Indonesian Journal of Electrical Engineering and Computer Science
Vol. 25, No. 3, March 2022, pp. 1803~1813
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v25.i3.pp1803-1813

An empirical evaluation of phrase-based statistical machine translation for Indonesia slang-word translator

Kyrie Cettyara Eleison, Sari Uli Inggrid Hutahaean, Sarah Christine Tampubolon, Teamsar Muliadi Panggabean, Ike Fitriyaningsih
Department of Information Technology, Faculty of Informatics and Electronic Engineering (FITE), Del Institute of Technology,
Toba Samosir, Indonesia

Article Info

Article history:

Received Aug 19, 2021
Revised Dec 24, 2021
Accepted Jan 17, 2022

Keywords:

BLEU score
Language
Machine translation
Phrase-based statistical machine translation
Questionnaire

ABSTRACT

The use of slang (non-standard language), especially in social media, is increasing. It causes reducing the level of understanding when communicating because not everyone understands slang (non-standard language). The purpose of this work is to develop a slang-word translator. The other objective is to find the minimum number of sentences and BiLingual Evaluation Understudy (BLEU) score used as a benchmark to determine that the translation is understandable. The approach used in this project is a phrase-based statistical machine translation (PBSMT) approach, suitable for low resource language, with a dataset of 100,000 sentences taken from the comments column of several online political news portals. The comments are then manually translated to produce a parallel corpus of non-standard language-standard language. The sample sentences are taken from the dataset then distributed using questionnaires to obtain the human understanding level regarding the translation result. The result of the implementation is a BLEU score of 64 and the minimum number of sentences to have an understandable machine translation is 500. The conclusion drawn from the distributed questionnaires is that humans can understand the sentences produced by the translation machine.

Corresponding Author:

Kyrie Cettyara Eleison
Department of Information Technology, Faculty of Informatics and Electronic Engineering (FITE)
Del Institute of Technology
P.I. Del Sitoluama St., Lagu Boti, Toba Samosir, Indonesia
Email: info@del.ac.id

1. INTRODUCTION

Language is the symbol used to express one's ideas, thoughts, and feelings to others [1]. As a form of conveying ideas and feelings, language continues to develop from generation to generation. Language development is influenced by language absorption and environmental factors. One result of language development is slang (non-standard language). Non-standard language is an informal language, is usually only known in certain social circles, and used among teenagers. Slang is an everyday language that modified in many ways [2].

Slang (non-standard) words usage, especially on social media, reduces the understanding level when communicating because not everyone understands slang (non-standard) words. Due to slang, the scattered information has a lot of noise that hinders the translation process and the text processing because the slang words are unrecognizable by the translation machine database [3]. The types of noise that occur are abbreviations, typos, and slang [4].

In previous research Sebastian and Nugraha [4], The test data taken from the Instagram comments column belonging to several users. Furthermore, the test data then tokenized. The words listed in the online Kamus Besar Bahasa Indonesia (KBBI) are labeled as formal words. Meanwhile, for non-standard words, crowdsourcing labeling will be carried out using weighted majority voting to obtain the standard form of the non-standard words.

Another research Pennell and Liu [5] conducted normalization of short message text (SMS). The corpus is built using status from twitter.com. The research is done with 2 phases of approach, character-level MT and language model (LM). The first phase, machine translation model is trained in the character level. The next phase, they decode the hypothesis using a language model. By combining these two phases, it will produce a translation model that is close to the best translation.

In this work we implemented a phrase-based statistical machine translation approach to develop a machine translation. Phrase-based statistical machine translation, a supervised machine learning type [6], is a corpus-based machine translation approach [7] that divides the given sentence into several phrases (phrase-level), then the translation process will be carried out on those phrases. We created the corpus by doing manual translation to the collected data from comment section in online news portal. Phrases that have been translated will be rearranged to become the result of the translation of the given sentence. The advantage of this approach is that it performs better than neural machine translation when the data has a lot of noise [8] and when the used language is low-resource [9].

The machine translation [10] is expected to be able to translate slang (non-standard) into standard language on the political sphere. The other objective of this work is to determine the minimum number of sentences and BLEU score that is used as a benchmark to show that the translation results are understandable. Also, to determine the human understanding level of the translation results produced by the machine translation model.

This journal consists of 4 main sections. The introduction section, explains the purpose and previous research. The research method section, describes the series of research methods used. The results and discussion section, discuss the results of the experiments carried out. The conclusion sections, contains the conclusions from the experiments that have been carried out.

2. RESEARCH METHOD

The main objective of this research is to develop the corpora and to use it to create a non-standard language to standard language machine translation model with Moses. Moses is a statistical machine translation system that can be used to automatically train translation model [11]. In the process, there is a training process that takes parallel data and uses coocurrences of words and segments (known as phrases) to conclude the translation correspondence between the two languages of interest. In phrase-based machine translation, these correspondences are simply between continuous sequences of words, whereas in hierarchical phrase-based machine translation or syntax-based translation, more structure is added to the correspondences. Moses also implements an extension of phrase-based machine translation know as factored translation which enables extra linguistic information to be added to a phrase-based systems.

