Experiment on a Transformer Model Indonesian-to-Sundanese Neural Machine Translation with Sundanese Speech Level Evaluation

Restu Bias Primandhika[1,*] Muhammad Nadzeri Munawar[2] Aceng Ruhendi Saifullah[1]

[1]*Universitas Pendidikan Indonesia*
[2]*GDP Labs Bandung*
[*]*Corresponding author. Email: restu@upi.edu*

ABSTRACT

Speech level is one of the essential Sundanese language elements. As Indonesian mixed within Sundanese language use, the usage of speech level is gradually degrading. Indonesian, in order to get correct word choice in Sundanese language, social contexts may refer to many sources such as a dictionary, or thesaurus. However, for better translation in syntax and context, machine translation is offered. Based on the fact, this experiment focuses on the problem when translating Indonesian to Sundanese and the evaluation of Sundanese speech level in the translated texts. Neural machine translation (NMT) was chosen as the current technology in machine translation, which worked by combining recurrent neural network encoder-decoder. The experiment started with building 50.000 Sundanese-Indonesian sentences as a parallel corpus to build and train NMT models. The experiment on sentence training in Transformer NMT without out-of-vocabulary (OOV) shows 42.72% BLEU Score, and Average Training Loss was 1.77 while for speech level was dominated by 56% *basa loma* (coarse) of the whole testing result.

Keywords: *Neural machine translation, Speech level, Sundanese*

1. INTRODUCTION

As one of the vernaculars in Indonesia, the Sundanese language is spoken by 40 million (Muamar, 2018). Most of the speakers are Indonesians that live in West Java. Sundanese speakers year by year have been gradually reduced. These are some factors why Indonesian find it hard to speak Sundanese, like their background, neutral language preference and the use of speech level. While speech level is an element that cannot be separated from Sundanese language use, ironically, it becomes a constraint for the speakers themselves. For Indonesian, to know good and right Sundanese language is essential.

The computer-based tools can be used as an effort to learn the Sundanese language and its preservation. According to the current technological developments, to overcome this problem, there are digital dictionaries and machine translation. Digital dictionaries have limitations in translating local languages into Indonesian because the approach used is to translate word for word and do not recognise which word is suitable to a context. Digital dictionaries that support Indonesia local languages can be accessed on kamusdaerah.com or kamuslengkap.com. Another alternative to dictionaries is a machine translation (Abidin, 2018).

There are some machine translators for Indonesian to Sundanese such as kamus-sunda.com, terjemahansunda.com and most popularly Google Translate. kamus-sunda.com, terjemahansunda.com use a dictionary-based, while Google translate is currently moved to their machine to Google Neural Machine Translation (GNMT). Still, there are some translation problems in the result, and there is no significant way to contribute to the context problem. Besides, GNMT uses English as a medium for language modelling that will affect Sundanese sentence structure and diction. Henceforth, we conducted this experiment. By now, we have not found yet any specific publications regarding the neural machine translation for Indonesian to Sundanese. However, from a previously done experiment by Suryani et al. (2015) on Sundanese to Indonesian SMT, we tried to get the depiction of what Sundanese translation problems are, especially in semantic. The neural machine translation (NMT) is chosen as the current technology in machine translation and based on four big wins, according to WMT (World Conference on Machine Translation); this approach is more context-aware than other machine translation model. Thus, we can assume that this experiment will initiate the project for developing context-aware Indonesian to Sundanese translation in NMT.

2. LITERATURE REVIEW

In Sundanese language translation, scientific research begins from Suryani et al. (2015) with an *Experiment on a Phrase-Based Statistical Machine Translation Using PoS Tag Information for Sundanese into Indonesian*. The statistical approach is chosen at that time because the method does not require any linguistic rules, so it is relatively faster to implement. Besides, any Sundanese language resources and tools that are ready to use could not be found. It can be concluded that the enrichment of Sundanese-Indonesian parallel corpus and its optimisation, is still needed until this time.

Another researcher that put concerns to local Indonesian languages is Abidin (2018) with his paper *Translation of Sentence Lampung-Indonesian Languages with Neural Machine Translation*. In his research, attention-based approach in machine translation is used. His findings prove that NMT can overcome the contextual meaning found in the Lampung language, such as several words which have different meanings depending on the context of the sentence or sentence in the sentence.

