COMMUNICATIONS IN
SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Machine translation of Indonesian: A review

Amalia Agung Septarina*, Faisal Rahutomo, M. Sarosa

*Electrical Engineering Department, Politeknik Negeri Malang, Malang, 65141, Indonesia

Article history:
Received: 14 December 2018 / Received in revised form: 30 May 2019 / Accepted: 02 June 2019

Abstract

Nowadays, machine translation has an important role in general communication. The need for machine translation system is higher in this era, resolving culture and nation boundary. Finding appropriate and optimal translation is not an easy task in language processing. Several machine translation systems already exist, but the quality of the translation is needed to be improved further. This paper discusses machine translation researches that involve Indonesian language to the other languages by systematic literature review. This paper exposes different approaches and tools for machine translation. The approaches also use various evaluation methods to measure performance. Moreover, this paper proposes several future works to improve the machine translation quality of Indonesian to other languages. The review results show that the attention-based approach is being increasingly used to improve the performance of neural machine translation. The translation performance quality depends on the number of the corpus, well-behaved aligned corpus, and the technique used.

Keywords: Machine Translation; Indonesian Review

1. Introduction

Machine translation is a sub-field of computational linguistics. Machine translation can be defined as the computerized system to translate from one language to another language. There are various approaches in machine translation, while Figure 1 shows the approaches briefly.

This paper aims to expose the researches of machine translation regarding Indonesian language. Indonesian language is the national language of Republic of Indonesia. Republic of Indonesia itself is a linguistically rich country. Sugiyono said that Indonesia has 726 traditional/local languages [1]. The three most used traditional languages are Javanese, Sundanese, and Maduranese. Several traditional languages are endangered from the lack of useness. Nowadays, not only traditional language is endangered, the Indonesian vocabulary itself is endangered. The use of vocabulary is decreased in daily communication [2]. The focus of discussion is needed since in computational linguistics different language will need different approaches. There are some researches in this area with focus in translation of Indonesian to the other languages.

This paper reviews the researches from a different angle of views: dataset, tools, method, and evaluation metric. Sometime, the research focuses on translation between Indonesian traditional language, such as [3]. The other research focuses on translation between Indonesian to foreign country language, such as [4]. The problem raises from lack of available dataset [3] with low resource of language pair. There are no Sundanese to Indonesian parallel corpus that ready to use. Then they collect the dataset manually from su.wikipedia.org and id.wikipedia.org. Several problems in the translation process are detected, such as low coverage corpus data, unknown word, and sentence reordering problem [4]. This paper does not discuss how to optimize the translation of Indonesian machine translation. This paper focuses on the approach used in Indonesian in recent years. The most researches use statistical machine translation. In the other hand, the attention-based neural machine translation is being increasingly performed in Indonesian machine translation system.

After introduction, this paper exposes the approaches to machine translation in chapter 2. Chapter 3 exposes the review itself with several subchapters. Table 1 contains a summary of the review. Chapter 4 concludes the review with several ideas for further work. Then chapter 5 provides the conclusion for this review paper.

2. Approaches to Machine Translation

Below are various approaches to machine translation that used in indonesian machine translation from recent years. Fig. 1 shows all various approaches to Machine Translation. First branch of the approaches are rule-based, hybrid, and empirical system. The approaches are developed further into various approaches.

2.1. Rule-Based Machine Translation (RBMT)

Rule-Based Machine Translation involves morphological, syntactic, and semantic rules about the source and target language [5][6]. This system can handle word-order problems

Amalia Agung Septarina Tel.:+62-822-3477-8840.
Email: amalia_agung_septarina@polinema.ac.id.

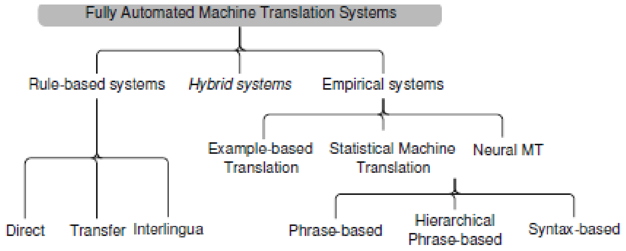
Fully Automated Machine Translation Systems
Rule-based systems
Hybrid systems
Empirical systems
Example-based Translation
Statistical Machine Translation
Neural MT
Direct
Transfer
Interlingua
Phrase-based
Hierarchical Phrase-based
Syntax-based