The two main components in Moses are the training pipeline and the decoder. There are also a variety of contributed tools and utilities. The training pipeline is really a collection of tools (mainly written in perl, with some in C++) which take the raw data (parallel and monolingual) and turn it into a machine translation model. The decoder is a single C++ application which, given a trained machine translation model and a source sentence, will translate the source sentence into the target language.

The proposed method divided into five parts; the first part is developing the non-standard languages and standard languages corpora; the second part is doing data preprocessing in Moses; the third part is performing the training process to the corpus provided; the fourth part is tuning processing; the fifth part is Evaluation processing to get the BLEU score. After the last part, the evaluation process, then the machine translation model is consumed into the user interface. The translation result is then also reviewed by respondents with a questionnaire provided. This questionnaire result will show the level of understanding of the translation results.

2.1. Developing non-standard languages to standard languages translation model

Figure 1 shows the system design. The dataset was scraped from the comment section of online news media in the politics section. The dataset was then preprocessed, trained, and tuned. The end result is a model machine translation that can be used to translate a non-formal Indonesian language to a formal Indonesian language. The next stage is evaluation, in this stage we provided a different dataset as a reference sentences to produce the BLEU score.

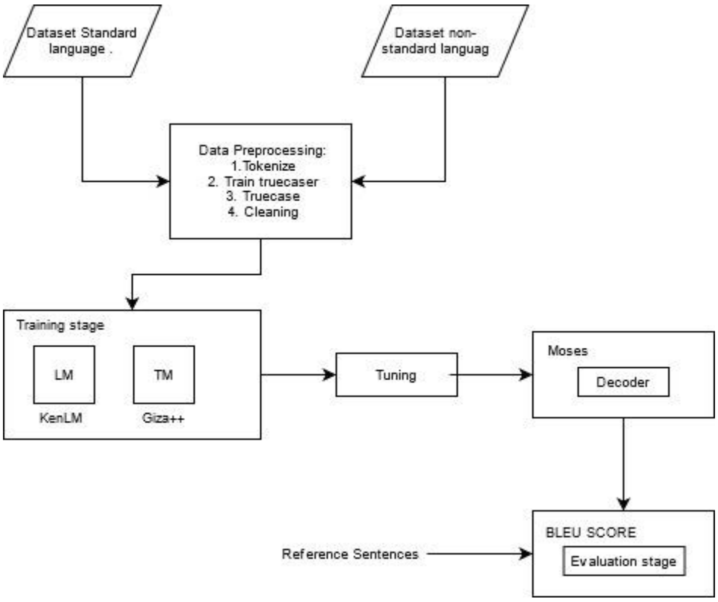


Figure 1. System design

2.1.1. Dataset

This research requires materials in non-standard form sentences that already have equivalent meanings in the standard form. The sentences used in this study are taken from online news media such as detik.com, kompas.com, and cnnindonesia.com. The data are taken from the comments column that comes from the news with political categories contained on these sites. However, the data is still in the form of non-standard sentences, so that further processing is needed so that the sentences have pairs in the form of standard sentences. Sentences of non-standard form and their translations into the standard form were made manually by correcting sentences due to the lack of materials. The dataset does not have any long sentences with the average length of sentences in the training dataset is 7.51 because long sentences can lead to a not good translation result [12]. This condition is also affect the result of the BLEU score [13].

We put the comments into an excel file which is then manually translated into a standard language. In this process, sentences with words that are not in the Kamus Besar Bahasa Indonesia (KBBI) replace with words or sentences that have the same or similar meaning. The results of these activities are non-standard sentences-standard sentences parallel datasets. Every non-standard sentence form has a pair in the standard sentence form. The process can be seen in Figure 2.

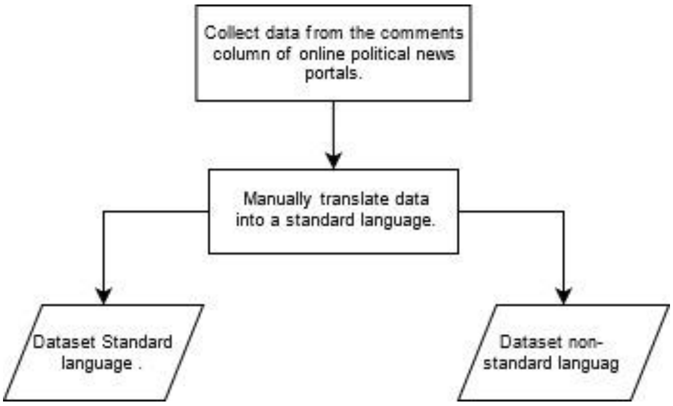

Collect data from the comments column of online political news portals.
Manually translate data into a standard language.
Dataset Standard language .
Dataset non-standard languag

Figure 2. Dataset development

2.1.2. Data preprocessing

The sentences in the dataset have many variations. This condition affects the corpora manufacturing process. The preprocess itself has four steps to do. The steps are written [14]:

- Tokenize: at the tokenize stage, the gap between words as well as the gap between words and punctuation is given.

