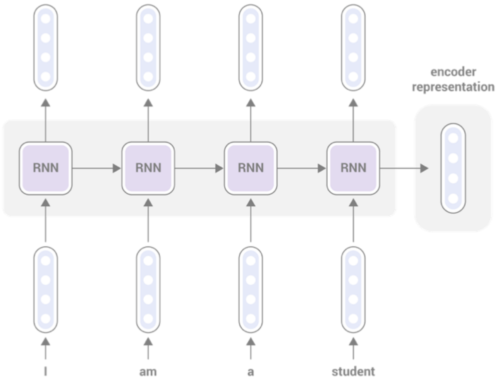


Figure 6 Encoder conceptual representation

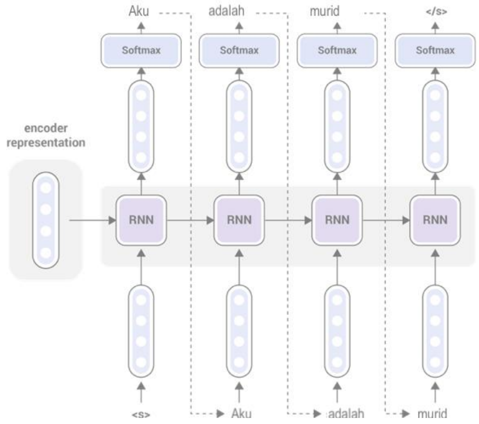


Figure 5 Vanilla decoder conceptual architecture

The inputs to the encoder are the tensors corresponding to the English tokens. For each GRU cell, it receives the previous hidden state from the prior GRU cell as shown in the picture. The input is fed into an embedding layer whose input dimensions are the English vocabulary size and the output dimensions are the hidden size which is a hyper parameter. The embedded tensor is fed into the GRU layer with the previous hidden state as the other input. The hidden state of the last GRU cell is known as the encoder representation, and this is the vector that is passed on to the decoder.

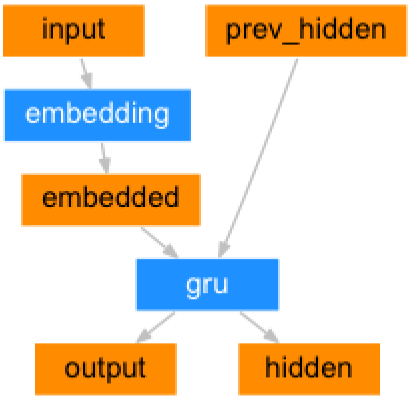


Figure 8 Encoder architecture

The decoder uses the previous hidden tensor and the input from the target Bahasa vocabulary if teacher forcing is enabled. The input is fed into an embedding layer whose input is the size of the Bahasa vocab and output is the hidden size which is a hyper parameter. ReLU activation is applied to the output of the embedding layer and fed into the GRU network as an input along with the previous hidden tensor. A linear layer of output dimension the size of the Bahasa vocabulary size is fed the output of the GRU network. **Log Softmax** activation is applied to this final linear layer. Training is performed to minimize the **negative log likelihood loss** using the SGD optimizer with learning rate of 0.0001.

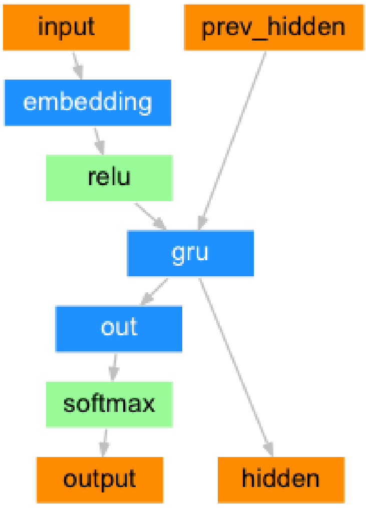


Figure 7 Vanilla Decoder architecture

4.1.1 Pros of this Model

This model is simple to train and understand among the different Seq2Seq architectures. It is a relatively simple baseline model that helps to understand the base performance of the translation task.

4.1.2 Cons of this Model

One of the main drawbacks of the plain vanilla Seq2Seq encoder-decoder model is that all the decoder knows about the encoder portion of the model is a fixed-length vector, no matter how long or complex the sentence is. This is a bottle-neck for the training, no matter how much the parameters of the model varies.

