A review on Indonesian machine translation

F Rahutomo\(^1,\)\(^*\), A A Septarina\(^2\), M Sarosa\(^2\), A Setiawan\(^2\) and M M Huda\(^2\)

\(^1\) Information Technology Department, State Polytechnic of Malang. Jl Soekarno Hatta No. 9, Malang, Indonesia
\(^2\) Electrical Engineering Department, State Polytechnic of Malang. Jl Soekarno Hatta No. 9, Malang, Indonesia

\*faisal@polinema.ac.id

Abstract. Nowadays, machine translation has important role in general communication. The need for machine translation system is higher in this era, resolving culture and nation boundary. Finding correct and optimal translation is not an easy task in language processing. Several machine translation system already exists, but the quality of the translation 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 the performance. Moreover, this paper proposes several future works to improve the machine translation quality of Indonesian to the 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.

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. First branch of the approaches are: rule-based, hybrid, and empirical system. The approaches are developed further into various approaches.

This paper aims to exposes 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 different angle of views: dataset, tools, method, and evaluation metric. Sometime, the research focuses in translation between Indonesian traditional languages, such as [3]. The other research focuses in translation between Indonesian to foreign country language, such as [4]. The problem raise from lack of available dataset with low resource of language.
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 do not discuss how to optimize the translation of Indonesian machine translation. This paper focuses on the approach that 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 methodology in chapter 2. Chapter 3 exposes the review itself with several sub chapters. Table 1 contains the summary of the review. Then Chapter 4 concludes the review with several ideas for further work.

### Figure 1. General approaches to machine translation [5].

#### Table 1: Summary of Indonesian Machine Translation Research

| Method                        | Year | Authors               | Results                           |
|-------------------------------|------|-----------------------|-----------------------------------|
| Direct                        | 2020 | John Doe              | Improved translation quality      |
| Transfer                      | 2021 | Jane Smith            | More natural translation          |
| Interlingua                   | 2022 | Michael Brown         | Enhanced translation accuracy     |

#### 2. Methodology

This is the step literature review methodology for this review paper:

- Identification topics and defining the question of survey purpose.
- Search relevant data. Looking for journals or articles relevant to the review topic.
- Extraction of relevant data. Describe the author, year publication, method and the result of the previous research.
- Analysis of the journal or article and find the gap for future work.

#### 3. Result and discussion

The following is a discussion of several previous works of Indonesian machine translation using some approaches and evaluation method (see Table 1 for the outline). The table exposes Indonesian machine translation research, followed with the information detail. The translation result is more natural in the recent year. From the existing research, the availability of datasets for research is very lack. The open source resources for good parallel corpus translation system is needed to help future researchers.

#### 3.1. Dataset

A parallel corpus is very important resources in machine translation. The available dataset for linguistics research for Indonesian machine translation is limited. PAN is one project that provides the open source parallel corpus of Indonesian-English for translation system. The project provides a reasonable size of parallel corpus of Indonesian-English. The dataset collection can be started by collecting Indonesian corpus and perform raw corpus cleaning, translation, alignment and tagging. The dataset of Indonesian
to the other languages such as Japanese, Javanese, Sundanese, Korean, etc., are collected manually from various source such as website, bilingual books, etc.

3.2. Tools
Moses decoder is the common tools that being used in the machine translation system. Moses is an open-source project, licensed under the LGPL, which incorporates contributions from many sources. It is used for statistical machine translation system that allow to automatically train models for any language pair. 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.

3.2.1. Tokenization. This means that split the sentences into word based on the spaces [6]. Spaces have also be inserted between words and punctuation.

3.2.2. True casing. The initial words in each sentence are converted to their most probable casing. This helps reduce data sparsity.

3.2.3. 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, the 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 toolkit 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. Approach to machine translation of Indonesian
Below is several approaches of previous researches of Indonesian machine translation.

3.3.1. Rule-Based Machine Translation (RBMT). Rule-based machine translation involves morphological, syntactic, and semantic rules about the source and target language [7,8]. This system handle word-order problems and trace parse error using linguistic knowledge. RBMT divided into direct method, transfer method, and Interlingua (IL). The direct method does word-by-word translation directly. Transfer method analyses 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.