- Train truecaser: At this stage, training is carried out on the truecaser model to extract statistics from the input file used. The result of this stage is the comparison of words whose initial letters use capital letters with words that do not use capital letters.
- True case: This stage utilizes the results of the true caser train stage. True case functions to convert each word into non-capital letters. The output of this stage is the sentences in which each word will have a capital letter and a non-capital letter that is determined based on the results of the truecaser train.
- Cleaning: Long and empty sentences will be removed because it can cause problems in the training pipeline, misaligned sentences will also be removed. At this stage, the allowed sentence length is not more than 80 words.

2.1.3. Training

KenLM will be used to build the n-gram language model [15] in both source and target domain. The key elements in language model are the probabilities of the word sequences written as *P(w₁, w₂,…,wₙ)*. The probability in n-gram language model can be calculated from the sum of n-gram frequencies:

$$P\left(w_i|w_{i-(n-1)},\dots,w_{i-1}\right)=\frac{\text{Count}\quad(w_{i-(n-1)},w_{i-1},\dots,w_i\ )}{\text{Count}\left(w_{i-(n-1)},w_{i-1}\right)}. \tag{1}$$

The following is an example of n-gram language model:
1. Unigram (1-gram): $P(W_1), P(W_2), \dots, P(W_n)$
2. Bigram (2-gram): $P(W_1), P(W_2|W_1), P(W_n|W_{(n-1)})$
3. Trigram (3-gram) : $P(W_{1,n}) = P(W_1)P(W_2|W_1)\dots P(W_n|W_{(n-2,n-1)})$

This process will generate a file in ARPA format. The ARPA file will contain probabilities and weights of the back-offs for each n-gram. Then the ARPA will be converted into a BLM file with the aim that the file can be processed faster. The translation model is used to pair the input text in the source language with the output text in the target language. Translation model built with tools Giza++ [16]. The translation modeling process by Giza++ produces a vocabulary corpus document, word alignment, and a phrase table [17]. The translation modeling process by Giza++ will also produce a translation model table consisting of a word table containing a set of words that matched between the source language and the target language with probability values.

2.1.4. Tuning

The tuning stage is done in Moses using minimum error rate training (MERT). MERT is a method that attempt to optimize the parameter of the model while considering a more complex evaluation than simply counting incorrect translation and attempt to train the model based on the method that will be used to evaluate the model [18]. The tuning process is done to develop better translation model than the one created on the training part [19]. The goal of MERT is to find a minimum error rate count on a representative corpus f_1^S with given translations $\hat{e}_1^S$ and a set of K different candidate translations $\mathbf{C}_s=\left\{\mathbf{e}_{s,1},\dots,\mathbf{e}_{s,K}\right\}$. To achieve the goal, MERT use this optimization criterion on the process:

$$\hat{e}(f_s;\lambda_1^M)=\underset{e\in C}{\operatorname{argmax}}\{\textstyle\sum_{m=1}^M\ \lambda_m h_m(e\mid f_s)\} \tag{2}$$

But the stated equation is not easy to handle because argmax operation is not possible to compute gradients and because there are many local optima of the objective function. To be able to compute gradients and smoothing the objective function, the following optimization criterion can be used. This optimization is a smoothed error count with parameter α to adjust the smoothness.

$$\hat{\lambda}_1^M=\underset{\lambda_1^M}{\operatorname{argmin}}\left\{\sum_{s,k}\ E\left(\mathbf{e}_{s,k}\right)\frac{p\left(\mathbf{e}_{s,k}|\mathbf{f}\right)^\alpha}{\sum_k p\left(\mathbf{e}_{s,k}|\mathbf{f}\right)^\alpha}\right\} \tag{3}$$

The smoothed error count is more stable and has fewer local optima than the unsmoothed one. And the result obtained by the smoothed equation does not significantly differ from the result of the unsmoothed equation of error count. Tuning requires a small amount of parallel data, separate from the training data. For tuning, the source language and target language tuning corpus are used, but tokenization and truecase are done first. The result of this tuning process is the MERT-WORK directory which contains a set of files. Inside this folder is the file moses.ini which is a configuration file for the decoder that has a number of default parameter settings.

2.1.5. Evaluation

Testing the translation results is done utilizing automatic testing of the machine translator. Automatic testing of the machine translator produces an output in the form of an accuracy value generated by Bilingual Evaluation Understudy (BLEU) [20]. This stage will use a different test set or parallel data set from the training data or tuning data. To run this testing process, tokenization and true case are first performed on the evaluation test set.

In the step-in automatic testing, the corpus to be tested first goes through an automatic translation step which will provide output in the form of a corpus in the target language that has been translated by the machine. After making the output in the form of automatic translation results from the machine translator, the next step is to get a score from the output by comparing the output with the target language manual corpus that has been made previously. BLEU will measures the modified n-gram precision score between automatic translation and reference translation and uses a constant called brevity penalty.