4.2 Seq2Seq Encoder-Decoder Model with Attention

In the Vanilla Seq2Seq model the single context vector is responsible for encoding the entire input sequence irrespective of the length of the input sequence. To aid the decoder to generate better translations, additional inputs in form of attention weights are generated that are multiplied with the outputs of the encoder to create a weighted combination.

The encoder for the Seq2Seq model with attention is the same as the encoder used for the Vanilla sequence to sequence model. The output of the encoder has dimension “hidden_size” which is a hyper parameter. The decoder network has 3 input tensors: the output from the encoder, the tensor corresponding to the token from the Indonesian vocab taken from the target if teacher forcing is used. Otherwise, the tensor for the token predicted in the earlier time step is used. The previous hidden state is initialized with the hidden output of the encoder. The input tensor based on the Indonesian vocabulary input is fed through an embedding layer of input dimension of Indonesian vocab size and output dimension corresponding to the “hidden_size”. Dropout is applied to the output to prevent overfitting to the training data.

The previous hidden tensor and the embedded tensor are concatenated into a linear layer with Softmax activation. The output which are the attention weights are multiplied with the encoder outputs as a matrix multiplication where the common dimension is the maximum estimated length of an output sentence set to 25. The attention applied tensor is concatenated with the embedded tensor generated from the embedding layer in a linear layer and ReLU activation is applied. This is fed to the GRU-based RNN along with the previous hidden tensor as the hidden state. A linear layer follows the RNN whose output dimension equals the Indonesian vocabulary size. **Log softmax activation** is applied to this final layer. Training is performed to minimize the **negative log likelihood loss**. Both SGD and the Adam optimizers are evaluated. Batch size is kept as 1 and we had a provision to keep the stopping criterion as the highest value of the BLEU score or the lowest value of validation loss observed with a patience of 10. So, if the BLEU score or the validation loss does not improve by a pre-defined threshold even after a pre-stated number of iterations of training, the training is stopped.

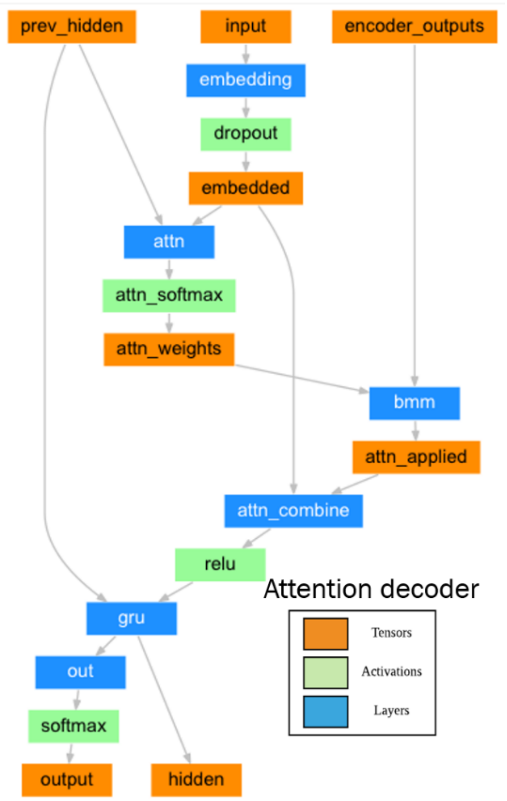


Figure 9 Attention Decoder architecture

4.2.1 Pros of this Model

Unlike the Vanilla decoder model, the attention mechanism ensures that the decoder is not constrained by just a fixed length representation vector generated by the encoder. The weights generated by the attention mechanism act as an additional information pathway from the encoder to the decoder.

4.2.2 Cons of this Model

Due to the sequential nature of the RNN based networks, the model performance deteriorates as the length of the sentence increases. So, we have limited our corpus to small sentences as mentioned in the pre-processing steps.

4.3 Fine-Tuning of the Above Models

Since the attention model performed better than the Vanilla sequence to sequence model, different combinations for the values of the hyperparameters are evaluated from the below list for the attention model.