In terms of social context, two researchers already managed to control the level of formality or politeness at the advanced level. In English to German NMT, there is Sennrich, Haddow and Birch (2016) with *Controlling Politeness in Neural Machine Translation via Side Constraints*, and in English to Japan NMT, there is Feely, Hasler and de Gispert (2019) with *Controlling Japanese Honorifics in English-to-Japanese Neural Machine Translation*. Both use side constraints that were added to the source side of a parallel text to provide control over the politeness of translation output. Those following papers suggest that this approach can be applied towards language with honorifics. However, our Sundanese parallel corpus is still needing enrichment to continue similar research.

3. METHOD

3.1. Neural Machine Translation

Neural Machine Translation is a machine translation approach that applies an extensive artificial neural network toward predicting the likelihood of a sequence of words, often in the form of whole sentences. The first successful neural machine translation model was the seq2seq encoder-decoder model with attention (Koehn, 2020 p. 126)

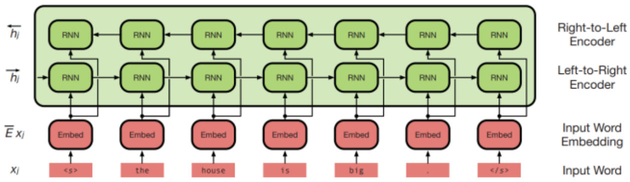


Figure 1 Neural machine translation model (input encoder)

The encoder consists of two recurrent neural networks, running right to the left and left to right (bidirectional recurrent neural network). The encoder states the combination of the two hidden states of the recurrent neural networks.

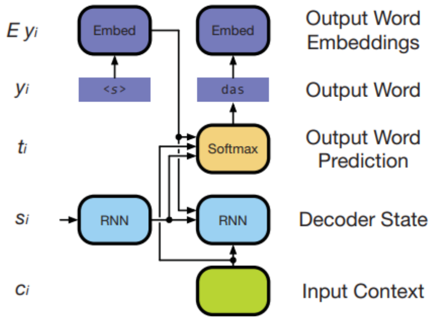


Figure 2 Neural machine translation model (output encoder). Given the context from the input sentence and the embedding of the previously selected word, new decoder states and word predictions are computed.

The advantage of the encoder-decoder architecture is that the system processes the entire input before starting to translate. This means that the decoder can use what has been generated and the entire source sentence when constructing the next word in translation.

To run our NMT, we use Google Colab Research Pro with the following specification:

- RAM: 25.51 GB
- Disk: 147.15 GB
- GPU: T4 & P100 GPU

3.1.1. Transformer Model NMT

The Transformer NMT model is an encoder-decoder model that has a self-attention mechanism. The model incorporates dependency relations into self-attention on both source and target sides. Self-Attention: Transformer (Koehn, 2020 p. 207). It considers associations between every input word and any output word and uses it to build a vector representation of the entire input sequence.

Transformer NMT model is a robust sequence-to-sequence modelling architecture capable of producing state-of-the-art neural machine translation (NMT) systems.

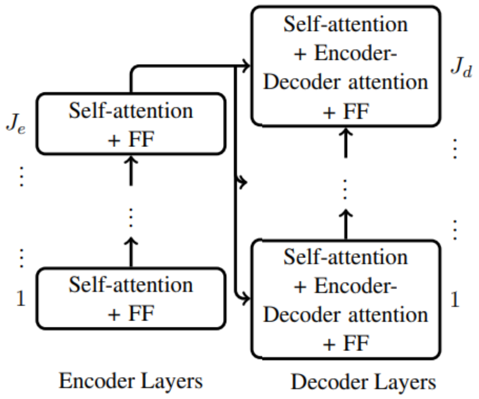


Figure 3 The outline of the Transformer NMT model

3.1.2. Transformer Model NMT

BLEU (**Bil**ingual **E**valuation **U**nderstudy) is a metric for evaluating machine-translated texts automatically. The BLEU score is a number between zero and one that measures the similarity of machine-translated text to a set of high-quality reference translations. A value of 0 means that the machine translation results do not overlap with the reference translation (low quality). In contrast, a value of 1 means that there is a complete overlap with the reference translation that implies high quality (Cloud, 2020).

$$BP = \{1\ e^{\left(1-\frac{r}{c}\right)} \qquad \begin{matrix} if\ c>r \\ if\ c\leq r \end{matrix}\ .$$

