Pivot Approaches in Machine Translation for Low-Resource Languages

by

Sari Dewi Budiwati

Graduate School of Science and Technology
Kumamoto University

July 22, 2021

Abstract

This thesis addresses a machine translation (MT) in two low-resource language pairs, namely, Kazakh to English (Kk-En) and Japanese to Indonesian (Ja-Id). The Kk-En and Ja-Id is considered as a low-resource language due to its limited parallel corpora. Low resource language is a state where two language pairs have limited parallel corpora and linguistic tools such as tokenizer, morphological tools, lemmatizer, and pos-tagger. As most languages in Asia are still considered as a low-resource, it becomes an essential task in MT to improve its translation quality.

In this study, we explore the pivot approach in Statistical Machine Translation (SMT) model to improve the translation quality of Kk-En and Ja-Id. Pivot approach is a strategy that uses a third language as a bridge to overcome the parallel corpora limitation. We explore two types of pivot approaches, viz., single and multiple pivots. The single pivot uses one language as a pivot, whereas multiple pivots use more than one languages. We employ three strategies in the single pivot, viz., cascade, triangulation, and interpolation. Whereas in multiple pivots, we employ interpolation strategy based on ascending and descending BLEU scores.

As a preliminary effort, we did two explorations, viz., single-pivot in Kk-En and Ja-Id, and multiple pivots in Ja-Id. We find that multiple pivots approach could outperform the direct and single pivot system. However, we find that our generated text followed the source language sentence pattern, i.e., Subject-Object-Verb (SOV), whereas the target language is Subject-Verb-Object (SVO). Thus, our generated text was not comprehensible and hard to understand.

Subsequently, we propose two strategies, viz., extending source–pivot (src–pvt) phrase table, and phrase table combination based on symmetrization of word alignment. The extending src–pvt phrase table is a merging of two phrase tables of src–pvt, viz., src–pvt *gdfand* and src–pvt *tgttosrc*. These techniques arise based on our finding that the *tgttosrc* phrase table has a candidate phrase pair that could not be obtained by the *gdfand*. We employ this technique in multiple pivots of Ja–Id. We also implement the *pre-ordering* of the Japanese dataset to overcome the issue of different word orders between Japanese and Indonesian languages. As a result, our generated text could be more understandable compared to the non *pre-ordered* one.

Our second proposed strategy is a phrase table combination based on symmetrization of word alignment. These techniques come based on the fact that the pivot approach comprises three direct translations, viz., src–trg, src–pvt, and pvt–trg, that obtain different BLEU scores when different symmetrizations are employed. Therefore, we did phrase table combinations based on the highest BLEU scores from the direct of system approach (DSA) for each phrase table of src–trg, src–pvt, and pvt–trg. We find that our strategy could be a competitive approach because it outperforms direct translation in Kk−En with absolute improvements of 0.35 and 0.22 BLEU points for 3-gram and 5-gram, respectively. Our proposed strategy also outperforms the direct translation of 3-gram in Ja−Id with an absolute improvement of 0.11 BLEU point.

In addition, we conducted NMT experiments to compare with SMT. We used two pre-trained models, viz., encoder-decoder and transformer. We present empirical results to show that SMT outperformed the NMT for Ja-Id as low-resource language pair. SMT system obtained the highest BLEU score, that is, 11.96, even with a small dataset, that is, 8.5K ALT dataset, while the NMT system obtained a BLEU score of 7.8 with an additional dataset, that is, 100K of the TEDTalk. Our results indicates that the SMT obtained better results that NMT, even with a small dataset.

Acknowledgements

I would like to say Alhamdulillahirobbilalamiin, thanks to Allah SWT for being forgiving, merciful, and give me the opportunity to have a meaningful life until now. I would like to express my special gratitude to my supervisors, Professor Masayoshi Aritsugi who gave me the opportunity to pursue my study at Kumamoto University. Deeply appreciate for your continuous support, excellent guidance, care, and patience to forge me as a researcher.

I would like to thank my thesis committee, Professor Tsuyoshi Usagawa, Professor Masahiro Iida, and Professor Masahiko Nishimoto, also all the professor in Computer Science and Electrical Engineering Department of Kumamoto University for their support and valuable knowledge. My appreciation extends to all DBMS lab members for their support during my study and for all memories that we shared together.

I would also like to thank my husband, daughter, parents, and all my family member for their care and support. I would also like to thank my colleagues in Telkom University, particularly Mrs. Arie Ardiyanti for all the discussion and suggestion.

Last but not least, deeply thankful for the Ministry of Education, Culture, Sports, Science and Technology (MEXT) Japan, who give me the funding for study through its scholarship.

Contents

List of Figures

List of Tables

List of Abbreviations

SMT	Statistical Machine Translation
NMT	Neural Machine Translation
LM	Language Model
TM	Translation Model
RM	Reordering Model
MERT	Minimum Error Rate Training
MIRA	Margin Infused Relaxed Algorithm
PRO	Pairwise Ranked Optimization
LI	Linear Interpolation
FI	Fillup Interpolation
MDP	Multiple Decoding Path
BLEU	Billingual Evaluation Understudy
WMT	Workshop on Machine Translation
DBMS-KU	Database Management System - Kumamoto University
Kk-En	Kazakh to English
En-Kk	English to Kazakh
Ja-Id	Japanese to Indonesian
Id-Ja	Indonesian to Japanese
SOV	Subject-Object-Verb
SVO	Subject-Verb-Object
VOS	Verb-Object-Subject
src-pvt	source to pivot
pvt-trg	pivot to target

src-trg	source to target
ALT	Asian Language Treebank
OPUS	Open Parallel corpus
OOV	Out Of Vocabulary
UNK	Unknown words
WoR	Without Reordering
WR	With Reordering
DSA	Direct System Approach
ISA	Interpolation System Approach
Std-ISA	Standard-Interpolation System Approach
H-ISA	Highest-Interpolation System Approach
NE	Named Entity
Lader	Latent Derivation Reorderer
RNN	Recurrent Neural Network
LSTM	Long Short-term Memory
ReLu	Rectified Linear unit

Chapter 1

Introduction

Machine Translation (MT) is a task of automatically translate a text from one natural language, i.e., English, to another language, i.e., Japan. The state-of-the-art of MT is Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) models. SMT is an approach that uses probabilistic models of faithfulness and fluency and then combining these models to choose the most probable translation (Jurafsky and Martin, 2009). In comparison, NMT based on a neural network model that consists of encoder-decoders. The encoder of neural network reads and encodes a source sentence into a fixed-length vector, while a decoder outputs a translation from the encoded vector (Bahdanau et al., 2015).

Koehn and Knowles (Koehn and Knowles, 2017) compared two models and stated that the NMT model still has to overcome various challenges, most notably in the performance of out-of-domain and low resource conditions. The out-of-domain experiment showed that the NMT perform poorly by a BLEU score of less than 1.0, whereas the SMT could be obtained by more than 2.1. The NMT also showed the lowest BLEU score compare to SMT, viz., 1.6 and 16.4, respectively, in the small dataset, i.e., 376K. Moreover, the experiment also showed that the NMT generated text is entirely unrelated to the input if the training dataset less than a few million words. Thus, SMT is a better option for low resource conditions.