Initializations: All the hidden tensors are initialized with 0’s, the embedding layers are

initialized using random normal distribution and the linear layers are initialized from a uniform distribution.
Optimizers: SGD, Adam
Learning rate: 0.0001, 0.001
Hidden size: 256, 512, 1024
Dropout: 0.1, 0.3, 0.5
Teacher forcing: 0.1, 0.5, 0.9
Total training iterations: 75000, 100000
Batch size: 1

The combination of the hyper-parameters that gave the best translations when combination of the BLEU score and the Turing test were considered is the below:
Model: Seq2Seq with attention
Optimizers: Adam
Learning rate: 0.0001
Hidden size: 512
Dropout: 0.5
Teacher forcing: 0.5
Total training iterations: 75000
Batch size: 1

We realized that the best model in terms of the generated translations was the one from the last iteration. This was verified via the Turing test.

5 Evaluation of Model

5.1 Evaluation using Adequacy, Fluency and Post-Editability

The below scores help to perform a Turing test that a native Bahasa speaker can use to score the generated translations.

1) 1 = Easier to translate from scratch than to edit
    a. Complete Garbage.
2) 2 = Requires the same time to edit as translating from scratch
    a. Very little of the meaning of the source sentence is captured.
    b. Grammatical structures is not present in the translation.
3) 3 = Requires some editing, easier to edit than translate from scratch
    a. Some meaning of the source sentence is captured.
    b. Acceptable grammatical structure
4) 4 = Requires light editing
    a. Almost all the meaning is captured.
    b. Only light grammatical cleaning is required (Past Tense and Present Tense)
5) 5 = Requires no editing
    a. Good to Go.

5.2 Evaluation Using BLEU Score

BLEU (bilingual evaluation understudy) is a metric for evaluating the quality of text which has been machine-translated from one natural language to another.
We used NLTK's sentence_bleu API in our experiment. By default, the *sentence_bleu()* scores calculate the cumulative 4-gram BLEU score, also called BLEU-4.. The cumulative and individual 1-gram BLEU use the same weights, e.g. (1, 0, 0, 0). The 2-gram weights assign a 50% to each of 1-gram and 2-gram. The 3-gram weights are 33% for each of the 1, 2 and 3-gram scores. The 4-gram weights are 25% for each of the 1, 2, 3, 4-gram scores. The cumulative scores for BLEU-1, BLEU-2, BLEU-3 and BLEU-4 were computed and in the end the average of all these scores have been considered as an evaluation criterion

5.3 Chosen Model and Justification

The Seq2seq model with attention decoder with optimal combination of hyper parameters as found in the earlier section is chosen as the final model since it generates translations that have the highest BLEU and Turing scores. The training and validation losses along with the progression of the BLEU scores are shown below for this model per 1000 iterations of training.

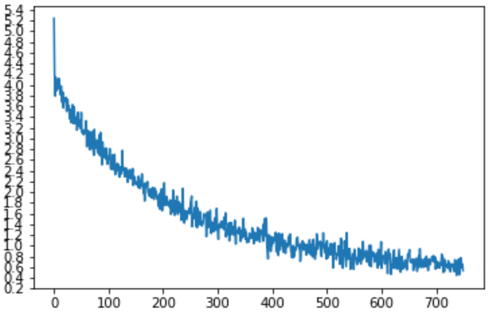


Figure 10 Training loss per 1000 iterations

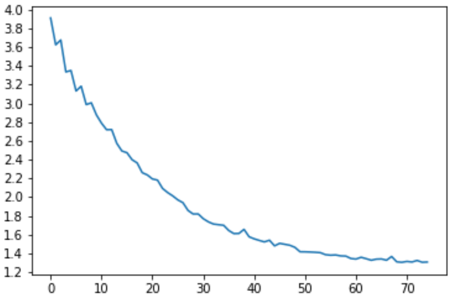


Figure 11 Validation loss per 1000 iterations

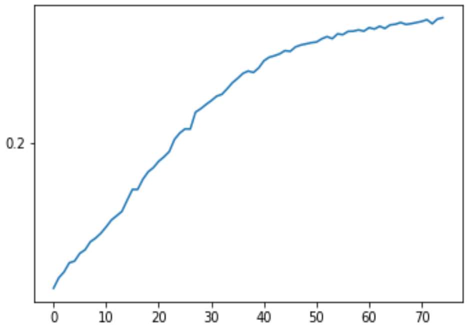


Figure 12 Bleu score per 1000 iterations

6 Lessons Learnt

Some of the lessons that were learnt:

a. Neural machine translation is a complex task and the metrics – validation loss, validation BLEU score to evaluate the model performance are not generally as indicative as accuracy in the case of CNNs for classification of images. Hence human judgement is still required for proper evaluation.