Fig. 1 General approach to machine translation [7].

and trace parse error using linguistic knowledge. Rule-based MT systems works based on the specification of rules for morphology, syntax, lexical selection and transfer and generation. RBMT depend on bilingual or multilingual lexicon that is manually built and the collection of rules. RBMT divided into direct method, transfer method, and Interlingua (IL) (see Fig. 2). Direct method does word-by-word translation directly. Transfer method analyzes of the syntactic structure of source language (SL) which results in an abstract representation of the sentences, then transferred to the abstract representation of the target language (TL), and the output generated from it using bilingual dictionaries and grammar rules. Interlingua method, abstract representation is assumed to be the same for all language and there is no need transfer step.

RBMT process the system word by word and can't handle ambiguity and idiomatic expression. Hence the resulting translation often not fluent and can't generate natural translation. The post-editing work is required to be adapted to the specific target audience and writing style.

2.2. Statistical Machine Translation (SMT)

Warren Weaver had introduced the idea of Statistical Machine Translation [5]. SMT is an approach to MT that is characterized by the use of machine learning methods [6]. SMT is one of the machine translation system using statistical approach which parameters are derived from the results of parallel corpus analysis. The statistical approach used is the concept of probability. The higher the probability value indicates that the translation results are well-formed sentences. There are three models in statistical approach, phrase based, syntax-based, and hierarchical phrase-based system.

SMT can handle morphology because it can separate suffixes that inflected word leading to meaning transfer. In other words, SMT can handle ambiguity. The system records phrase-based translations with their frequency of occurrence on phrase table. Thus, the translation result generates more fluent and natural than RBMT. One weakness of SMT is the challenge of translating material that is not similar to content from the training corpora [8]. It gives poor accuracy of the translation result. So that, to achieved good translation, the corpus should be customized for a specific style. SMT does not work well between languages that have significantly different word orders e.g. Japanese-Indonesian.

2.2.1. Phrase-Based

The fundamental unit of phrase-based is a phrase or sequence of words but is not necessarily a linguistic element. The phrasemes found using statistical method from corpora. Phrasemes or multi-word expression utterance at least one of whose components is selectionally constrained or restricted by linguistic convention such that it is not freely chosen [9]. The input of phrase is segmented into phrases, translated one-to-one into phrases target and possibly reordered.

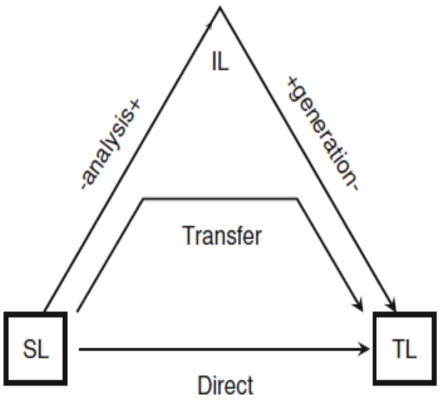

Fig. 2 Rule-based machine translation [9].

2.2.2. Syntax-Based

The basic idea of syntax-based is the translation rule by synchronous grammar between source and target language. The rule for translation consists of sequence of words, syntax tree, and vector of feature value which describe the language pair. Synchronous grammars are learned from parallel corpus and that makes the approach very slow in comparison to the PBSMT systems [7].

2.2.3. Hierarchical-Phrase-Based

Hierarchical phrase-based systems combine a balance between pure lexical phrase-based and syntax-based translation. A hierarchical phrase consists of words and subphrases and this hierarchy is intended to capture reordering among phrases [7]. The hierarchical phrase pairs use synchronous context-free grammar (CFG) rules learned from parallel corpora without syntactic information.

2.3. Hybrid Machine Translation

This approach is a combination of the multiple machine translation approach. Often associated with "statistical" and "rule-based" approaches (see fig.3). Developing hybrid machine translation stems from the failure of any single technique to achieve a satisfactory level of accuracy. There are several types of hybrid system such as multi-engine, statistical rule generation, multi-pass and Confidence-based.