The BLEU value is obtained from the product of the brevity penalty with the geometric mean of the modified precision score. The higher the BLEU value, the more accurate the reference is. The value of Pn is in the range of 0 to 1. A translation will reach a value of 1 if the translation is identical to the reference translation. Therefore, even with human translation, it is not possible to produce a value of 1. To produce a high BLEU value, the length of the translated sentence must be close to the length of the reference sentence, and the translated sentence must have the same word and order as the reference sentence. The BLEU formula is as follows [13]:

$$\text{Brevity Penalty } = \begin{cases} 1 & \text{if } c > r \\ e^{\left(1-\frac{r}{c}\right)} & \text{if } c \leq r \end{cases} \tag{4}$$

$$p_n = \frac{\sum_{C\in\{\text{ Candidates }\}} \sum_{n-\text{gram}\in C} \text{ Count}_{\text{clip}}\,(n\text{ -gram })}{\sum_{C'\in\{\text{ Candidates }\}} \sum_{n-\text{gram}'\in C'} \text{ Count }(n-\text{ gram }')} \tag{5}$$

$$\text{BLEU} = \text{BP} \cdot \exp\,\left(\textstyle\sum_{n=1}^{N} w_n \log\, p_n\right) \tag{6}$$

where,

BP = brevity penalty
c = length of the candidate translation
r = the reference corpus length
Pn = modified precision score
wn = 1/N (the standard value for BLEU is 4)

2.2. Consuming the translation model into user interface

After we evaluate the machine translation model, the model is then integrated with the designed user interface. The purpose is so the users could access the translation model by using a web-based application. The translation model, which can translate non-standard languages into standard languages, is then integrated with the user interface using Python-based FlaskAPI [21]. Figure 3 shows the design interface of web application.

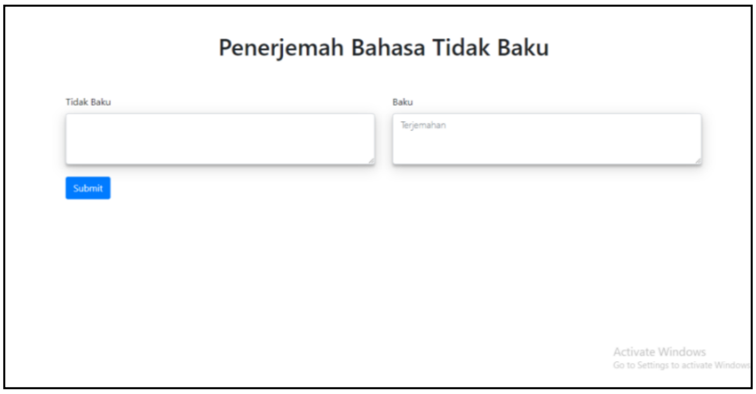
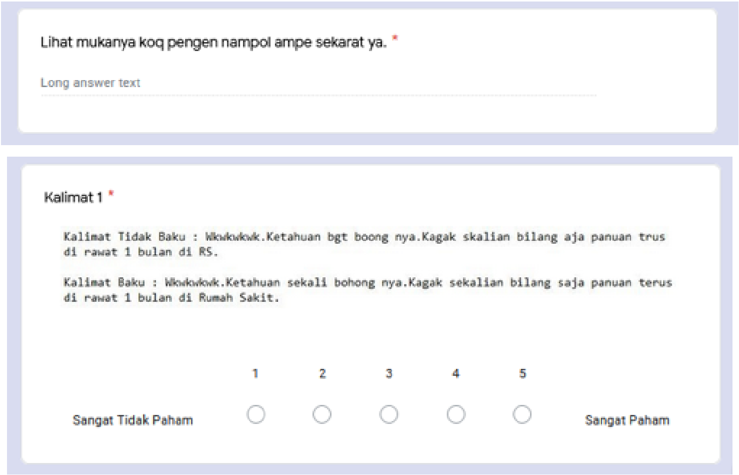

Penerjemah Bahasa Tidak Baku
Tidak Baku
Baku
Terjemahan
Submit
Activate Windows
Go to Settings to activate Windows

Figure 3. Web application design

2.3. Providing questionnaire for respondents

The purpose of distributing questionnaires is to find out the level of human understanding of the translation results obtained by the translation model, which is based on the previously calculated BLEU score from the evaluation stage. Figures 4 and 5 are the designs of prepared questionnaires. Filling out the questionnaire have been done by taking samples from the population of Del Institute of Technology (Students, Staff, and Lecturers). Based on existing data, it is known that there are 146 lecturers, 173 staff, and 1,473 students. By comparing the number of each entity with the total, we get the percentage of 8% lecturers, 10% staff, and 82% students. Based on these percentages, a sample of 4 lecturers, 5 staff, and 41 students were taken as questionnaire respondents.