Low resource condition is a state where two language pairs have limited parallel corpora. The low resource also implies that a particular language still has limited linguistic tools, i.e., tokenizer, morphological tools, lemmatizer, and pos-tagger. The linguistic tools used in the pre-processing before dataset trained and evaluated. The tools are useful for word separation and segmentation, particularly for agglutinative language that words are formed by joining suffixes, i.e., Kazakh, Indonesian. Some researchers showed that the BLEU score obtained higher by using segmentation or pos-tag dataset (Assylbekov and Nurkas, 2014; Simon and Purwarianti, 2013; Sulaeman and Purwarianti, 2015). Nevertheless, these tools often unpublished in their research.

With the limited parallel corpora, there are two strategies to achieve high-quality translations in SMT, namely building parallel corpora and utilizing existing corpora (Trieu, 2017). Building parallel corpora are challenging since it can be time-consuming and expensive, and needs experts (Wołk and Wołk, 2018). Therefore, researchers have focused on utilizing existing corpora, i.e., using pivot approaches (Ahmadnia et al., 2017; Dabre et al., 2015; El Kholy et al., 2013; Habash and Hu, 2009; Paul et al., 2009; Trieu and Nguyen, 2017; Utiyama and Isahara, 2007). Instead of direct translation between a language pair, pivot approaches use the third language as a bridge to overcome the parallel corpora limitation. Pivot approaches arise as a preliminary assumption that there are enough parallel corpora between source–pivot (src–pvt) and pivot–target (pvt–trg) languages.

In this study, we explore pivot approaches in the SMT model for two low-resource languages, viz., Kazakh to English (Kk–En), and Japanese to Indonesian (Ja–Id). The available open parallel corpora of Kk–En is 953,240, whereas Ja–Id is 1,468,155 parallel sentences. Furthermore, the linguistic tools of Kazakh and Indonesian still limited or unpublished. It makes both language pairs considered as low-resource language pairs. We explore three pivot approaches, viz., cascade, triangulation, and interpolation. The cascade approach is a technique that uses two systems, namely src–pvt and pvt–trg. The triangulation technique combines the src–pvt and pvt–trg phrase tables called the triangulation phrase table. Whereas the interpolation is a combination of src–trg and triangulation phrase tables. The interpolation is also known as the phrase table combination technique. In addition, we conducted NMT experiments to compare with SMT. We used OpenNMT-py framework (Klein et al., 2017) with two pre-trained models, viz., encoder-decoder and transformer. The encoder-decoder model is a default pre-trained model of OpenNMT-py that uses multiple recurrent neural network (RNN) cells and attention types (Bahdanau et al., 2015; Luong et al., 2015). The transformer model is a pre-trained model of OpenNMT-py based on the Google transformer model (Vaswani et al., 2017).

1.1 Objectives

This work aims to apply pivot approaches and examine issues in two low-resource language pairs, viz., Kk–En and Ja–Id. To the best of our knowledge, the pivot approaches never implemented on that two low-resource language pairs, except the Ja–Id that uses the cascade approach (Paul et al., 2013). We explored two types of pivot approaches in preliminary works, namely single and multiple pivots. The single pivot applied to Kk–En and Ja–Id, whereas the multiple pivots applied to Ja–Id.

Another objective from this work is to propose a technique that could improve the translation quality compare to the direct translation as the Baseline system. In this work, we measure the translation quality by BLEU score and observe the generated text using POS (Part of Speech)-tag from a particular target language, i.e., Indonesian Pos-tagger for Indonesian.

1.2 Contribution

The main contributions of this work can be presented as follows:

1. In multiple pivots of Ja–Id, we proposed an extending src–pvt phrase table before the phrase table combination process. These techniques arise based on our finding that the non-standard symmetrization has candidate phrase pair that could not be obtained by the standard one. Therefore, we merge two phrase tables of src–pvt, viz., src–pvt *gdfand* and src-pvt *tgttosrc*, called extending phrase table. In multiple pivots, we also employed the *pre-ordering* process for Ja dataset to overcome the issue of different word order between Japanese and Indonesian languages, i.e., SOV and SVO, respectively.

2. In the single pivot, we proposed a strategy, i.e., phrase table combination, that uses symmetrization of word alignment, which obtains the highest BLEU scores. We named our strategy as highest-interpolation system approach (H-ISA). We find that the H-ISA could be a competitive approach because it outperforms direct translation in Kk-En. Our strategy also outperforms the direct translation of 3-gram in Ja-Id.

1.3 Dissertation Outline

The rest of this thesis is organized and continued with Chapter 2, which describes the SMT model, pivot approaches and NMT model that used in the experiments. Chapter 3 focuses on the single pivot and multiple pivots approach for Kk–En and Ja–Id. Subsequently, Chapter 4 describes the word reordering in multiple pivots for Ja–Id. We also describe our comparison result of phrase table combination experiments between Kk–En and Ja–Id. At last, Chapter 5 discusses the conclusion and future work of our study.

Chapter 2

Basic Theory and Related Work

In this chapter, we present necessary background knowledge of the main topic and methods in this dissertation, including SMT, pivot approach, and NMT models. We also describe the current research result of Kk–En and Ja–Id language pairs.

2.1 Statistical Machine Translation

Statistical Machine Translation (SMT) is an approach that uses probabilistic models of faithfulness and fluency and then combining these models to choose the most probable translation (Jurafsky and Martin, 2009). Language model and translation model are the two-part that form the basis of SMT. Language model is used to determine the fluency of translation output or generated text. Whereas, translation model is used as a connection between source and target language or known as phrase pair.

Figure 2.1 illustrates the architecture and process in SMT. The SMT architecture consists of translation model (TM), language model (LM), reordering model (RM), and decoder. At the same time, the SMT process consists of training, tuning, and evaluation. TM or phrase table obtain from parallel corpora then mapped by word alignment model between the source and the target words (Jurafsky and Martin, 2009). The baseline word alignment model does not allow a source word aligned with more than one target word (Koehn et al., 2005). Therefore, it needs to train from both directions, i.e., source–to–target and target–to–source. Additionally, the symmetrization technique was performed to increase the quality of word alignment using various combination methods. Let A_{TS} and A_{ST} represent the two alignments in target–to–source (TS) and source–to–target (ST), respectively, then the various combination methods are (Och and Ney, 2003; Wu Hua, 2007):

- Intersection (I) is a method that preserved the word alignment points occurred in both alignments ($A_{\mathrm{TS}} \cap A_{\mathrm{ST}}$).

- Union (U) is a method that joins the word alignment points from both alignments ($A_{\mathrm{TS}} \cup A_{\mathrm{ST}}$).
- Grow is a method that adds word alignment points from left, right, top or bottom neighborhood.
- Grow diagonally (grow-diag) is a method that adds word alignment points from the diagonal neighborhood.
- Grow-diag-final (gdf) is an additional method from grow-diag that adds the non-neighboring alignment points between words, of which at least one is currently unaligned.
- Grow-final (gf) is an additional method from grow that adds the non-neighboring alignment points between words, of which at least one is currently unaligned.
- Source to target *(srctotgt)* is a method that only considers word-to-word alignments from the source-target GIZA++ alignment file [1];
- Target to source *(tgttosrc)* is a method that only considers word-to-word alignments from the target-source GIZA++ alignment file [1].

The symmetrization technique above then will be aligned and resulting phrase-pair with the phrase translation probabilities as follows (Jurafsky and Martin, 2009):

$$\theta(\overline{f},\overline{e}) = \frac{count(\overline{f},\overline{e})}{\sum_{\overline{\mathrm{f}}} count(\overline{f},\overline{e})} \tag{2.1}$$

where $count(\overline{f},\overline{e})$ describes the frequency of the phrase $\overline{f}$ is aligned with the phrase $\overline{e}$ in the parallel corpus.