2.3.1. Hybrid with Multiple Approaches

Hybrid takes advantage of the combination of multiple approaches (SMT and RBMT). In some cases, the rule-based approach implemented in the first step by built the lexicon and implement the grammar rules and other rules. Then followed by correcting the output using SMT approach. In other cases, rules are used to pre-process the input data as well as post-

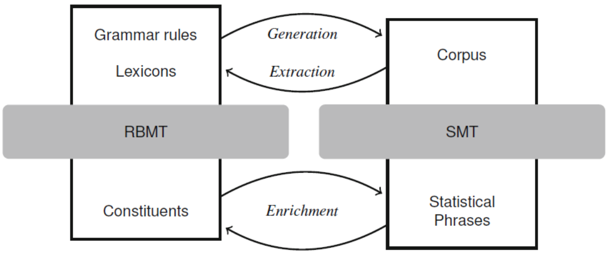

Fig. 3 Resources exchanges between RBMT and SMT. Some hybrid system are built by taking advantage of these relations[7].

process the statistical output of a statistical-based translation system. This technique is better than the previous and has more power, flexibility, and control in translation [10]*Multi-Engine*

Multi-engine to hybrid machine translation implement multiple machine translation system in parallel. The output obtained from combining all the system selected. C. Hogan and R. E. Frederking [11] combining example-based, transfer based, knowledge-based and statistical translation sub-systems into one machine translation system (see fig. 4). First analysis the morphology then processed using some of the methods in RBMT and SMT.

2.3.2. Statistical Rule Generation

This method uses statistical data to obtain lexical and syntactic rules. The input processed used a rules-based engine.

2.3.3. Multi-Pass

This approach involves respectively processing the input multiple times. The most common technique used in multi-pass machine translation systems is to pre-process the input with a rule-based machine translation system. The output of the rule-based pre-processor is passed to a statistical machine translation system, which produces the final output. This technique is used to limit the amount of information a statistical system need consider, significantly reducing the processing power required. It also removes the need for the rule-based system to be a complete translation system for the language, significantly reducing the amount of human effort and labour necessary to build the system.

2.4. Neural Machine Translation

This approach uses large artificial network technology to predict a possible sequence of words in a single integrated model. Neural machine translation is widely used by researchers to the proposed translation system. The structure of the models is simpler than phrase-based models.

In 201, Nal Kalchbrenner and Phil Blunsom were at first typically done using a recurrent neural network (RNN). NMT with NN-based encoder-decoder to address sequence-to-sequence model to prediction problem. NMT doesn’t need reordering model, translation model, and language model, but just a single sequence model that predicts one word at a time.

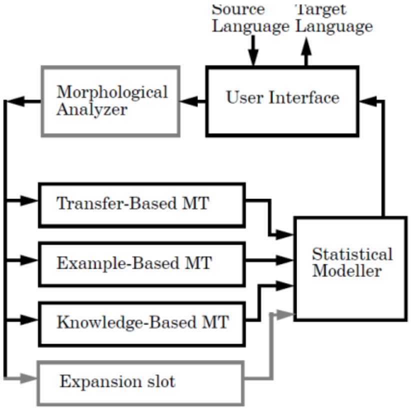

Fig. 4 Architecture of multi-engine machine translation [11].

This model will encode a given source text into a continuous vector using Convolutional Neural Network (CNN), and then use Recurrent Neural Network (RNN) as the decoder to predict the word in the target language. Encoder-decoder still has a problem with long sequences of text to be translated.

In 2014, Sutskever et al. and Cho et al. introduced RNN with Long Short-Term Memory (LSTM). This model can handle "long-distance reordering" problem in a sentence much better. Another challenge for NMT is "fixed-length vector". The neural network needs to compress the source sentence into a fixed-length vector, which will lead to increasing complexity and uncertainties during decoding especially when the source sentence is long [12].

Yosua Bengio’s group introduced the "attention-based" model to NMT in 2014. The attention-based approach is being increasingly used to improve the performance of neural machine translation (NMT). The neural machine translation with attention is currently the state-of-the-art on some benchmark problems for machine translation. Most of the best MT systems were using neural network such as Google, Facebook, Amazon, Microsoft, SYSTRAN, etc [12]. There are some toolkits for neural machine translation i.e. OpenNMT, Xnmt, Nematus, Sockeye, T2T, and Marian [13].

NMT can be breakthrough over previous technology because NMT systems understand similarities between words, NMT Systems Consider Entire Sentences, and NMT Systems Learn Complex Relationships between Languages [14].

3. Machine Translation of Indonesian

3.1 *Dataset*

Parallel corpus is very important resources in machine translation. PAN build open source parallel corpus Indonesia-English for translation system in their project with reasonable size of Parallel Corpus Indonesia-English. Started by collecting Indonesian corpus and perform raw corpus cleaning, translation, alignment and XML tagging (see fig. 5).