b. A large corpus of sentence pairs is required with an equal distribution of words from both the vocabularies to have a decent performance, else the models might put more focus on the words that occur more often in the vocabulary.

c. The translation task is highly sensitive to the number of words in the input English sentence. The longer the input sentence, the poorer the performance of the translation for the trained model.

7 References

[1] S. Millward. (2014). *Indonesia to be world's fourth-largest smartphone market by 2018*. Available: https://www.techinasia.com/indonesia-worlds-fourth-largest-smartphone-2018-surpass-100-million-users

[2] KPMG. (2018). *ASEAN Business Guide*. Available: https://home.kpmg/content/dam/kpmg/sg/pdf/2018/07/ASEAN-GUIDE-Indonesia.pdf

[3] M. I. Office. (2016). *Unlocking Indonesia's digital opportunity* Available: https://www.mckinsey.com/~/media/McKinsey/Locations/Asia/Indonesia/Our%20Insights/Unlocking%20Indonesias%20digital%20opportunity/Unlocking_Indonesias_digital_opportunity.ashx

[4] J. Errens. (2015). *The past, present and future of machine translation*. Available: https://www.monotype.com/resources/articles/the-past-present-and-future-of-machine-translation/

8 Examples of sentences generated by different models

8.1 Vanilla sequence to sequence model

The English and Bahasa translations of some sentences along with the evaluation score provided by a native speaker is provided in the below table. These sentences were not in the training or the validation sets. The scores are provided by a native Bahasa speaker according to the criterion introduced in section 5.

ID	English Sentence	Bahasa Sentence	Score
1	tom is playing with ball.	tom sedang dengan dengan . .	2
2	she is standing there.	dia ada di sana sana.	3
3	he is a bad man.	dia adalah pria yang .	3
4	he wants to sleep.	dia ingin menjadi .	2

5	I can’t see you crying.	aku harap kalian bertemu denganmu .	2
6	My dog is running around.	Ayah adalah besar besar .	1
7	It is very popular.	Ini sangat . .	2
8	she speaks american english to tom's father	Dia dapat mengatakan bahwa dia mungkin untuk .	2
9	Please eat lunch in the afternoon	Tolong tolong nyalakan di di . .	1
10	I see red roses in the garden	Aku harap saya adalah orang asing.	1

The average score of vanilla Seq2Seq obtained is **1.9.** The final BLEU score of vanilla Seq2Seq is 0.3308.

8.2 Sequence to sequence model with attention

The below table captures the best translations for the English sentences for the sentence pairs that are not present in the training or the validation sets.

ID	English Sentence	Result from Google	Bahasa Sentence	Score
1	tom is playing with ball.	tom bermain dengan bola.	tom bermain dengan temannya .	4
2	she is standing there.	dia berdiri di sana	dia berdiri di sana .	5
3	he is a bad man.	dia orang jahat.	dia adalah pria yang baik .	4
4	he wants to sleep.	dia ingin tidur.	dia ingin tidur .	5
5	I can’t see you crying.	Aku tidak bisa melihatmu menangis.	aku harap kali kamu . .	1
6	My dog is running around.	Anjing saya sedang berlarian.	ibuku saya meninggal .	1
7	it is very popular.	Ini sangat populer.	itu sangat . .	2
8	she speaks american english to tom's father	Dia berbicara bahasa Inggris Amerika kepada ayah Tom	dia berbicara bahasa bahasa tom tom ayah .	4
9	Please eat lunch in the afternoon	Silakan makan siang di sore hari	tolong berbicara siang siang siang .	2
10	I see red roses in the garden	Saya melihat mawar merah di kebun	saya melihat membantu mawar di di . .	3

The Average score obtained is **3.1** and BLEU score obtained is 0.339.