Lihat mukanya koq pengen nampol ampe sekarat ya. *

Long answer text

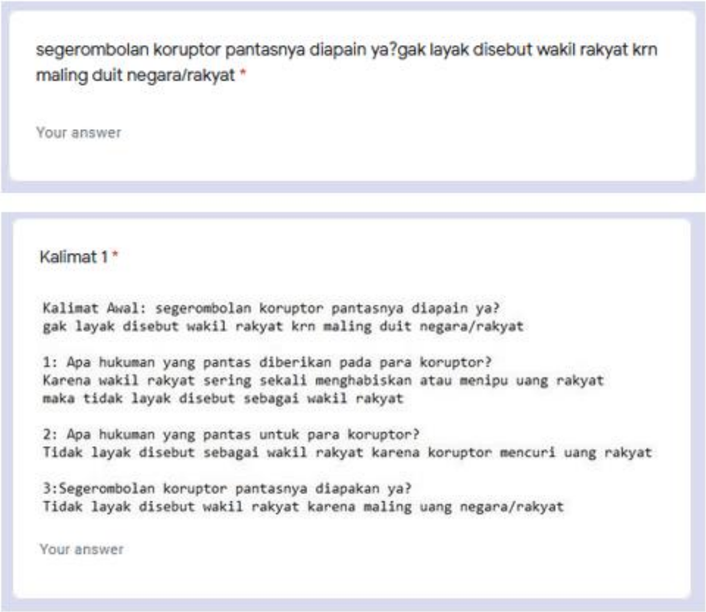
Kalimat 1 *

Kalimat Tidak Baku : Wkwkwkwk.Ketahuan bgt boong nya.Kagak skalian bilang aja panuan trus di rawat 1 bulan di RS.

Kalimat Baku : Wkwkwkwk.Ketahuan sekali bohong nya.Kagak sekalian bilang saja panuan terus di rawat 1 bulan di Rumah Sakit.

	1	2	3	4	5	
Sangat Tidak Paham	○	○	○	○	○	Sangat Paham

Figure 4. Questionnaire design for general respondents



segerombolan koruptor pantasnya diapain ya?gak layak disebut wakil rakyat krn maling duit negara/rakyat *

Your answer

Kalimat 1 *

Kalimat Awal: segerombolan koruptor pantasnya diapain ya?
gak layak disebut wakil rakyat krn maling duit negara/rakyat

1: Apa hukuman yang pantas diberikan pada para koruptor?
Karena wakil rakyat sering sekali menghabiskan atau menipu uang rakyat
maka tidak layak disebut sebagai wakil rakyat

2: Apa hukuman yang pantas untuk para koruptor?
Tidak layak disebut sebagai wakil rakyat karena koruptor mencuri uang rakyat

3:Segerombolan koruptor pantasnya diapakan ya?
Tidak layak disebut wakil rakyat karena maling uang negara/rakyat

Your answer

Figure 5. Questionnaire design for linguist

Questionnaires were distributed to 50 predetermined respondents. Because of the large amount of testing data, in this work the number of sample was obtained by using Slovin formula which has the ability to obtain a smaller number of samples but represents the entire population. Therefore, the number of sample sentences used is 100 sentences. The sample was obtained using the Slovin formula with a population of

20,000 sentences taken from testing data with a margin of error of 10% [22]. The sentences was taken using random sampling [23]. Each respondent was expected to fill in 2 types of questions. In the first type, respondents are expected to enter a translation of 4 non-standard sentences given. The second type, which is a Likert scale type [24], respondents are expected to choose the level of understanding of the four translations presented. One questionnaire is filled by two respondents. So, each questionnaire could have two different perspectives. Also, each questionnaire contains four translation results so it will not be too many and respondents could fill the questionnaire thoughtfully.

There is also a special questionnaire for linguists which contains 10 sentences with the lowest level of understanding based on the results of the general questionnaire. This questionnaire has 2 types of questions. The first type is in the form of stuffing, linguists are asked to translate the non-standard sentences given. The second type is also an entry, linguists are asked to sort sentences that are considered the most standard to the least standard.

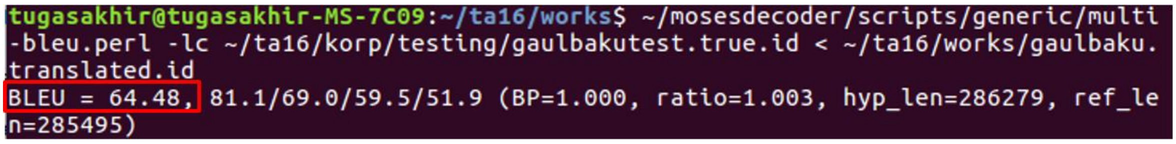
3. RESULTS AND DISCUSSION

In this section, the results of the research are explained and at the same time, a comprehensive discussion is given. The sub-section will contain the results and discussion of the machine translation model using 100000 datasets, the distributed questionnaires, the experiment using different corpus, and the experiment using different mean lengths of sentences. The experiment was done using previously collected dataset. The dataset in this work does not include any long sentences because long sentences can cause problem [6].

3.1. Machine translation using 100000 dataset

In this section, we will discuss the result of machine translation model that have been made using 100000 dataset which are divided into 60000 training data, 20000 tuning data, and 20000 testing data. The average length of sentences in the training dataset is 7.51. The machine translation model is generated through the language modeling stage until the tuning stage as described in previous chapter. This machine translation model is built using a non-standard language corpus and a standard language corpus that has been prepared previously, so that the model can translate from non-standard language into standard language.