Along with the word alignment in the training phase, the language model also trained to determine the fluency of the generated text. The LM obtain from monolingual corpora that relatively easy to collect compared to TM. The standard LM approach in SMT is *n*-gram that can be used by various order, viz., unigram (1-gram), bigram (2-gram), trigram (3-gram). The choice of the LM order will determine the translation quality and model size (Liu et al., 2014). The longer the LM order then the model size will be bigger. The *n*-gram probabilities score is measured based on w as a word sequence, as follows (Jurafsky and Martin, 2009):

$$P(w_{\mathrm{n}}|w_{\mathrm{n\text{-}1}}) = \frac{C(w_{\mathrm{n\text{-}1}}w_{\mathrm{n}})}{C(w_{\mathrm{n\text{-}1}})} \tag{2.2}$$

After the training process, the tuning algorithm was tuned to find the best feature weights for the decoding process. The decoding process is a process to

[1] `http://www.statmt.org/moses/?n=FactoredTraining.AlignWords`

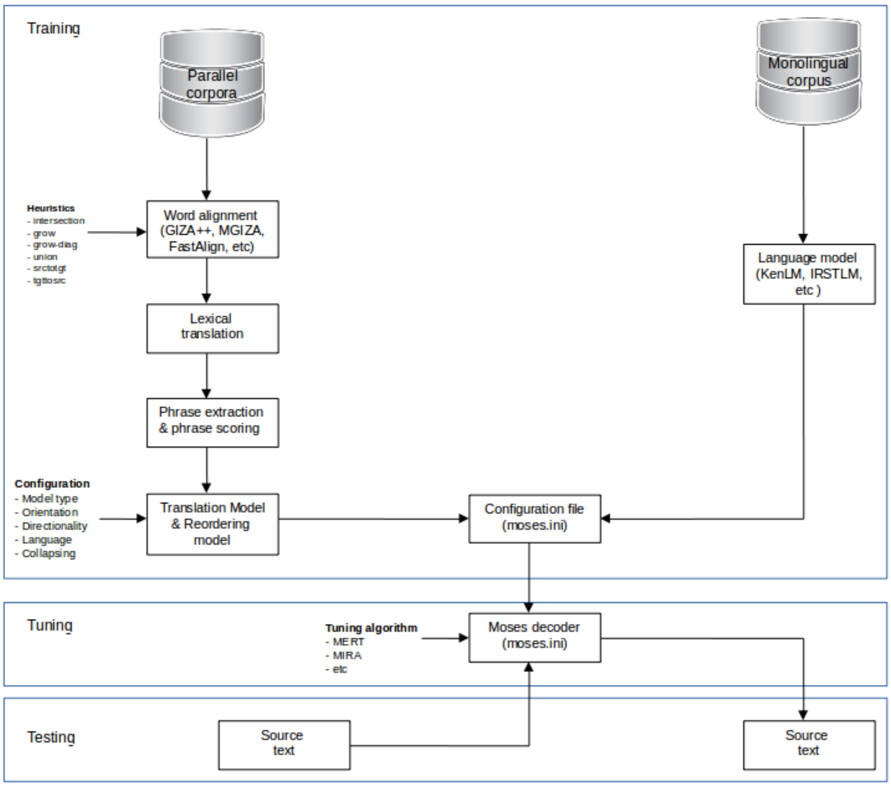


Figure 2.1: SMT architecture

find maximum value by multiplying the feature functions weight with TM and LM probability scores. The SMT system outputs the best target translation $\mathrm{t}_{\mathrm{best}}$ as follows:

$$\begin{aligned} t_{\text{best}} &= \arg\max_t p(t|s) \\ &= \arg\max_t \sum_{m=1}^{M} \lambda_{\mathrm{m}} h_{\mathrm{m}}(t|s) \end{aligned} \tag{2.3}$$

where $h_m(t|s)$ represents feature function, and λ_{m} is the weight assigned to the corresponding feature function (Wu and Wang, 2007). The feature function $h_m(t|s)$ consist of language model probability of target language, phrase translation probabilities (both directions), lexical translation probabilities (both directions), a word penalty, a phrase penalty, and a linear reordering penalty. The weight (λ_{m}) can be set by tuning algorithm such as Minimum Error Rate Training (MERT)(Och, 2003), MIRA (Margin Infused Relaxed Algorithm) (Chiang, 2012), PRO (Pairwise ranked optimization) (Hopkins and May, 2011), and k-best MIRA (Cherry and Foster, 2012). The process called tuning, as shown in Figure 2.1.

2.2 Pivot approach

Pivot approach is a translation from a source language (src) to a target language (trg) through an intermediate pivot language (pvt) (Paul et al., 2009). Pivot approach arise as a preliminary assumption that there are enough parallel corpora between src–pvt and pvt–trg languages. Several pivot approaches are sentence translation, triangulation and interpolation.

2.2.1 Sentence translation

The sentence translation strategy or cascade uses two independently trained SMT systems (Utiyama and Isahara, 2007). These two independently systems are src–pvt and pvt–trg systems. First, given a source sentence s, then translate it into n pivot sentences p_1, p_2, ..., p_{n} using an src-pvt system. Each p_{i} has eight scores namely language model probability of the target language, two phrase translation probabilities, two lexical translation probabilities, a word penalty, a phrase penalty, and a linear reordering penalty. The scores are denoted as $h^{\mathrm{e}}{}_{\mathrm{i1}}$, $h^{\mathrm{e}}{}_{\mathrm{i2}}$, ..., $h^{\mathrm{e}}{}_{\mathrm{i8}}$. Second, each p_{i} is translated into n target sentences t_{i1}, t_{i2}, ..., t_{in} using a pvt-trg system. Each t_{ij} (j= 1, ..., n) also has the eight scores, which are denoted as $h^{\mathrm{t}}{}_{\mathrm{ij1}}$, $h^{\mathrm{t}}{}_{\mathrm{ij2}}$, ..., $h^{\mathrm{t}}{}_{\mathrm{ij8}}$. The situation is as follows:

$$\begin{aligned} \textit{SRC-PVT} &= p_{\mathrm{i}}(h^{\mathrm{e}}{}_{\mathrm{i1}}, h^{\mathrm{e}}{}_{\mathrm{i2}}, ..., h^{\mathrm{e}}{}_{\mathrm{i8}}) \\ \textit{PVT-TRG} &= t_{\mathrm{ij}}(h^{\mathrm{t}}{}_{\mathrm{ij1}}, h^{\mathrm{t}}{}_{\mathrm{ij2}}, ..., h^{\mathrm{t}}{}_{\mathrm{ij8}}). \end{aligned} \tag{2.4}$$