Then

$$BLEU = BP\ \cdot exp\left(\sum_{n=1}^{N} w_n\, log\ p_n\right)\cdot$$

The ranking behaviour is more immediately apparent in the log domain,

$$log\ BLEU = \left(1-\frac{r}{c},0\right) + \textstyle\sum_{n=1}^{N} w_n\, log\ p_n\ \cdot \quad (1)$$

According to Eq. (1), BP is a brevity penalty, which counters the length of the translation result. Whereas *r*, *c* and P*n* refer to the length of reference, the length of translation results, and precision-recall of each n-gram. To understand BLEU Score, we can see in Table 1 (Cloud, 2020)

Table 1. Interpretation of BLEU Scores

BLEU Score (%)	**Interpretation**
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High-quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

3.1.3. Datasets

Since there is no open-source parallel corpus for Indonesian Sundanese, we build parallel datasets for our NMT experiments. The data taken from various sources, starting from movie subtitles on Subscene, bilingual online bible and Sundanese-Indonesian websites. In these experiments, we use the standard training and test sets for each parallel corpus. The data then were pre-processed before training to clean text from noises such as spelling mistakes, slangs and abbreviations. The following example of four sentences with different levels of formality in Sundanese (in order): Neutral, *Loma*, *Sedeng, Lemes.*

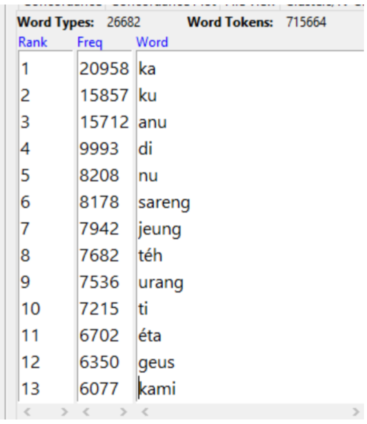
Table 2. Parallel Corpus in Indonesia-Sundanese

Indonesian	**English**	**Sundanese**
Uangnya ternyata tinggal segini	The money turns out to be this much	*Artos téh geningan nyésa sakieu*
Dia seharusnya tak memukulmu, Jenny.	He shouldn't hit you, Jenny.	*Manéhna teu sakuduna neunggeul didinya, Jenny.*
Aku tidak pernah makan barang haram ataupun najis	I have never eaten anything that is haram or unclean	*Sim kuring tara neda barang haram atanapi anu najis*
Amat sedang makan di restoran	Amat is eating at restaurant	*Amat nuju tuang di réstoran*

3.1.4. Pre-processing

After the data had been collected, there were several steps to do in order to cleanse noises on the text. Pre-processing was an essential process for text documents to undergo a suitable "cleaning" and normalisation process before they could be processed by machine algorithms (Torres-Moreno, 2014). We tried to clean misspelt or mistakenly attached words, typographic inconsistencies, ungrammatical sentences, strange characters or they are in a different coding language.

This process started with analysing the whole word in our text documents with concordance tools, *AntConc*, a freeware, multi-platform, multi-purpose corpus analysis toolkit (Anthony, 2016). We collected misspelt words in both languages to be later replaced in batches.



Word Types: 26682 Word Tokens: 715664

Rank	Freq	Word
1	20958	ka
2	15857	ku
3	15712	anu
4	9993	di
5	8208	nu
6	8178	sareng
7	7942	jeung
8	7682	téh
9	7536	urang
10	7215	ti
11	6702	éta
12	6350	geus
13	6077	kami

Figure 4 The illustrations of Sundanese words concordance in AntConc

Based on the analysis, there were in Indonesian Sundanese cases. For, we have 1) Collected words that contained acute e (é) (as Sundanese signify /e/ in reading) 2) Collected common mistake spelling in Indonesian (*ejaan tidak baku*) to be replaced each word correction according to EBI (*Ejaan Bahasa Indonesia*) 2) collected mostly abbreviated words in Indonesian and its replacement. As analysis was done, we had collected the misspelt words and their replacement. The software to support this process was *SarAnt*, a freeware batch search and replace tool. Some examples can be seen in the following table:

Table 3. Word Replacement Format in SarAnt

Elements	Example in *SarAnt* formats

Sundanese	*Acute e (é) issue*
	hade => hadé (*good*)
	mere => méré (*give*)

Indonesian	*Abbreviated words*
	yg => yang (*which*)
	dgn => dengan (*with*)
	
	Word Standardisation
	praktek => praktik (*practise*)
	nasehat => nasihat (*advice*)