3.2 *Tools*

Moses Decoder one of common tools used in machine translation system. Moses is an open-source project, licensed

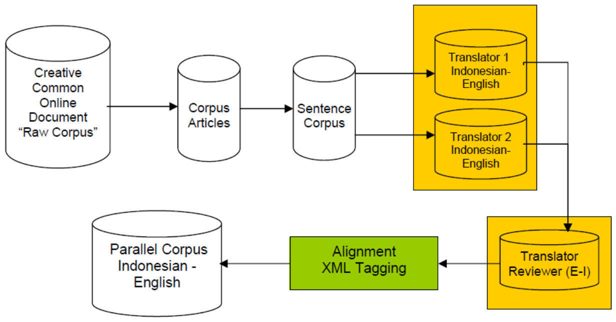
Creative Common Online Document "Raw Corpus"
Corpus Articles
Sentence Corpus
Translator 1 Indonesian-English
Translator 2 Indonesian-English
Translator Reviewer (E-I)
Alignment XML Tagging
Parallel Corpus Indonesian - English

Fig 5. Overview of PANL BPPT parallel corpus [15].

under the LGPL, which incorporates contributions from many sources. It used for statistical machine translation system that allows you to automatically train models for any language pair. Moses decoder finding the highest score in target language. In statistical machine translation, you need parallel corpus as the dataset. Parallel corpus is collection sentence between two different languages that each sentence in one language have related translation to others. There are two main parts in moses, training and decoder. Moses is mainly written in perl and some C++. The following steps have to perform before train the data.

(a). Tokenisation: This means that spaces have to be inserted between (e.g.) words and punctuation.

(b). Truecasing: The initial words in each sentence are converted to their most probable casing. This helps reduce data sparsity.

(c). Cleaning: Long sentences and empty sentences are removed as they can cause problems with the training pipeline, and obviously misaligned sentences are removed.

There are three steps to train the data, word alignment, language model and tuning. The first step for train the data is word alignment, typically using GIZA++. Word alignment is used to extract phrase translations or hierarchical rules, and corpus-wide statistics on these rules are used to estimate probabilities. Then, language model, a statistical model built using monolingual data in the target language and used by the decoder to try to ensure the fluency of the output. Moses relies on external tools for language model building. Moses supports several language model tool kits such as KenLM, SRILM, IRSTLM, RandLM.

The final step in the creation of the machine translation system is tuning, where the different statistical models are weighted against each other to produce the best possible translations. Decoder ranked list of the translation candidates, and also to supply various types of information about how it came to its decision (for instance the phrase-phrase correspondences that it used).

3.3 Previous Indonesian Machine Translation

Some journal discusses how optimizing translation Indonesian-English using statistical machine translation and also other approaches will discuss in this section (see Table 2). One of the advantages of using statistical machine translation is with a larger corpus, it will learn the "context" of phrase if it occurs enough, and hence it produces a more appropriate translation [16]. A good parallel corpus should meet the requirement below [4]:

(a). It should contain naturally occurring language data;

(b). It should be representative of its domain;

(c). Alignment process should be done with high accuracy;

(d). It should have a reasonable length per sentence pair.

T. Mantoro et.al [16] discuss translation process from English to Indonesian using statistical machine translation by considering four parameter i.e. translation model (w_t), language model (w_l), distortion/reordering (w_d), word penalty (w_w).

The well-behaved aligned parallel corpus as the training data is used to increase the evaluation score. The parallel corpus used for training was collected from domains-newspaper, the websites of commercial, government, educational institution, and Penn TreeBank corpus licensed from PAN Localization. The training corpus has 25,715 parallel sentence which includes 563,666 English words and 525,102 Indonesian words.

For the evaluation metrics, they use BLEU (Bilingual Evaluation Understudy) and NIST (National Institute of Standart and Technology).