We define the score of t_{ij}, $S(t_{ij})$, as

$$S(t_{\mathrm{ij}}) = \sum_{m=1}^{8} (\lambda^{\mathrm{e}}{}_{\mathrm{m}} h^{\mathrm{e}}{}_{\mathrm{im}} + \lambda^{\mathrm{t}}{}_{\mathrm{m}} h^{\mathrm{t}}{}_{\mathrm{ijm}}) \tag{2.5}$$

where $\lambda^{\mathrm{e}}{}_{\mathrm{m}}$ and $\lambda^{\mathrm{t}}{}_{\mathrm{m}}$ are weights set by performing minimum error rate training (Och, 2003). Finally, $\mathrm{t}_{\mathrm{best}}$ will be

$$t_{\mathrm{best}} = \arg\max_{t_{\mathrm{ij}}} S(t_{\mathrm{ij}}). \tag{2.6}$$

2.2.2 Triangulation

Triangulation, or known as phrase table translation is an approach for constructing an src–trg translation model from src–pvt and pvt–trg translation models (Hoang and Bojar, 2016a). First, we train two translation models for src–pvt and pvt–trg, respectively. Second, we build an src–pvt translation model with **p** as a pivot language. The src–pvt translation model also known as triangulation translation

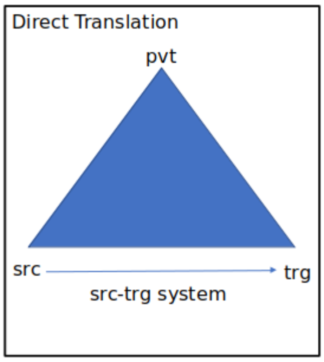


Figure 2.2: Direct translation approach

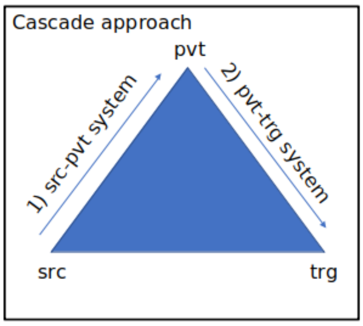


Figure 2.3: Cascade approach

model or triangulation phrase table, as shown in Figure 2.4.

Given a sentence **p** in the pivot language, the pivot translation model can be formulated as follows (Wu and Wang, 2007):

$$\begin{aligned} p(\mathbf{s}|\mathbf{t}) &= \sum_{p}(p(\mathbf{s}|\mathbf{t},\mathbf{p}))p(\mathbf{p}|\mathbf{t}) \\ &\approx \sum_{p}(p(\mathbf{s}|\mathbf{p}))p(\mathbf{p}|\mathbf{t}) \end{aligned} \tag{2.7}$$

where **s** and **t** are source and target translation model, respectively.

The triangulation translation model is often combined with src–trg translation model, called interpolation approach. The interpolation also known as phrase table combination. We used two types of interpolation in this study, namely Linear Interpolation (LI) and Fillup Interpolation (FI). The LI is performed by merging the tables and computing a weighted sum of phrase pair probabilities from each phrase table giving a final single table. Fillup Interpolation does not modify phrase probabilities but selects phrase pair entries from the next table if they are not present in the current table.

More than one pivot language can be used to improve the translation quality, called multiple pivots. If we use n pivot languages and combine with src–trg translation model, then the estimation of phrase translation probability and the lexical weight are as follows (Ahmadnia et al., 2017):

$$P(s|t) = \sum_{i=1}^{n} \alpha_{\mathrm{i}} P_{\mathrm{i}}(s|t) \tag{2.8}$$

$$P(s|t,\alpha) = \sum_{i=1}^{n} \beta_{\mathrm{i}} P_{\mathrm{i}}(s|t,\alpha) \tag{2.9}$$

where $P(s|t)$ and $P(s|t,\ \alpha)$ are the phrase translation probability and the lexical weight trained with src–trg corpus estimated by using pivot language, while α_{i} and

β_{i} are interpolation coefficients. Last, $\sum_{i=1}^{n} \alpha_{\mathrm{i}} = 1$, and $\sum_{i=1}^{n} \beta_{\mathrm{i}} = 1$.

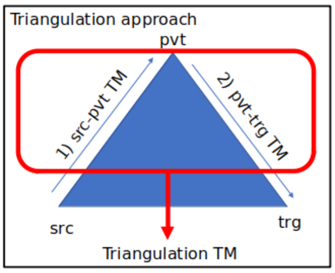


Figure 2.4: Triangulation approach

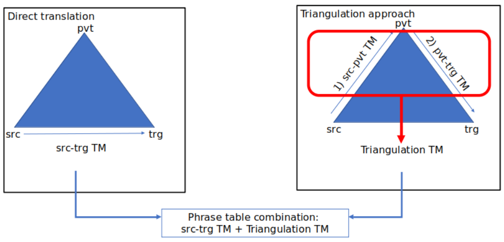


Figure 2.5: Phrase table combination approach

2.3 Neural Machine Translation

Neural machine translation (NMT) is an approach based on a neural network model that consists of encoder-decoders. Early models of NMT are convolutional, and sequence-to-sequence models. Then, the addition of the attention mechanism become competitive models with a few refinements, such as byte pair encoding and back-translation of target monolingual data (Koehn, 2020). An encoder neural network reads and encodes a source sentence into a fixed-length vector, while a decoder then outputs a translation from the encoded vector (Sutskever et al., 2014), as shown in Figure 2.6. The whole encoder-decoder system is jointly trained to maximize the probability of a correct translation given a source sentence. In this study, we used two NMT models in our experiments, namely recurrent neural network (RNN) encoder-decoder, and transformer.

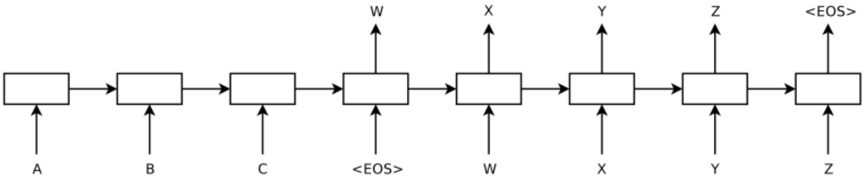


Figure 2.6: An encoder-decoder model (Sutskever et al., 2014).

2.3.1 RNN encoder-decoder

We describe a recurrent neural network (RNN) encoder-decoder, based on (Bahdanau et al., 2015) and (Luong et al., 2015). The RNN is a natural generalization of feed forward neural networks to sequences (Sutskever et al., 2014). In the encoder-decoder framework, an encoder reads the input sentence, a sequence of vectors $x = (x_1, ..., x_{\mathrm{T}x})$, into a vector c. The RNN approach is as follows:

$$h_(t) = f(x_{\mathrm{t}}, h_{\mathrm{t\text{-}1}}) \tag{2.10}$$

and

$$c = q(\{h_1, ..., h_{\mathrm{T}x}\}) \tag{2.11}$$

where $h_{\mathrm{t}} \in \mathbb{R}^{\mathrm{n}}$ is a hidden state at time t, and c is a vector generated from the sequence of the hidden states. f and q are some nonlinear functions, i.e., long short-term memory (LSTM) Sutskever et al. (2014).

The decoder is used to predict the next word $y_{\mathrm{t'}}$ given the context vector c and all the previously predicted words $y_1, ..., y_{\mathrm{t'\text{-}1}}$. The decoder defines a probability over the translation y by decomposing the joint probability into the ordered conditionals:

$$p(y) = \prod_{t=1}^{T} p(y_{\mathrm{t}} \mid \{y_1, ..., y_{\mathrm{t\text{-}1}}\}, c) \tag{2.12}$$

where $y = (y_1, ..., y_{\mathrm{T}y})$. With an RNN, each conditional probability is modeled as

$$p(y_{\mathrm{t}} \mid \{y_1, ..., y_{\mathrm{t\text{-}1}}\}, c) = g(y_{\mathrm{t\text{-}1}}, s_{\mathrm{t}}, c) \tag{2.13}$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_{t}, and s_{t} is the hidden state of the RNN.