Using the translation machine model that has been built, then a BLEU score calculation is carried out to determine the closeness between the machine translation results and human translation understanding. The BLEU score obtained is 64.48, which means that the translation results have great quality [25]. The obtained BLEU score can be seen on Figure 6. The machine translation model that has been able to translate non-standard languages into standard languages is then connected to the application interface using the Python-based FlaskAPI. Figure 7 is the interface of Slang-word translator application.



```
tugasakhir@tugasakhir-MS-7C09:~/ta16/works$ ~/mosesdecoder/scripts/generic/multi
-bleu.perl -lc ~/ta16/korp/testing/gaulbakutest.true.id < ~/ta16/works/gaulbaku.
translated.id
BLEU = 64.48, 81.1/69.0/59.5/51.9 (BP=1.000, ratio=1.003, hyp_len=286279, ref_le
n=285495)
```

Figure 6. BLEU Score result

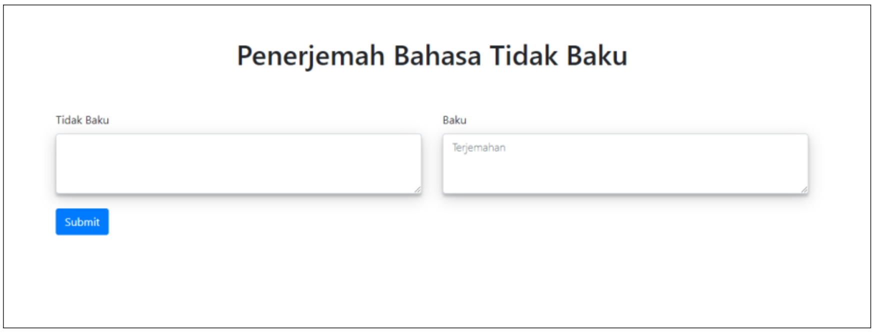


Figure 7. Application user interface

The interface formed by two text-box and one button. Users can use the slang-word translator application with phrase-based statistical machine translation approach by typing non-standard sentences in the text area of non-standard sentences (left text box) and pressing the 'Translate' button. Then the user input of non-standard sentences will be translated into standard languages by the machine translation model that has been built before.

3.1.1. Result of questionnaire

In this sub-chapter, we will discuss the results of the general questionnaire which was distributed to 50 respondents. Each respondent is expected to give an opinion on 4 sentences in one questionnaire. There are 25 questionnaires provided and 1 questionnaire filled out by 2 respondents with the aim of having different assumptions in each sentence. So, each questionnaire will get 2 responses for 1 sentence. Responses to the survey on the level of human understanding of the results of machine translation can be seen in Table 1.

Table 1. Questionnaire responses

Level of Understanding	Responses
Not understand at all	10
Do not understand	28
Quite understand	40
Understand	69
Understand very well	53
Total	200

Based on the results of the questionnaire distributed there are five levels of understanding, namely do not understand at all, do not understand, quite understand, understand, and understand very well. At the level of understanding do not understand at all obtained as many as 10 opinions. At the level of understanding do not understand obtained as many as 28 opinions. At the level of understanding quite understand obtained as many as 40 opinions. At the level of understanding understand obtained as many as 69 and at the level of understanding very understand very well obtained as many as 53 opinions. It can be concluded that the majority of machine-translated sentences can be understood by the respondents.

Figure 8 illustrates the level of human understanding of the given machine translated sentences. This diagram shows that the largest percentage is the opinions with the level of understanding ‘understand’ and the smallest percentage is the opinions with the level of understanding ‘do not understand at all’. We calculated the BLEU score of the translated sentences input by the general respondents and the translated sentences from the machine translation. The BLEU value is obtained using a python-based program. The translated sentences input by the general respondents produced a BLEU score of 26.6. The translated sentences from the machine translation produced a BLEU score of 54.48. Based on the two BLEU scores, it can be concluded that the results of machine translation are better than the results of translations provided by the general public. In addition, based on the results of the linguist's questionnaire containing 10 sentences with the lowest comprehension score, it can be concluded that the machine translated sentences cannot be understood, in line with the results obtained from the general questionnaire. The machine-translated sentences with a low level of understanding were then compared with the translations of linguists and it was found that the machine was still unable to translate non-standard sentences that had very irregular punctuation marks and sentences with OOV (Out of Vocabulary), which are sentences containing words which are not included in the corpus.

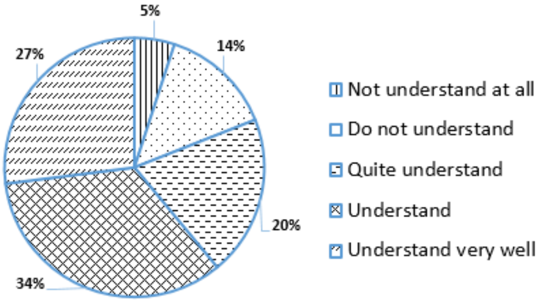


Figure 8. Pie-chart of understanding level

3.2. Experiment using different corpus quantity

By using the same steps, we created 7 machine translation models different corpus quantity. Each machine translation model produced different BLEU scores. The results of the experiment can be seen in Figure 9. Based on the results, we concluded that by using 500 datasets, good BLEU score result is obtained. We reckon this happened because the dataset is taken only in the political category so that the vocabulary contained in the corpus is less diverse. Assessment is taken based on the results of the BLEU score obtained from each test.