3.1.5. Tokenisation

Tokenisation is the process of separating a specific text into a continuous sequence of lexical units to represent a given text. Tokenisation can take place at various granularities such as at the phrase level, word level, subword level or character level, considering the level of textual representation required for a task (Riedl & Biemann, 2018). In neural machine translation, each sentence should be correctly space-separated for each word. For example, "*Teu, tibang saukur bagja.*" should be tokenised to "*Teu, tibang saukur bagja .*". As seen in the example, comma and period are space-separated to be identified by machine as different words. To do this, we use a subword tokeniser. There are different modes for tokeniser:

- Aggressive mode is standard tokenisation but only keeps sequences of the same character type (e.g., "2,000" is tokenised to "2, 000", "soft-landing" to "soft - landing", etc.)

- Conservative mode is standard tokenization (e.g., "2,000" still keeps to "2,000", "soft-landing" still keeps to "soft - landing", etc.)

- Char mode is character tokenisation (e.g. "2,000" is tokenised to "2, 0 0 0", etc)

Due to neural machine translation requiring sentences with meaning for each word, we use subbed word tokeniser with conservative mode. An example of sub word tokeniser with conservative mode can be seen in Table 4.

Table 4. Tokenisation

Before Tokenisation	After Tokenisation	English
Abdi térang yén sagala rupi aya watesna, nanging aturan-aturan Gusti mah sampurna.	*Abdi térang yén sagala rupi aya watesna , nanging aturan-aturan Gusti mah sampurna .*	I know that everything has a limit, however; God's rule is impeccable.
Palataran sisi wétan, anu aya lawangna, rubakna 22 méter.	*Palataran sisi wétan , anu aya lawangna , rubakna 22 méter .*	The east area, where their door was, has 22 metre width.

3.1.6. Data Splitting

The train-validation-test split is a technique for evaluating the performance of machine learning algorithms. Training and testing on the same data is a methodological error: a model that will only repeat the label of the sample it has just seen will have a perfect score but fail to predict anything useful on those unseen data. To avoid this, it is common practice to split some of the available datasets.

The procedure involves taking the dataset and splitting it into three subsets. The first subset (train) used to fit the model and referred to the training dataset. The second subset (validation) is not used to train the model; instead, it is a set of examples used to adjust the parameters of the classifier to find the optimal number of hidden units or to determine stopping points. The third subset (test) is a set of examples used only to assess the performance of classifiers which are fully trained to estimate the error rate after we have selected the final model. To split the dataset into three subsets, we use cross-validation method.

3.1.7. Building Vocabulary

Building vocabulary is a process of collecting vocabulary from a parallel corpus. The result of building vocabulary is a simple text file with one token per line and ordering it according to the most frequent tokens. We collected 50000 tokens based on the most frequently occurring. The result of building vocabulary consisted of two vocabulary files from each Indonesian and Sundanese corpus. This process was needed as a reference during the training and testing process.

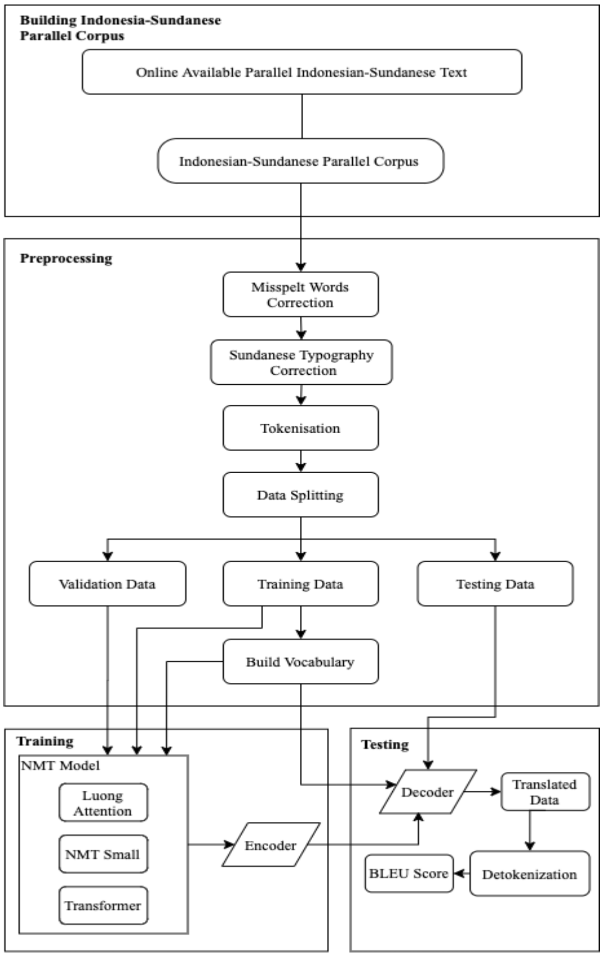


Figure 5. Overall Process Block Diagram

3*.2. Sundanese Speech Level*

This section will discuss general Sundanese language social context and speech level conversion due to its social context.