Two Properties of the BLEU metrics are reliance on higher n-gram and the Brevity Penalty (BP). The value of BLEU metric is between 0 and 1, with 1 being the candidate translation with high accuracy. The BLEU formula is shown below:

$$BP_{BLEU} = \begin{cases} 1, \ if\ c > r \\ e^{(1-r/c)}, \ if\ c \leq r \end{cases} \quad (1)$$

$$P_{n=} \frac{\sum_{C \in corpus n-gram \in C} \sum count_{clip}\,(n-gram)}{\sum_{C \in corpus n-gram \in C} \sum count\,(n-gram)} \quad (2)$$

$$BLEU = BP_{BLEU} \bullet e^{\sum_{n=1}^{N} W_n \log P_n} \quad (3)$$

NIST based on BLEU metric with some changes. NIST calculate how informative a particular n-gram is. The formula to calculate NIST score is shown below:

$$BP_{NIST} = e^{\left\{\beta log^2\left[min\left(\frac{L_{sys}}{L_{ref}},1\right)\right]\right\}} \quad (4)$$

$$NIST = \sum_{n=1}^{N} \left\{ \frac{\begin{matrix} \sum Info(w_1 \ldots\ldots w_n) \\ all\ w_1 \ldots\ldots w_n \\ that\ co-occur \end{matrix}}{\begin{matrix} \sum(1) \\ all\ w_1 \ldots\ldots w_n \\ in\ sys\ output \end{matrix}} \right\} \cdot BP_{NIST} \quad (5)$$

They also compared their system (statistical machine translation) results with the previous work rule-based machine translation results. Their system presents better translation than rule-based machine translation as it does not miss keywords, synonyms, and enough data in corpus.

A. A. Suryani et.al [3] discussed Sundanese into Indonesian translation using phrase-based statistical machine translation with PoS Tag Information. They use three kinds of translation model. Table 1. shows the translation model and its descriptions.

Table 1. Translation model with PoS tag information

Model	Description
0-0,1	PoS Tag added to Indonesian
0,1-0	PoS Tag added to Sundanese
0,1-0,1	PoS Tag added to Sundanese and Indoesaian

They use 75, 150, 250, and 350 parallel corpora that collected manually from Wikipedia which are Sundanese and Indonesian. The evaluation metric that they used is BLEU. They compared surface form, PoS Tag, Google Translate. In the dataset 150, 250, 300 achieved better BLEU score than baseline. The translation model with PoS Tag Information gives better performance than only surface form.

J. Pranata [17] experiment Indonesian to Javanese translation using phrase-based statistical machine translation approach. Total 4500 datasets from bible. BLEU score as the evaluation metric for this experiment shows that their machine translation gives higher BLEU than Google Translate with 29.79% range value.

M. A. Sulaeman and A. Purwarianti [4] discuss Indonesian-Japanese using statistical machine translation. Research about Indonesian-Japanese also has several problems, such as low coverage parallel corpus, unknown words, and sentence reordering. The two methods to handle these problems i.e. lemma translation and additional post-process. A total of 1132 sentences corpus was collected from JLPT (Japanese Language Proficiency Test) level 3 and tatoeba.org. Its number increase in domain coverage compared with parallel corpus used in previous research.

PoS Tag is needed for obtaining surface form used in lemma translation process. Hierarchical reordering was added after PoS Tag model. It has been evident as a better model than word-based and phrase-based reordering. They set several rules for untranslated katakana and unknown words synonym substitution at additional post-process. They also use KBBI to checked the generated words in Indonesia. The synonym of the unknown words is searched on Japanese WordNet. The result shows that 116% increased value of BLEU on Japanese to Indonesian and 26% increased value of BLEU on Indonesian to Japanese.

BPPT [15] discuss statistical machine translation for Indonesian-English using dataset from BBPT (Badan Pengkajian dan Penerapan Teknologi). They use 500.000 words into 9 blocks of word Indonesian-English and English-Indonesian. The results show that BLEU score was improved by increasing size of words. They obtain 92.1% translation quality for English-Indonesian. A higher BLEU score represents better translation.

H. Sujaini et.al [18] present a PoS (Part-of-Speech) method for statistical approach. Collected 27K parallel corpus Indonesian-English used in this experiment. They compared translation system uses Grammar PoS, Computational PoS, and without PoS. In the test step, a total of 1500 sentences consist of 5 groups with word length 10, 15, 20, 25, and 30. The results show that Computational PoS achieved higher BLEU score with an average value of 52.95%. While using without PoS and Grammar PoS represent 49.74% and 50.85%. Accuracy of the translation system in short sentence (10 words) results in high BLEU score.

C. O. Mawalim et.al [19] focus on Indonesia to Korean translation using statistical machine translation approach with PoS-based reordering rules. A total 11,155 parallel corpus from movie/drama subtitles and Korean language books used in their experiment. They using PoS Tag and word alignment information, then apply 150 reordering rules for Korean-Indonesian and 50 reordering rules for Indonesian-Korean. This method increases the quality of translation by BLEU score of 1.25% for Indonesian-Korean and 0.83% for Korean-Indonesian. They also apply this reordering rules with Korean verb formation rules for Indonesian-Korean. That increased BLEU score from 38.07 to 49.46.