Luong et al. (2015) experimented a multiple recurrent neural network and attention types, as shown in Figure 2.7, with formulation is as follows:

$$J_{\mathrm{t}} = \sum{}_{(}x, y) \in D - \log p(y|x) \tag{2.14}$$

where D is a parallel training corpus, and $\log p(y|x)$ is a conditional probability of a source to a target sentence. The model is a feature of encoder-decoder model in an OpenNMT-py (Klein et al., 2017) framework. They used two attention types, namely, global and local attentions. Global attention considered all source words, whereas local attention only considered a subset of source words at a time.

2.3.2 Transformer model

The transformer is a model which eschewed recurrence and relied on an attention mechanism to draw global dependency between input and output (Vaswani et al., 2017). The transformer used stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, as shown in Figure 2.8. The encoder is composed of 6 identical layers (N_{x}). Each layer has two sub-layers, namely, a multi-head self-attention mechanism, and a position-wise fully connected feed-

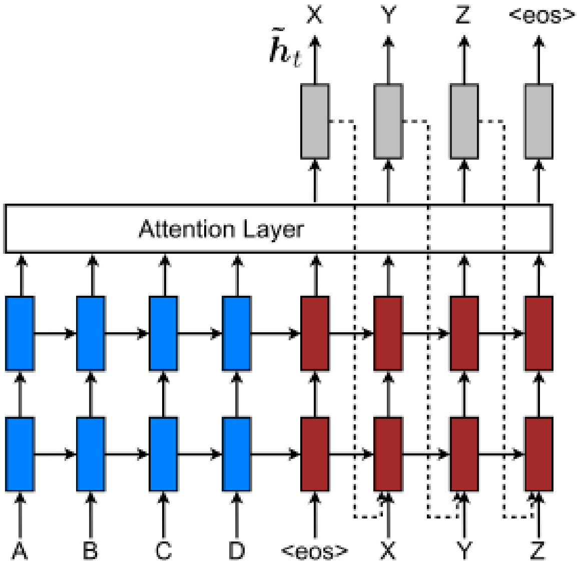


Figure 2.7: A stacking recurrent encoder-decoder model (Luong et al., 2015).

forward network. The residual connection (Add) and layer normalization (Norm) is employed in every two sub-layers. The decoder also consists of 6 identical layers and two sub-layers. In addition, one sub-layer is added that employed a masked multi-head self-attention. This attention is a modified multi-head self-attention to ensure that the predictions for position i can depend only on the known outputs at positions less than i.

An attention function is a process of mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The transformer model used two attentions, namely, scaled dot-production and multi-head attention. Scaled dot-production attention consists of queries and keys of dimension d_{k}, and values of dimension d_{v}. Then these values computed by dot products of the query with all keys, divide each by $\sqrt{d_{\mathrm{k}}}$, and apply a softmax function to obtain the weights on the values. Vaswani et al. (2017) compute the attention function on a set of queries simultaneously, packed together into a matriq Q. The keys and values are also packed together into matrices K and V, as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}})V \tag{2.15}$$

Further, Vaswani et al. (2017) performed a linearly project the queries, keys, and values h times with different, learned linear projections to d_{q}, d_{k}, and d_{v} dimensions, respectively. These linearly projection called multi-head attention. On each projection, they performed the attention function in parallel, yielding d_{v}-dimensional output values. Then, the dimensional output values are concatenated. Multi-head attention allows the model to jointly attend to information from different represen-

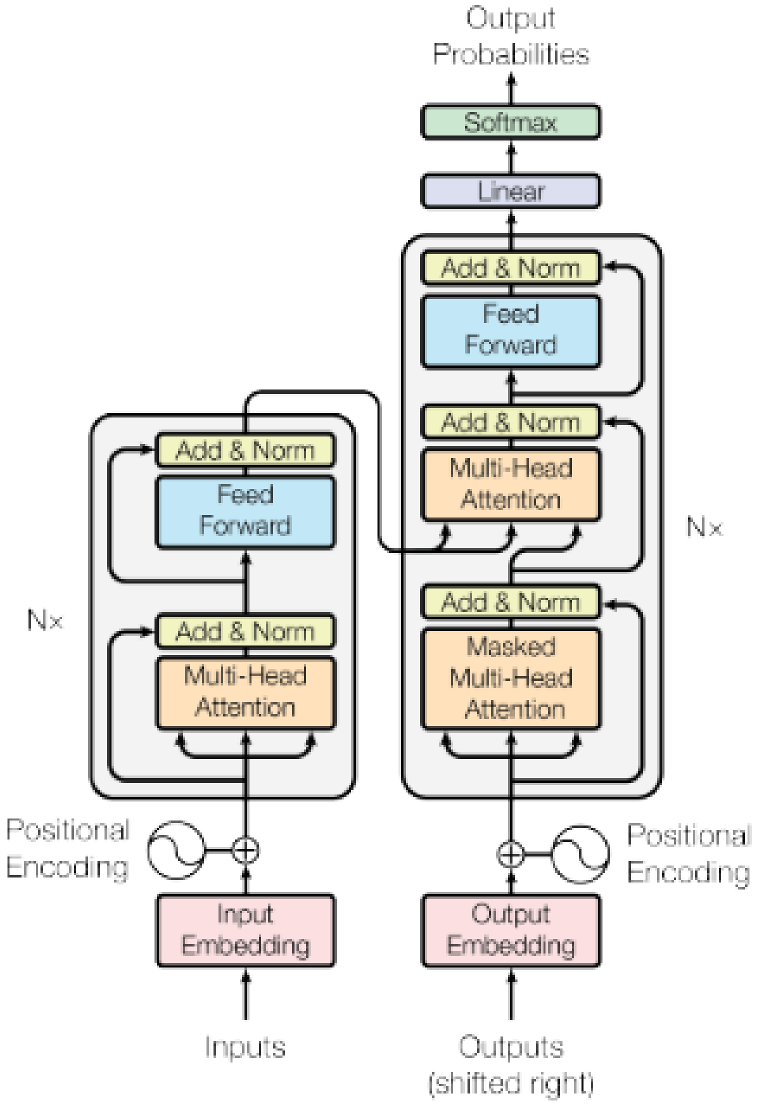


Figure 2.8: The transformer model (Vaswani et al., 2017).

tation subspaces at different positions, as follows:

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, ..., head_{\mathrm{h}})W^{\mathrm{O}} \\ head_{\mathrm{i}} &= Attention(QW_{\mathrm{i}}{}^{\mathrm{Q}}, KW_{\mathrm{i}}{}^{\mathrm{K}}, VW_{\mathrm{i}}{}^{\mathrm{V}}) \end{aligned} \tag{2.16}$$

Where the projections are parameter matrices $W_{\mathrm{i}}{}^{\mathrm{Q}} \in \mathbb{R}^{\mathrm{d_{model}x\ d_k}}$, $W_{\mathrm{i}}{}^{\mathrm{K}} \in \mathbb{R}^{\mathrm{d_{model}x\ d_k}}$, $W_{\mathrm{i}}{}^{\mathrm{V}} \in \mathbb{R}^{\mathrm{d_{model}x\ d_k}}$.