Using formality distinctions in Sundanese language is essential. The choice of words determines what speech level and is socially deictic. Levinson (1983 p. 63) portrayed social deixis as the predetermination of social differences that are according to participant-roles, central aspects of the social correlation possessed between the speaker and addressee(s) or speaker and some referent. Hence, speakers are always making a choice of what level of formality to use depending on the social context. In Sundanese, making the conversation by noticing social context means paying attention to *undak-usuk.*

According to LBSS's Sundanese General Dictionary (LBSS: *Lembaga Basa jeung Sastra Sunda*), *undak-usuk* means the speech level that is related to politeness. One that uses Sundanese language in conversation without using the speech level can be labelled as impolite. Based on the definition, Sundanese speech level plays an essential role in daily conversation to show honorifics and to value each one another in the society.

The existence of *undak usuk* is one of the constraints in using Sundanese language good and right. People tend to argue that this speech level system as a reason to use other language rather than Sundanese in conversation; as stated by in *Polemik Undak Usuk Bahasa Sunda* (Rosidi & Djiwapradja, 1987 p. 80) that Sundanese language that has the stages of speech level becomes the constraint for its native speaker to speak or to write in Sundanese. The reason is affraid of mistaken or to be blamed. "

Based on social context, Sundanese speech level (*undak usuk*) according to Faturohman (1982) is divided into three levels: 1) *Basa Loma* (coarse) 2) *Basa Sedeng* (neutral) and 3) *Basa Lemes* (refined). For instance, when speaking with friends of equal or lower social status, *bahasa loma* is used. When speaking to family, superiors, strangers or older individuals, *Sedeng* and *Lemes* are used. In this case, self-honorifics (*Basa Sedeng*) and giving honorifics to others need to be distinguished and cannot be replaced in use.

Table 5. Undak-Usuk Bahasa Sunda

Formality	Sundanese sentence
Loma	Urang keur dahar di imah (*I am eating at home)*
Sedeng	Abdi nuju neda di rorompok (*I am eating at home)*
Lemes	Sim kuring nuju tuang di bumi (*I am eating at home)*

There are several ways to apply formality from Bahasa Loma; the following are the examples of changing *Basa loma* to have formality or politeness:

Table 6. Conversion Method from Loma to Lemes

Conversion Method	Example
Substituting the suffixes (*Éngang Panungtung*)	*Kirim → Kintun (send)* *Maju → Majeng* (move forward)
Changing the vowels, (e.g. from u to i)	*Murah → Mirah (cheap)*
Inserting an infix (e.g "-in-")	*Sareng → Sinareng (along with)*
Change the whole word	*Nénjo → Ningali (see)*
Find related word	*Beuteung → Patuangan (belly)*
Find the semantically equivalent word in other languages	*Peuting → wengi* (Jv). (night)

4. FINDINGS AND DISCUSSION

The main goal of this experiment is to determine which translation model gives the best result and to identify the issue of social context in Indonesian to Sundanese text translation. Besides that, speech level analysis is also one of the parameters to be measured.

4.1. Experiment Scenario

In this research, we used three scenarios for different types of NMT models, Luong Attention (Luong, Pham, & Manning, 2015), NMT Small (Kreutzer, Bastings, & Riezler, 2019), and Transformer (Vaswani et al., 2017). Learning rate, training loss, and testing BLEU score were used for choosing the best NMT model for case Indonesia-Sundanese translation.

As shown in Table 7, for each corpus, 51027 sentences were split into three parts, training data with 43627 sentences, validation data with 4848 sentences, and testing data with 2552 sentences.