K. M. Shahih and A. Purwarianti [20] discuss handling utterance disfluency in Indonesian-English translation. They apply hybrid approach that combines statistical-based and rule-based. The experiment compared using CRF model with 5 labels (0, FL, RC, NC, G) and CRF model with 3 labels (0, FL, NC) and the extension of rule-based. The label 0 for fluent words, FL for filled pause and discourse marker, RC for repetition, NC for restart phenomenon, G for stutter problem. They use variations of lexical features such as word, pos, dist_word, dist_pos, sim, token_position in every testing experiment. The results show BLEU score for an original dataset is 10.10, scenario1 word, dist_word, sim 12.70, scenario2 word, token_position, dist_word, sim 12.71, scenario3 word_pos, dist_word, dist_post, sim 12.60, oracle (all disfluencies removed) 13.87.

A. A. Suryani et.al [21] proposed a method to fill unknown translation of English into Javanese and Sundanese which occurred in the phrase translation in Translator-Gator system using phrase-based approach. Translator-Gator is a language game created by the United Nation Global Pulse to support the research initiative in Indonesia. They use Indonesian as a pivot. There are 1324 unique English keyword form transactional data translation by more than 100 Translator-Gator users. From these keywords, they get 1340 pair of initial Indonesian-Javanese dictionary and 460 pair initial Indonesian-Sundanese dictionary. They set two rules which are searching the keyword in phrase translation table and if it isn't in translation table, then check whether the word was a borrowed word or a phrase. If it was a borrowed word, then the Indonesian translation is a result. If it is a phrase break into N-word and translate word by word using the step of translation rules. The translation evaluation defined using Slovin Formula. The experiment results show relatively low translation accuracy by 37% correct translation of Indonesian-Javanese and 46% of Indonesian-Sundanese. They also apply weighting formula based on the number translation occurrence, the number of users that agreed and disagreed the translation. The results show that formula increases the translation accuracy which 65% proper phrase translation for Indonesian-Javanese and Indonesian-Sundanese.

Other approach used some researcher to increase translation quality. A few years later after Mohammad Anugrah Sulaeman et.al study about Indonesian-Japanese using statistical approach, M. T. Models [4] present translation Japanese-Indonesian using Neural Machine Translation. They use more data corpus parallel. There is 725,495 corpus parallel from Open Subtitle 2016, Asian Language Treebank, Tanzil, Global Voices, and Tatoeba. In their experiment, they use three methods i.e. RNNenc (simple RNN encoder-decoder), mRNNa(multi-layer RNN with attention), biRNN

Table 2. Comparison of the results using different approach and method

Researcher	Year	Language	Dataset	Tools	Approach & Method	Result
BPPT [15]	2009	Indonesian -English	BPPT	Moses Decoder	Statistical Machine Translation	BLEU
T. Mantoro et.al [16]	2013	English-Indonesian	Penn Treebank (PAN Localization)	Moses Decoder	Statistical Machine Translation (weight variable: translation model, language model, reordering, word penalty)	NIST; BLEU
H. Sujaini et.al [18]	2014	English-Indonesian	27K sentences	Moses Decoder	Statistical Machine Translation (Grammar PoS, Computational PoS)	BLEU
A. Hermanto et.al [22]	2015	English-Indonesian	BPPT	Cygwin	Recurrent Neural Network	BLEU 24.5; RIBES 76.3
A. A. Suryani et.al [3]	2015	Sundanese - Indonesian	(su.wikipedia.org and id.wikipedia.org)	Moses Decoder	Phrase-based SMT	BLEU
M. A. Sulaeman et.al [4]	2015	Indonesian -Japanese	1132 sentences (tatoeba.org and Japanese Language Profeciency Test Level 3)	Moses Decoder	Statistical Machine Translation (Lemma Translation and Post-Process)	BLEU
J. Pranata et.al [17]	-	Indonesian -Javanese		Moses Decoder		BLEU
K. M. Shahih et.al [20]	2016	Indonesian -English	27K Sentences	Moses Decoder, CRF++	Hybrid approach (statistical-based and rule-based)	BLEU
A. A. Suryani et.al [21]	2016	English-Sundanese English-Javanese	1,340 pair Indonesian-Javanese, 460 pair Indonesian-Sundanese. (Transactional data translation by more than 100 users Translator-Gator)	Moses Decoder	Rule-based	Slovin Formula
C. O. Mawalim et.al [19]	2017	Indonesian -Korean	11,155 segments (Subtitle drama/movie and korean language book for indonesian)	Moses Decoder	Phrase-based Translation Model	BLEU
M. T. Models [23]	2017	Japanese-Indonesian	725,495 sentences (Open subtitle 2016, asian language Treebank, globalvoices.org, tanzil, tatoeba)	Moses Decoder	Neural Machine Translation	BLEU
Zaenal Abidin et.al [19]	2018	Lampung-Indonesia Language	3,000 sentences (Lampung reference book)	THUMT-Theano	Neural Machine Translation	BLEU 51.96