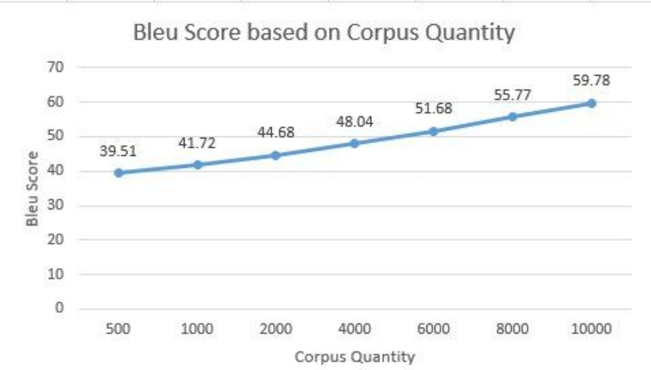


Figure 9. BLEU Score based on corpus quantity

3.3. Experiment using different sentence lengths

By using the same steps then 4 machine translation models were built using a different sentence length. Each machine translation model produced different BLEU scores. The results of the translation result on the corpus quantity can be seen in Figure 10. Based on the results of the tests that have been carried out, it was found that the sentence length in the corpus affects the resulting translation results. Model A and Model B produced a low BLEU score because the sentence is too-short which affects the probability of each word. Model D produces a better BLEU score than Model A and Model B, but not better than Model C. This is due to the too-long sentences contained in the corpus in Model D.

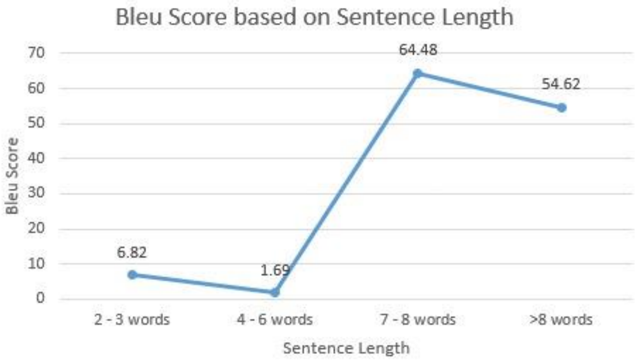


Figure 10. BLEU score based on sentence length

4. CONCLUSION

Statistical machine translation approach is a method used to implement the process of translating non-standard Indonesian into standard Indonesian sentences in the political sphere. This model is built using 100,000 pairs of non-standard sentences-standard sentences that have been prepared manually by humans and the results of the BLEU parameter produce a human understandable value of 64.48. We tested that 500 sentence pairs could produced a machine translation with a BLEU score of 39.51 which can still be understood by humans. We also tested the different sentence length and found that a too-long and too-short sentences produced a low BLEU score.

The results of the translation produced by machine translation are then analyzed by distributing questionnaires to 50 respondents to determine human understanding of the translation results. From the questionnaire, it is known that human understanding of the translation results given is appropriate. This is proven based on the opinions obtained, namely opinions with a level of machine understanding (34%) and very understanding (27%). It can be seen that most of the translation results can be reached by the respondents.

In this research, the data used only covers political topics and cannot cover other topics. Therefore, we suggest using augmentation techniques, which could create additional data or synthesize data using existing data, in future studies in the hope of expanding the scope of the topics used. Based on the results of our experiments, we found that too-short or too-long sentences lead to significant changes in the BLEU score. This could also be used as material for further research.

ACKNOWLEDGEMENTS

The authors would like to thank Institut Teknologi Del for financing this research, Mr. Lit Malem Ginting, S.Si, MT, and Mrs. Monalisa Pasaribu, SS., M.Ed (TESOL) for giving guidance towards this research.

REFERENCES

[1] S. Rabiah, “Language as a Tool for Communication and Cultural Reality Discloser,” *INA-Rxiv*, pp. 1–11, 2018, doi: 10.31227/osf.io/nw94m.

[2] V. de Klerk, "Slang, Sociology,” *Encyclopedia of Language and Linguistics*, pp. 407-412, 2006, doi: 10.1016/B0-08-044854-2/01303-1.

[3] L. Wu, F. Morstatter, and H. Liu, “SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification,” *Language Resources and Evaluation*, vol. 52, no. 3, pp. 839-852, 2018, doi: 10.1007/s10579-018-9416-0.

[4] D. Sebastian and K. A. Nugraha, "Text Normalization for Indonesian Abbreviated Word Using Crowdsourcing Method," *2019 International Conference on Information and Communications Technology (ICOIACT)*, 2019, pp. 529-532, doi: 10.1109/ICOIACT46704.2019.8938463.

[5] D. Pennell and Y. Liu, “A Character-Level Machine Translation Approach for Normalization of SMS Abbreviations,” *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 2011, pp. 974–982.

[6] Y. S. A. Mustofa, M. M. Ismail, and H. T. Lin, Learning from Data, AMLbook USA, 2012.

[7] P. Li, “A Survey of Machine Translation Methods,” *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 11, no. 12, pp. 7125–7130, 2013, doi: 10.11591/telkomnika.v11i12.2780.

[8] R. Zens, F. J. Och, and H. Ney, “Phrase-Based Statistical Machine Translation Phrase-Based Statistical Machine Translation,” *Annual Conference on Artificial Intelligence*, 2014, doi: 10.1007/3-540-45751-8.