Figure 2.8 shows a feed forward sub-layers in each encoder and decoder. Feed forward network consists of two linear transformations with a rectified linear unit (ReLu) activation function in between, as follows (Vaswani et al., 2017):

$$FFN(x) = max(0, xW_1 + b_1) + W_2 + b_2 \tag{2.17}$$

2.4 Previous work

2.4.1 Phrase table combination

Phrase table combination has been used to improve the translation quality in direct translation and pivot approaches. In direct translation, researchers have combined two or more phrase tables of direct translation with various symmetrizations of word alignment (Singh, 2015; Wu Hua, 2007). Wu Hua (2007) used two approaches in their direct translation of Spanish–English: the direct translation of single symmetrization and phrase table combination. The first approach used three symmetrizations: gdfand, grow-diag, and intersection. The second approach combined three phrase tables produced by gdfand, grow-diag, and intersection. Their experimental results showed that phrase table combination outperformed the direct translation of a single symmetrization approach. Singh (2015) explored massive experiments, also arriving at the same conclusion. They combined three to eight phrase tables of direct translation in the Chinese–English language pair. The phrase tables were produced by several symmetrizations of word alignment: gdfand, intersection, union, grow-diag-final, grow-diag, grow, srctotgt, tgttosrc, and grow-final. Their experimental results showed that the phrase table combination of several symmetrizations outperformed the Baseline.

Phrase table combination could also be used in pivot approaches to overcome the data scarcity between src–trg language pairs (Ahmadnia and Serrano, 2017; Ahmadnia et al., 2017, 2018; Budiwati and Aritsugi, 2019; Dabre et al., 2015; Utiyama and Isahara, 2007; Wu and Wang, 2007). The phrase table combination outperformed direct translation or other pivot approaches, namely, sentence translation and triangulation. We identified that the phrase table combination in the pivot approaches uses the same symmetrization of word alignment, i.e., gdfand (Ahmadnia and Serrano, 2017; Ahmadnia et al., 2017, 2018; Budiwati and Aritsugi, 2019; Dabre et al., 2015; Utiyama and Isahara, 2007; Wu and Wang, 2007). This is because the gdfand is a standard symmetrization of word alignment in statistical machine translation (SMT) (Girgzdis et al., 2014). Then, Kholy and Habash (2014) proposed the phrase table combination that combined two symmetrization of word alignments: GDFA (grow-diag-final-and) and GDFA_R (grow-diag-final-and_Relaxation), for the Hebrew–Arabic language pair. GDFA is a symmetrization of word alignment that starts with an intersection of two alignments. Then, the symmetrization adds alignment points between two unaligned words (Koehn et al., 2005). GDFA is a common symmetrization of word alignment used in word alignment tools such as GIZA++. GDF_R was constructed by Kholy and Habash (2014) using the following steps:

1. Creating a list of all possible pivot unigrams using the intersections of the source–pivot and pivot–target corpora.

2. Building two-directional alignments model using grow-diag-final-and in pivot–target, namely, pivot-to-target $\overrightarrow{A_{pt}}$ and target-to-pivot $\overleftarrow{A_{pt}}$, subsequently combining the two-directional alignment models. Last, obtaining the final alignment metrics of a new phrase obtained by removing a given word unlisted in the pivot unigram. This second part was also applied to source–pivot.

The system combination of the GDFA and GDF_R symmetrizations obtained an absolute improvement of 0.8 BLEU points compared to GDFA_R, which verified the superiority of the phrase table combination.

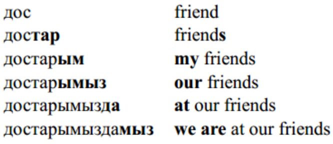
From those studies, the application of the phrase table combination could improve the translation quality. Nevertheless, the symmetrization of word alignment can vary across language pairs and datasets (language- and dataset-specific). Koehn et al. (2005) explored five symmetrizations (final, final-and, grow-diag, grow, and intersection) in five language pairs (Arabic–English, Japanese–English, Korean–English, Chinese–English, and English–Chinese). Given an example of Japanese–English, they obtained the highest BLEU score of 45.1 when using intersection, and the lowest BLEU score of 39.0 when using grow-diag. Stymne et al. (2014) explored five symmetrizations (intersection, grow-diag, grow-diag-final-and, grow-diag-final, and union) in German–English. Their experimental result showed that the highest BLEU score was obtained by grow-diag with a score of 20.9, and the lowest BLEU score was obtained by intersection with a score of 19.1. Those investigations showed that the best symmetrization of word alignment can differ for each language pair. This conclusion confirmed the language-specific characteristic of the symmetrization of word alignment.

Symmetrization of word alignment is also dataset-specific because the same language pairs with different datasets can obtain different BLEU scores (Singh, 2015; Wu Hua, 2007). Wu Hua (2007) explored two types of test sets: in-domain and out-of-domain. The exploration was conducted using six symmetrizations: grow-diag-final, grow-final, union, grow-diag, grow, and intersection. In-domain means the test set type was the same domain as the training dataset. In contrast, out-of-domain uses a different type of test set from the training dataset. Their experiment was implemented in three language pairs, Spanish–English, French–English, and Dutch–English. In an an example of Spanish–English, a higher BLEU score of 30.63 was obtained using grow-final in the in-domain test set. Using the same symmetrization, Spanish–English obtained a lower BLEU score of 25.00 in the out-of-domain test set. Their results suggested that the symmetrization of word alignment can differ for each type of test set (dataset-specific). Singh (2015) also suggested the same conclusion. They explored six datasets: NIST2002, NIST2003, NIST2004, NIST2005, NIST2006, and NIST2008. The test was conducted using nine symmetrizations (grow-diag-final-and, intersection, union, grow-diag-final, grow-diag, grow, srctotgt, tgttosrc, and grow-final) in the Chinese–English language pair. Given an example of a baseline system that used the same symmetrization, i.e., grow-diag-final-and, in two datasets (NIST2002 and NIST2005), the BLEU scores obtained for both datasets were 31.56 and 25.82. Their results confirmed the dataset-specific characteristic, in agreement with Wu Hua (2007).

The symmetrization of word alignment can be applied either in direct translation (Singh, 2015; Wu Hua, 2007) or pivot approaches (Ahmadnia and Serrano, 2017; Ahmadnia et al., 2017, 2018; Budiwati and Aritsugi, 2019; Dabre et al., 2015; Utiyama and Isahara, 2007; Wu and Wang, 2007). No studies have compared the different strategies in direct translation and pivot approaches. Considering this gap, we applied a phrase table combination that uses different symmetrizations of word alignment in pivot approaches based on the highest BLEU scores. Our consideration was based on (Koehn et al., 2005; Singh, 2015; Stymne et al., 2014; Wu Hua, 2007), which stated that the symmetrization of word alignment is language-specific and dataset-specific as different language pairs with different datasets have different BLEU scores. Lastly, our strategy employs the available parallel corpora of src–trg, src–pvt, and pvt–trg without removing the phrase pair that could be a potential candidate in the translation process. We explain further of our strategy in Section 4.2.