(bidirectional RNN).

The NMT models use layer size, word embedding dimension, and attention mechanism. The results show BLEU score with removed unknown words achieved higher BLEU. RNNenc with all data 4.45 and without unknown words 4.96. mRNNa with all data 4.57 and without unknown words 5.16. biRNN with all data 4.85 and without unknown words 6.45. They also calculated the accuracy of the translation system using phrase-based SMT and achieved 8.78 BLEU score with all data and 9.34 without unknown words. A large number of unbalanced translation from parallel corpus may disturb NMT to learn the correct translation.

A. Hermanto et.al [22] present Recurrent Neural Network Language approach for English to Indonesian machine translation. A total 10462 training sentence from BPPT used in Bilingual Evaluation Score) RIBES for evaluation. This research compared Recurrent Neural Network (RNN) with Statistical Machine Translation. The result shows that RNN-based produce higher BLEU and RIBES value. Statistical-based has 23.4 BLEU score and 74.7 RIBES score. RNN-based has 24.5 BLEU score and 76.3 RIBES score.

Z. Abidin et.al [24] also use Neural Machine Translation (NMT) based on attention for Lampung-Indonesian Language. They collected 3000 parallel corpora manually from Lampung language reference book. They perform three experiments for NMT model attention. First, NMT model attention with the size of the hidden layer (n) 500 and word embedding (m) 310. Second, the size of the hidden layer (n) 1000 and dimensional vector of word embedding (m) 620. Third, the size of the hidden layer (n) 1500 and dimensional vector of word embedding (m) 930. The evaluation result using BLEU score achieved 51.96% accuracy value. NMT model attention on second experiment get the best configuration dimension.

4. Future Work for Indonesian Machine Translation

The need for machine translation is getting higher in this information era. Neural machine translation is widely used by researchers to the proposed translation system. NMT with RNN-based encoder-decoder to address sequence-to-sequence model to prediction problem. NMT doesn't need reordering model, translation model, and language model, but just a single sequence model that predicts one word at a time. Encoder-decoder still has a problem with long sequences of text to be translated. The attention-based approach is being increasingly used to improve the performance of neural machine translation (NMT). The neural machine translation with attention is currently the state-of-the-art on some benchmark problems for machine translation.

5. Conclusion

The review result describes the research conducted on machine translation focused on Indonesian to another language. Most of the researcher of the Indonesian language implements the statistical approach in their study and they manually collect the parallel corpus for the data training. Several translation machine of Indonesian to other language needs to be improved by generating good translation. The attention-based approach is being increasingly used to improve the performance of neural machine translation (NMT). Accuracy of the translation system is influenced by many factors. The number of parallel corpus can increase evaluation score.

Table 2 shows that the open spaces can be investigated further by Indonesian machine translation researcher. Hundreds of traditional languages can be explored further. Building new datasets and providing it freely for research community are interesting to work. The other open space is to try the other methods of already available work and compare the performance result. Developing a new tool especially work in Indonesian machine translation is the other space. The researcher can also try different performance metrics or develop new performance metric in this research area.