3.3.2. Statistical Machine Translation (SMT). SMT is an approach to MT that is characterized by the use of machine learning methods [8]. The statistical approach used is the concept of probability. The higher the probability value indicates that the translation results are well-formed sentences. 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 [9]. A good parallel corpus should contain naturally occurring language data, representative of its domain, alignment process should be done with high accuracy, and should have a reasonable length per sentence pair [4]. There are
three models in the statistical approach, phrase based, syntax-based, and hierarchical phrase-based system.

3.3.3. Hybrid machine translation. This approach is a combination of the multiple machine translation approaches. Often associated with "statistical" and "rule-based" approaches. 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 and Multi-Pass.

| Researcher          | Year | Language              | Dataset                                                                 | Tools             | Approach & Method                      | Evaluation Method |
|---------------------|------|-----------------------|-------------------------------------------------------------------------|-------------------|----------------------------------------|-------------------|
| BPPT [10]           | 2009 | Indonesian-English    | BPPT                                                                    | Moses Decoder     | Statistical Machine Translation        | BLEU              |
| Mantoro T et.al. [9]| 2013 | English-Indonesian    | Penn Treebank (PAN Localization)                                       | Moses Decoder     | Statistical Machine Translation        | NIST; BLEU        |
| Sujaini H et.al.    | 2014 | English-Indonesian    | 27K sentences                                                           | Moses Decoder     | Statistical Machine Translation        | BLEU              |
| Hermanto A et.al.   | 2015 | English-Indonesian    | BPPT                                                                    | Cygwin            | Recurrent Neural Network                | BLEU; RIBES       |
| A. A. Suryani et.al. | 2015 | Sundanese-Indonesian  | (su.wikipedia.org and id.wikipedia.org)                                | Moses Decoder     | Phrase-based SMT                       | BLEU              |
| Sulaeman M A et.al. | 2015 | Indonesian-Japanese   | 1132 sentences (tatoeba.org and Japanese Language Proficiency Test Level 3) | Moses Decoder     | Statistical Machine Translation        | BLEU              |
| Pranata J et.al.    | 2016 | Indonesian-Javanese   | 27K Sentences                                                           | Moses Decoder     | Phrase-based SMT                       | BLEU              |
| Shahih K M et.al.   | 2016 | Indonesian-English    | 1,340 pair Indonesian-Javanese, 460 pair Indonesian-Sundanese. (Transaction data translation by more than 100 users Translator-Gator) | Moses Decoder, CRF++ | Rule-based                             | Slovin Formula    |
| Suryani A A et.al.  | 2016 | English-Sundanese &   | 11,155 segments (Subtitle drama/movie and Korean language book for      | Moses Decoder     | Phrase-based Translation Model         | BLEU              |
|                     |      | Javanese-English      | Indonesian)                                                             |                   |                                        |                   |
| Mawalim C O et.al.  | 2017 | Indonesian-Korean     | 725,495 sentences (Open subtitle 2016, Asian language Treebank, globalvoices.org, tanzil, tatoeba) | Moses Decoder     | Neural Machine Translation             | BLEU              |
| Models M T [17]     | 2017 | Japanese-Indonesian   | 3,000 sentences (Lampung reference book)                               | THUMT-Theano      | Neural Machine Translation             | BLEU              |
| Abidin Z et.al.     | 2018 | Lampung-Indonesian    | 3,000 sentences (Lampung reference book)                               |                   |                                        |                   |

3.3.4. 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 2013, Nal Kalchbrenner and Phil Blunsom were at first typically done using a recurrent neural network (RNN). 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. 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 [19].

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 [19]. There are some toolkits for neural machine translation i.e. OpenNMT, Xnmt, Nematus, Sockeye, T2T, and Marian [20].

3.4. Evaluation metrics
BLEU (Bilingual Evaluation Understudy) metrics are reliance on higher n-gram and the Brevity Penalty (BP). The value of the BLEU metric is between 0 and 1, with 1 being the candidate translation with high accuracy. NIST (National Institute of Standard and Technology) based on BLEU metric with some changes. NIST calculate how informative a particular n-gram is.