2.4.2 Kazakh to English machine translation

Kk–En and Ja–Id are considered low-resource language pairs due to the scarcity of available parallel corpora. The available parallel corpora of Kk–En are open source parallel corpus (OPUS) (Tiedemann, 2012) and news-commentary, which have 953,240 parallel sentences in total. The morphological segmentation approach produced a progressive improvement in the translation quality of Kk–En in the SMT model (Assylbekov and Nurkas, 2014; Kartbayev, 2015b). The morphological segmentation is an approach that breaks words into morphemes. The approach was implemented because Kazakh is considered an agglutinative and highly inflected language. Figure 2.9 shows Kazakh words without morphological segmentation. Word *"дос/friend"* was joined by various suffixes and corresponded to English phrases of various lengths (Assylbekov and Nurkas, 2014). Assylbekov and Nurkas (2014) showed the morphological segmentation approach could obtain an absolute gain of up to 1.05 BLEU points in the 636K dataset. Kartbayev (2015b) also showed an absolute improvement of 1.43 BLEU points when using the morphological segmentation approach in 60K dataset. Both research results (Assylbekov and Nurkas, 2014; Kartbayev, 2015b) showed that the morphological segmentation is an important consideration in the SMT model for Kazakh.



дос	friend
дос**тар**	friend**s**
достар**ым**	**my** friends
достар**ымыз**	**our** friends
достарымыз**да**	**at** our friends
достарымызда**мыз**	**we are** at our friends

Figure 2.9: Example of Kazakh suffixation. The left side is Kazakh words with various suffixes. The right side is English translation with various phrase lengths (Assylbekov and Nurkas, 2014)

Kk–En machine translation was introduced as a shared task for low-resource language pairs in the Workshop on Machine Translation (WMT) 2019 (Barrault et al., 2019). Most participants used the NMT model with several approaches: back translation, transfer learning, multilingual transfer learning, and sequence-2-sequence. The transfer learning approach is a similar technique to the pivot approaches in the SMT model. Transfer learning uses a high-resource language pair to train the parent model, and then the parent training data are replaced with the training data of low-resource language pairs (Kocmi and Bojar, 2019). We identified three submission systems that use transfer learning: the NICT (Dabre et al., 2019), CUNI (Kocmi and Bojar, 2019), and UMD systems (Briakou and Carpuat, 2019). The NICT and CUNI systems use Russian–English as the parent model and obtained BLEU scores of 26.2 and 18.5, respectively. In comparison with both systems, the UMD system uses Turkish–English as the parent model and obtained a BLEU score of 9.2. Their experimental results (Briakou and Carpuat, 2019; Dabre et al., 2019; Kocmi and Bojar, 2019) showed that the third language was still needed to improve the Kk–En translation quality.

2.4.3 Japanese to Indonesian machine translation

The available parallel corpora of Ja–Id are Asian language treebank (ALT) (Riza et al., 2016), TUFS Asian language parallel corpus (TALPCo) (Nomoto et al., 2018), and OPUS (Tiedemann, 2012), which contain 1,468,155 parallel sentences in total. Unlike Kk–En, the SMT model still outperformed NMT with an absolute improvement of 3.93 BLEU points in Ja–Id (Adiputra and Arase, 2017). Moreover, the pivot approaches of the SMT model obtained a higher BLEU score than direct translation in Ja–Id (Budiwati and Aritsugi, 2019; Paul et al., 2009). Paul et al. (2009) showed that the single pivot approach produced an absolute gain of up to 0.23 BLEU points compared to direct translation. Budiwati and Aritsugi (2019) showed that single and multiple pivot approaches obtained absolute improvements of 0.31 and

0.41 BLEU points, respectively, compared to direct translation.

Besides model and techniques, Simon and Purwarianti (2013) and Sulaeman and Purwarianti (2015) found several morphological issues in Ja–Id: word order problems, incorrectly defined phrases, and words with affixes. In particular, Simon and Purwarianti (2013) proposed several techniques: using pos-tag, increasing the LM dataset, stemming for the Indonesian dataset, removing Japanese particles for the Japanese dataset, and removing the named entity (NE). The BLEU score with the Japanese particles removal outperformed the baseline by 0.02. Then, Sulaeman and Purwarianti (2015) proposed several techniques: the pos-tag model, the hierarchical model, lemmatizer, and post-processing. The pos-tag model and lemmatizer outperformed the baseline by 0.1 BLEU points. However, Simon and Purwarianti (2013) and Sulaeman and Purwarianti (2015) did not show their proposed generated text results. Consequently, it is hard to compare the pre- and post-proposed generated texts.

Several experimental results show that the SMT model still a primary option to improve the translation quality for Ja-Id language pair, mainly using the pivot approach. However, it needs an additional technique to overcome the morphological issue that previous works could not achieve.

Table 2.1: BLEU score comparison of related work for Ja-Id.

Experiments	Paul et al., (2009)		Simon et al., (2013)		Sulaeman et al., (2015)		Adiputra et al., (2017)
	Ja-Id	Id-Ja	Ja-Id	Id-Ja	Ja-Id	Id-Ja	Ja-Id
Baseline	52.90	55.52	0.06364	0.10424	0.0065	0.1369	9.34
Proposed	53.13	54.12	0.08806	0.08342	0.172	0.1652	6.45

Table 2.2: Proposed approaches and dataset of the related works for Ja-Id.

Experiments	Paul et al., (2009)	Simon et al., (2013)	Sulaeman et al., (2015)	Adiputra et al., (2017)
Baseline	SMT	SMT	SMT	SMT
Proposed approaches	SMT with single pivot Cascade	SMT with stemmer	SMT with reordering model	NMT with biRNN
Dataset	160K of BTEC	500	1,132 of JLPT	725,495 of OPUS and ALT

Chapter 3

Single and multiple pivots in low-resource languages

This chapter discusses our preliminary work in Kk–En and Ja–Id by using pivot approaches. We applied a single pivot on Kk–En, whereas we applied a single pivot and multiple pivots on Ja–Id. The Kk–En single pivot is explained based on our first publication. The Ja–Id single pivot and multiple pivots explained based on our second publication.

3.1 Single pivot approach in Kazakh to English

In this part, we explain our participation in the WMT19 (Workshop on Machine Translation 2019) shared task as a preliminary study of the Interpolation approach. We choose the *news* translation task and focus on Kk–En (and vice versa) as low-resource language. We built several systems and called our system as DBMS-KU (Database Management System - Kumamoto University) Interpolation as we use our laboratory and university name, as well as utilize the Interpolation method in our experiments.

Kk–En is a new shared task for this year, that is, no experience system description from previous WMT. Kk–En considered a low-resource language pair due to the limitation of parallel corpora and morphological tools. Additionally, another challenge is the difference in the writing system between Kazakh and English languages. Kazakh uses Cyrillic letters, while English uses the alphabet. Different writing system between language pair needs specific attention in the tokenization step because of its segmentation results that affect the BLEU-cased score.

Kk–En machine translation explored in Statistical Machine Translation (SMT) (Assylbekov and Nurkas, 2014; Kartbayev, 2015a,b; Kuandykova et al., 2014) and Neural Machine Translation (NMT) (Myrzakhmetov and Kozhirbayev, 2018). Assylbekov and Nurkas (2014) have shown an interesting result that different n-gram and neural LSTM-based language models were able to reduce the perplexity score, i.e., giving better translation results. For this reason, we consider investigating different n-gram language model order in this work.