Table 7. Number of Sentences used for Experiment

Data Types	∑ Sentences
Training Data	43627
Validation Data	4848
Testing Data	2552

4.2. Experiment Results and Analysis

As shown in Fig. 6 and table VIII, the learning rate for the Transformer model increases significantly compared to Luong Attention and NMT Small. It shows that the Transformer model has a reasonable learning rate for the case of Indonesia-Sundanese translation. As shown in Fig. 7 and table IX, the Transformer model decreases significantly compared to the others. It shows that the Transformer model aligns with the training goal to decrease training loss for each step.

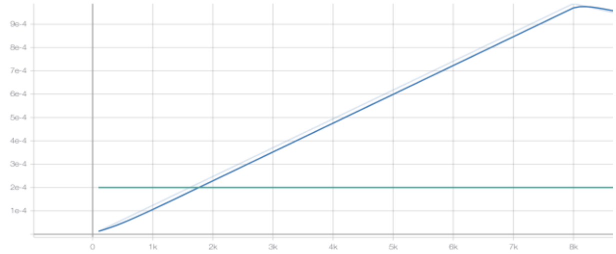


Figure 6. The graphics of the learning rate between 3 models. Luong Attention (green line), NMT Small (light blue line), and Transformer (blue line)

Table 8. Learning Rate Comparison

Model	Average Learning Rate
Luong Attention	0.0002
NMT Small	0.0002
Transformer	0.0005

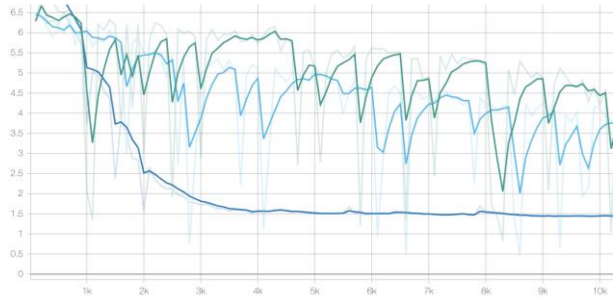


Figure 7. The graphics of training loss between 3 models. Luong Attention (green line), NMT Small (light blue line), and Transformer (blue line)

Table 9. Training Loss Comparison

Model	Average Training Loss
Luong Attention	4.57
NMT Small	4.15
Transformer	1.77

Table 10. BLEU Score Comparison

Model	BLEU Score
Luong Attention	26.32 %
NMT Small	36.22 %
Transformer	42.72 %

4.3. Indonesian-Sundanese Social Context Translation

When translating from Indonesia into Sundanese, a translator must determine one level of formality or another. This raises a challenge for Indonesian Sundanese NMT, for a translation result that needs to be adequate and fluent (Koehn, 2020). It needs to both capture the significance of the source sentence and utilise the best possible degree of custom. In machine interpretation from a language with fewer honorifics (for example, Indonesian), it is hard to predict the suitable honorific. However, users might need to control the degree of politeness in the output.

Sundanese social context affects not only the verb, like English tenses but also subject, pronoun and even the noun, as illustrated previously in Table VI. Based on that, NMT translation results should suggest various diction in a sentence to satisfy the Sundanese speech level. This means more training datasets are needed. The translation result should be controlled by users in selecting the desired translation output. In this case, we are going to evaluate the translation result based on our NMT model.

As shown in Table X, the Transformer model has 42.72 % of BLEU score that indicates the best result compared to Luong Attention and NMT Small. Based on Table I, that BLUE score indicates high-quality translation. From the transformer model, the result text was taken to be tokenised in order to measure the percentage of speech level. Accordingly, we refer to the digitised *Kamus Undak Usuk Bahasa Sunda* in which each word is labelled by 1) *loma* 2) *lemes* 3) *sedeng 4) kasar* excluding neutral words such as words with no speech level conjunctions or nouns. The following table shows a summary of the most frequent word appears based on the translated text:

Table 11. The Summary of Tokenized Words

Freq.	words	level
598	*Sareng (along with)*	*Sedeng (neutral)*
183	*Jeung (along with)*	*Loma (coarse)*
164	*Abdi (I)*	*Sedeng (neutral)*
153	*Bakal (will)*	*Loma (coarse)*
120	*Tuluy (continously)*	*Loma (coarse)*
	etc.	

Then, the tokenised word is put into a percentage to see overall speech level depiction in the translated text.

Table 12. The Percentage of Speech Level

Level	words	%
Kasar (very coarse)	44	1%
Lemes (refined)	451	10%
Sedeng (neutral)	1442	33%
Loma (coarse)	2476	56%

4.3.1. Translation Tendency

As *Basa Loma* dominates the language model, this affects a translation result. We tried to translate 30 simple sentences without out-of-vocabulary (OOV). The result shows there was speech level change mostly in T/V distinction, which is an honorific for which both the referent and the target of the expression of relative social status (Levinson, 1983 p. 90).