References

1. Tempo.co, 3 Bahasa Terpopuler di Indonesia, 2012. [Online]. Available: https://nasional.tempo.co/read/435218/3-bahasa-terpopuler-di-indonesia/full&view=ok. [Accessed: 20-Aug-2018].
2. F. Rahutomo, et al., *Computational analysis on rise and fall of Indonesian vocabulary during a period of time*, 6th Int. Conf. Inf. Commun. Technol. ICoICT, 2018, pp. 75–80.
3. A. A. Suryani, et al., *Experiment on a phrase-based statistical machine translation using PoS Tag information for Sundanese into Indonesian,* Int. Conf. Inf. Technol. Syst. Innov. ICITSI- Proc., 2016.
4. M. A. Sulaeman and A. Purwarianti, *Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process,* Proc. - 5th Int. Conf. Electr. Eng. Informatics Bridg. Knowl. between Acad. Ind. Community, ICEEI 2015, pp. 54–58.
5. K. H. Khayat, G. Ballivy, and M. Gaudreault, *High-performance cement grout for underwater crack injection*, Can. J. Civ. Eng. 24 (1997) 405–418.
6. J. Oladosu, A. Esan, I. Adeyanju, B. Adegoke, O. Olaniyan, and B. Omodunbi, *Approaches to Machine Translation: A Review*, 1 (2016) 120–126.
7. M. R. C. R. Rapp, P. Lambert, and K. Eberle, *Hybrid Approaches to Machine Translation*, 2016.
8. U. L. Group, The Pros and Cons of Statistical Machine Translation. [Online]. Available: https://unitedlanguagegroup.com/blog/pros-and-cons-statistical-machine-translation/.
9. Phrasemes, Available: https://en.wikipedia.org/wiki/Phraseme. [Accessed: 12-Jun-2018].
10. M. D. Okpor, *Machine Translation Approaches: Issues and Challenges*, IJCSI Int. J. Comput. Sci. Issues. 11 (2014) 159–165.
11. C. Hogan and R. E. Frederking, *An Evaluation of the Multi-engine MT Architecture*, Proc. Third Conf. Assoc. Mach. Transl. Am. Mach. Transl. Inf. Soup, 1998, pp. 113–123.
12. Synced, History and Frontier of the Neural Machine Translation, 2017. Available: https://medium.com/syncedreview/history-and-frontier-of-the-neural-machine-translation-dc981d25422d. [Accessed: 12-Nov-2018].
13. N. Machine and T. Philipp, *Statistical Machine Translation*, 2017.
14. G. Dino, "3 Reasons Why Neural Machine Translation is a Breakthrough, 2017. Available: https://slator.com/technology/3-reasons-why-neural-machine-translation-is-a-breakthrough/.
15. BPPT, Final Report on Statistical Machine Translation for Bahasa Indonesia-English and English- Bahasa Indonesia, no. Technical Report PAN Localization Project, 2009.
16. T. Mantoro, J. Asian, R. Octavian, and M. A. Ayu, *Optimal translation of English to Bahasa Indonesia using statistical machine translation system*, 5th Int. Conf. Inf. Commun. Technol. Muslim World, 2013, pp. 1–4.
17. J. Pranata, T. Informatika, F. I. Komputer, and U. D. Nuswantoro, *Mesin penerjemah bahasa indonesia- bahasa jawa*, no. 5, pp. 1–5.
18. H. Sujaini, A. A. Arman, and A. Purwarianti, A *Novel Part-of-Speech Set Developing Method for Statistical Machine A Novel Part-of-Speech Set Developing Method for Statistical Machine Translation*, no. 2015, 2014.
19. C. O. Mawalim, D. P. Lestari, and A. Purwarianti, *POS-based Reordering Rules for Indonesian – Korean Statistical Machine Translation*, 6th International Conference on Electrical Engineering

and Informatics (ICEEI) 2017, pp. 1-6.

20. K. M. Shahih and A. Purwarianti, *Utterance disfluency handling in Indonesian-English machine translation*, 4th IGNITE Conf. Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA, 2016, pp. 1–4

21. A. A. Suryani, I. Arieshanti, B. W. Yohanes, M. Subair, S. D. Budiwati, and B. S. Rintyarna, *Enriching English into Sundanese and Javanese translation list using pivot language*, Proc. 2016 Int. Conf. Inf. Commun. Technol. Syst. ICTS, 2016, pp. 167–171.

22. A. Hermanto, T. B. Adji, and N. A. Setiawan, *Recurrent Neural Network Language Model for English-Indonesian Machine Translation : Experimental Study*, Int. Conf. Sci. Inf. Technol., 2015, pp. 132–136.

23. M. T. Models, *Performance of Japanese-to-Indonesian*, no. C, pp. 7–10, 2017.

24. Z. Abidin, A. Sucipto, and A. Budiman, *Penerjemahan Kalimat Bahasa Lampung-Indonesia Dengan Pendekatan Neural Machine Translation Berbasis Attention* , vol. 06, no. 02, pp. 191–206, 2018.