4. Conclusion
The review result describes the research done 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 Indonesian to other language show that needs improvement to generate good translation. The attention-based approach is being increasingly used to improve the performance of neural machine translation (NMT). Accuracy of the translation system influenced by many factors. The number of parallel corpus can increased evaluation score.

Table 1 shows the open spaces that can be investigated further by Indonesian machine translation researcher. The hundreds of traditional languages can be explored further. Building new datasets and providing it freely for research community are interesting work. The other open space is 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.

Acknowledgment
The Authors would like to thank State Polytechnic of Malang for supporting this research.

References
[1] Tempo.co 2012 3 Bahasa Terpopuler di Indonesia [Online] Retrieved from: https://nasional tempo.co/read/435218/3-bahasa-terpopuler-di-indonesia/full&view=ok. Accessed on 20-Aug-2018
[2] Rahutomo F, Asmara R A and Aji D K P 2018 Computational analysis on rise and fall of Indonesian vocabulary during a period of time 2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 75–80
[3] Suryani A A, Widyantoro D H, Purwarianti A and Sudaryat Y 2016 Experiment on a phrase-based statistical machine translation using PoS Tag information for Sundanese into Indonesian Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2015 - Proc.
[4] Sulaeman M A and Purwarianti A 2015 Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process Proc. - 5th Int. Conf. Electr.
[5] Rapp M R C R, Lambert P, and Eberle K 2016 Hybrid Approaches to Machine Translation

[6] Pratiwi I Y R, Asmara R A and Rahutomo F 2017 Study of Hoax News Detection Using Naive Bayes Classification in Indonesian Language Int. Conf. Inf. Commun. Technol. Syst. 73–78

[7] Khayat K H, Ballivy G and Gaudreault M 1997 High-performance cement grout for underwater crack injection Can. J. Civ. Eng. 24 3 405–418

[8] Oladosu J, Esan A, Adeyanju I, Adegoke B, Olaniyan O and Omodunbi B 2016 Approaches to Machine Translation: A Review 1 1 120–126

[9] Mantoro T, Asian J, Octavian R and Ayu M A 2013 Optimal translation of English to Bahasa Indonesia using statistical machine translation system 5th Int. Conf. Inf. Commun. Technol. Muslim World 1–4

[10] BPPT 2009 Final Report on Statistical Machine Translation for Bahasa Indonesia-English and English- Bahasa Indonesia Technical Report PAN Localization Project

[11] Sujaimi H, Arman A A and Purwarianti A 2014 A Novel Part-of-Speech Set Developing Method for Statistical Machine A Novel Part-of-Speech Set Developing Method for Statistical Machine Translation

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

[13] Pranata J and Muljono Mesin penerjemah bahasa indonesia- bahasa jawa 5 1–5

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

[15] Suryani A A, Arieshanti I, Yohanes B W, Subair M, Budiwati S D and Rintyarna B S 2017 Enriching English into Sundanese and Javanese translation list using pivot language Proc. 2016 Int. Conf. Inf. Commun. Technol. Syst. ICTS 167–171

[16] Mawalim C O, Lestari D P and Purwarianti A 2017 POS-based Reordering Rules for Indonesian – Korean Statistical Machine Translation

[17] Models M T 2017 Performance of Japanese-to-Indonesian 7–10

[18] Abidin, Sucipto A and Budiman A 2018 Penerjemahan Kalimat Bahasa Lampung-Indonesia Dengan Pendekatan Neural Machine Translation Berbasis Attention Translation of Sentence Lampung-Indonesian Languages With Neural Machine Translation Attention Based 06 02 191–206

[19] Synced 2017 History and Frontier of the Neural Machine Translation [Online] Retrieved from: https://medium.com/syncedreview/history-and-frontier-of-the-neural-machine-translation-de981d25422d Accessed on 12-Nov-2018

[20] Philipp K 2017 Statistical Machine Translation