[9] B. Ahmadnia and B. J. Dorr, “Low-resource multi-domain machine translation for Spanish-Farsi: Neural or statistical?,” *Procedia Computer Science*, vol. 177, pp. 575–580, 2020, doi: 10.1016/j.procs.2020.10.081.

[10] P. Bhattacharyya, *Machine Translation*, 1st ed. CRC Press, 2015.

[11] P. Koehn *et al.*, “Moses: open source toolkit for statistical machine translation,” *Proceedings of the ACL 2007 Demo and Poster Sessions*, 2007, pp. 177-180, doi: 10.3115/1557769.1557821.

[12] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, “Neural versus phrase-based machine translation quality: A case study,” *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, 2016, pp. 257–267, doi: 10.18653/v1/d16-1025.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2001, vol. 22176, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.

[14] F. Nurifan, R. Sarno, and C. S. Wahyuni, “Developing corpora using word2vec and wikipedia for word sense disambiguation,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 3, pp. 1239–1246, 2018, doi: 10.11591/ijeecs.v12.i3.pp1239-1246.

[15] K. Heafield, “KenLM: Faster and Smaller Language Model Queries,” *Proceedings of the 6th Workshop on Statistical Machine Translation.*, 2011, pp. 187-197.

[16] F. J. Och and H. Ney, “A Systematic Comparison of Various Statistical Alignment Models,” *Computational Linguistics*, vol. 29, pp. 19–51, 2003, doi: 10.1162/089120103321337421.

[17] P. Passban, Q. Liu, and A. Way, “Enriching phrase tables for statistical machine translation using mixed embeddings,” *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap.*, 2016, pp. 2582–2591.

[18] F. J. Och, “Minimum Error Rate Training,” *Proc. 41st Annu. Meet. Assoc. Comput. Linguist.*, 2003, pp. 160–167.

[19] A. R. Nabhan and A. Rafea, "Tuning statistical machine translation parameters using perplexity, "*IRI -2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005.*, 2005, pp. 338-343, doi: 10.1109/IRI-05.2005.1506496.

[20] E. Reiter, “A Structured Review of the Validity of BLEU,” *Comput. Linguist.*, 2018, doi: 10.1162/coli_a_00322.

[21] V. R. Vyshnavi and A. Malik, “Efficient Way of Web Development Using Python and Flask,” *Int. J. Recent Res. Asp.*, vol. 6, no. 2, pp. 16–19, 2019.

[22] A. M. Adam, “Sample Size Determination in Survey Research,” *Journal of Scientific Research and Reports*, vol. 26, no. 5, pp. 90–97, 2020, doi: 10.9734/jsrr/2020/v26i530263.

[23] J. Kelleher, B. Mac Namee, and A. D’Arcy, Fundamentals Machine Learning For Predictive Data Analysis, *The MIT Press Cambridge*, 2015.

[24] S. E. Harpe, “How to analyze Likert and other rating scale data,” *Currents in Pharmacy Teaching and Learning*, vol. 7, no. 6, pp. 836–850, 2015, doi: 10.1016/j.cptl.2015.08.001.

[25] Google Cloud, “Evaluating models.” 2021. Accessed: 16, October 2021. [Online]. Available: https://cloud.google.com/translate/automl/docs/evaluate

BIOGRAPHIES OF AUTHORS

Kyrie Cettyara Eleison iD 8 SC P is now a third year student of Information Technology at Del Institute at Technology. Her current research interests focus on Software Engineering, Intelligent Systems and Business Process Management. She can be contacted at e-mail: kyriecettyara@gmail.com.

Sari Uli Ingrid Hutahaean iD 8 SC P is now third year student of Information Technology at Del Institute of Technology. She is an assistant lecturer in the discrete mathematics course in semester 2 and an assistant lecturer in linear algebra in semester 6. Her interests are about the world of technology and its development. She can be contacted at e-mail: sariulihutahaean@gmail.com.

Sarah Christine Tampubolon iD 8 SC P is now third year student of Information Technology at Del Institute of Technology. She received Beasiswa Prestasi Scholarship from Institut Teknologi Del in her first, second, fourth and fifth semester. She was a Student Teaching Assistant of Introduction to Database course in 4th semester. Her interests are about Technology, Application Development, Artificial Intelligent (AI), Quality Assurance and Data Science. She can be contacted at e-mail: sarah.ch.tampubolon@gmail.com.

Teamsar Muliadi Panggabean iD 8 SC P is a member of Faculty of Informatics and Electronic Engineering (FITE) at Del Institute of Technology. His interests in Machine Learning, NLP, Machine Translation, Distributed System, and High-Performance Computing. Hold a certification in Associate Big Data Engineer from Data Science Council of America (DASCA). He can be contacted at e-mail: teamsar.panggabean@del.ac.id.

Ike Fitriyaningsih iD 8 SC P is an Assistant Professor at the Study Program Diploma III Technology Information of Institut Teknologi Del. She received her Bachelor and Master in Statistics Program of Brawijaya University in 2012 and Institut Teknologi Sepuluh Nopember in 2015. Her research interests in prediction, spatial regression and machine learning. She can be contacted at e-mail ike.fitri@del.ac.id or ike.fitriyaningsih@gmail.com.