Table 13. T/V Distinction Issue

Human Translation on Training Data	NMT result	English
[1]*Sim kuring ménta dihapunten (Sedeng)*	Urang ménta *dihampura (Loma)*	I beg your pardon
[2]*Hidep teu cocok jeung anak Ki Lurah (Loma)*	Anjeun teu cocok sareng anak Ki Lurah. (Sedeng)	You don't get along with that Ki Lurah's son/daughter.

The first example is the translation of the Indonesian sentence "*Aku minta maaf*" (Eng: I beg your pardon) that should have a subject "*Sim Kuring*" (I) in Sundanese but resulting *urang* instead. This makes the sentence change its speech level from *sedeng (neutral)* to *loma (coarse).*

There is another case when *loma* was converted into *sedeng*. The second example is a translation from "*Kau tidak cocok dengan anak Ki Lurah*". As can be seen in the table, word "*hidep" (coarse)* turned into "*anjeun*" (neutral). The word "*hidep*" had low probability to be appeared in the (*hidep* appears 0.06% from language model while *anjeun* 0,38%). As a replacement, the NMT translated the word 'Kau' (you) in Indonesia to "Anjeun".

5. CONCLUSION

Before going to the context issue, this experiment started with language modelling three times, by adding more sentences in parallel text gradually and doing some optimisation. This was done to enhance the quality of syntax translation beforehand. This is shown by the

Transformer's BLEU score of 42.72 % and Average Training Loss 1.77. Accordingly, we found that the number of parallel corpus affects the quality of translation results. Transformer Model NMT is better at the syntactic problem but still lacking for context-aware machine translation. As the analysis had done, the translation result generated most tends to be in *Basa Loma (coarse).* Even though the text in the target side corpus needs to be balanced; otherwise, there will be bias in language level. We suggest that this kind of experiment can be done with other translation model and more parallel corpus as a training data.

REFERENCES

Abidin, Z. (2018). Translation of sentence Lampung-Indonesian languages with neural machine translation attention based approach. *Inovasi Pembangunan: Jurnal Kelitbangan*, *6*(2), 191–206. doi: 10.35450/jip.v6i02.97

Anthony, L. (2016). AntConc [computer software]. Retrieved from http://www.laurenceanthony.net/software/antconc/

Cloud, G. (2020). *Evaluating models | automl translation documentation | Google Cloud*. Retrieved from https://cloud.google.com/translate/automl/docs/evaluate#interpretation

Faturohman, T. (1982). *Tata basa sunda*. Bandung: Jatnika.

Feely, W., Hasler, E., & de Gispert, A. (2019). Controlling Japanese honorifics in English-to-Japanese neural machine translation. *Proceedings of the 6th Workshop on Asian Translation*, 45–53. doi: 10.18653/v1/D19-5203

Koehn, P. (2020). *Neural machine translation*. Cambridge: Cambridge University Press.

Kreutzer, J., Bastings, J., & Riezler, S. (2019). Joey NMT: A minimalist NMT toolkit for novices. *Computation and Language (cs.CL)*, *1*(1), 1-17.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Computation and Language (cs.CL)*, *1*(1), 1-11.

Muamar, A. (2018). Mempertahankan eksistensi bahasa sunda. *Pikiran Rakyat*, 27.

Riedl, M., & Biemann, C. (2018). Using semantics for granularities of tokenisation. *Computational Linguistics*, *44*(3), 483–524. doi: 10.1162/coli_a_00325

Rosidi, A., & Djiwapradja, D. (1987). *Polemik undak usuk basa Sunda*. Bandung: Mangle Panglipur.

Sennrich, R., Haddow, B., & Birch, A. (2016). Controlling politeness in neural machine translation via side constraints. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 35–40. doi: 10.18653/v1/N16-1005

Suryani, A. A., Widyantoro, D. H., Purwarianti, A., & Sudaryat, Y. (2015). Experiment on a phrase-based statistical machine translation using pos tag information for Sundanese into Indonesian. *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, 1–6. doi: 10.1109/ICITSI.2015.7437678

Torres-Moreno, J.-M. (2014). *Automatic text summarisation*. California: John Wiley & Sons.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.