Interpolation has been used in language model (Allauzen and Riley, 2011; Heafield et al., 2016; Liu et al., 2013) and translation model (Bisazza et al., 2011; Rosa et al., 2015; Sennrich, 2012). Also, the interpolation has been used in pivot language as a strategy to overcome the limitation of parallel corpora (Dabre et al., 2015; Hoang and Bojar, 2016b; Kunchukuttan et al., 2017). Pivot strategy arises as a preliminary assumption that there are enough parallel corpora between src–pvt and pvt–trg languages. Currently, English as lingua franca has more datasets compared to other languages. Thus, pivot researchers commonly use English as a bridge between source to target (Ahmadnia et al., 2017; Dabre et al., 2015; El Kholy et al., 2013; Paul et al., 2013; Trieu, 2017). However, Paul et al. (2013) and Dabre et al. (2015) have shown that using non-English as pivot language could be a better option to improve the translation results for particular language pair. Since Kk–En is categorized as low resource language pair, we adopt the pivot and interpolation strategies in our translation model.

In this preliminary work, we consider examining two systems, namely, Baseline and Interpolation. The Baseline system is a direct translation between each language pair, while Interpolation one is a combination of pivot and direct translation models. We use Russian as our pivot language with 3-gram and 5-gram language model orders in each system. Our experimental results are encouraging and indicate that using the Interpolation system could obtain a better BLEU-cased score than employing Baseline one when translating both Kazakh to English (Kk–En) and English to Kazakh (En–Kk).

3.1.1 Dataset and experimental setup

In this section, we describe the dataset, and experimental setup of this study.

Dataset and preprocessing

We used a dataset provided by the WMT19 organizer. Thus, our system was considered as a constrained system. To prepare parallel datasets, we cleaned the dataset by using our script because the original dataset had blank lines and unsynchronized sentences between source and target parallel corpora. In the Interpolation system, we used a Russian–English dataset from WMT18. The dataset statistics of training *(train)* and development (*dev*) for Baseline and Interpolation systems are given in Table 3.1.

Dataset	Sentences	Average Sentence Length	Vocab
Baseline system			
Train			
news-commentary-v14.en-kk.kk	9,619	18.0857	29,142
news-commentary-v14.en-kk.en	9,619	22.1487	16,742
Dev			
newsdev2019-enkk.kk	2,068	18.0164	11,389
newsdev2019-enkk.en	2,068	22.2316	7,726
Language Model			
news-commentary-v14.kk	12,707	17.2109	-
news-commentary-v14.en	532,560	21.5762	-
Interpolation system			
Train			
news-commentary-v14.kk-ru.ru	7,230	23.6836	27,819
news-commentary-v14.kk-ru.kk	7,230	20.1187	24,627
news-commentary-v14.en-ru.en	97,652	23.0416	51,566
news-commentary-v14.en-ru.ru	97,652	21.3508	126,476
Dev			
news-commentary-v14.kk-ru.ru	2,000	20.8755	11,841
news-commentary-v14.kk-ru.kk	2,000	18.048	10,561
newstest2018-ruen.dev.en	3,000	20.975	10,108
newstest2018-ruen.dev.ru	3,000	17.3293	17,091
Language Model			
news-commentary-v14.kk	12,707	17.2109	
news-commentary-v14.en-ru.ru	114,375	21.2678	
news-commentary-v14.en-ru.en	114,375	22.9811	

Table 3.1: Dataset statistics for Baseline and Interpolation systems

After cleaning the dataset, we followed dataset preprocessing as in (Myrzakhmetov and Kozhirbayev, 2018), namely, tokenizing, normalizing punctuation, recasing, and filtering the sentences. Tokenizing was used to separate the token and punctuation by inserting spaces. Our tokenization results were based on words. Thus, the obtained sentences of the tokenization results were longer than the original sentences. Since long sentences could cause problems in the training process, we removed the sentences with a length of more than 80 words, the process called filtering the sentences. Normalizing punctuation was to convert the punctuation for being recognized by the decoder system. Recasing was to change the initial words into their most probable casing to reduce the data sparsity. All preprocessing steps were done by using scripts from Moses (Koehn et al., 2007).

Experimental setup

We used an open-source Moses (Koehn et al., 2007) and Giza++ for word alignment, Ken-LM (Heafield, 2011a) for language model, and MERT (Och, 2003) for tuning the weight. The translation results measured by five automatic evaluations provided by the organizer, namely BLEU, BLEU-cased, TER, BEER 2.0, and CharacTER. In this work, we used the BLEU-cased because it is the main comparison metric in the evaluation system[1].

We built two systems, namely, Baseline and Interpolation. The Baseline system is a direct translation between Kk–En and vice versa. Meanwhile, the Interpolation system is the combination of direct translation with a pivot phrase table. Pivot phrase table produced by merging the source to pivot (src–pvt) and pivot to target (pvt–trg) by using the Triangulation method (Hoang and Bojar, 2015). We built

[1] http://matrix.statmt.org/

the Interpolation phrase table as follows:

- Constructing a phrase table from src–pvt and pvt–trg systems and pruning the phrase table with *filter-pt* (Johnson et al., 2007). The pruning activity intended to minimize the noise of src–pvt and pvt–trg phrase tables.
- Merging two pruned phrase tables by using the Triangulation method (Hoang and Bojar, 2015). The result called `TmTriangulate` phrase table.
- Combining `TmTriangulate` and direct translation model with *dev* phrase table as references. We used linear interpolation with backoff mode and exploited *combine-ptables* tools (Bisazza et al., 2011). The result was called `Interpolation` phrase table.

3.1.2 Results and discussion

In this section, we show the obtained automatic evaluation results using a BLEU-cased score. We also discuss the effect of the different language model order with the BLEU-cased score. Furthermore, we analyze the perplexity score on the Interpolation system.

Language model effects on BLEU-cased score

In this work, we conducted experiments for two language model orders, i.e., 3-gram and 5-gram, and two systems, viz., Baseline, and Interpolation. As shown in Table 3.2, the 5-gram language model order had more significant influence than the 3-gram one on the BLEU-cased score for Kk–En translation in both Baseline and Interpolation systems. The improvement in Kk–En obtained by +0.3 and +0.7 points for Baseline and Interpolation systems, respectively. However, the BLEU-cased score for En–Kk could not be improved in terms of the language model order. These results might indicate that the language model order influenced the BLEU-cased score.

In terms of the translation system, the Interpolation system obtained a higher BLEU-cased score than the Baseline one for all language models and translation directions. The improvement of the BLEU-cased score from Baseline to Interpolation system for Kk–En using 3-gram and 5-gram was +0.1 and +0.5 points, respectively. Meanwhile, the improvement from Baseline to Interpolation System for En–Kk was +0.1 for both 3-gram and 5-gram orders. These results indicated that the use of pivot language in the Interpolation system combined with a longer language model also had a significant influence on the BLEU-cased score.

We found that the Kk–En obtained a higher BLEU-cased score than the En–Kk in terms of the translation direction. This result might be influenced by the number of target LM dataset En–Kk had 12,707 sentences. The translation direction of Kk–En, that is, having almost 42 times larger sentences than En–Kk, could obtain a higher BLEU-cased score than En–Kk. This result indicated that the number of target LM dataset in the experiments might improve the BLEU-cased score.

Although our obtained BLEU-cased score was relatively low, we showed that by combining Baseline and pivot parallel corpora with different LM order was a valuable effort compared with using direct parallel corpora only. Moreover, the BLEU-cased score improvement could be influenced by the language model order, the translation system, and the target monolingual LM dataset.

Table 3.2: BLEU-cased score results

Language Pair	3-gram LM	5-gram LM
KK-EN		
1. Baseline system	2.6	2.9
2. Interpolation system	2.7	3.4
EN-KK		
1. Baseline system	0.8	0.8
2. Interpolation system	0.9	0.9

Table 3.3: Perplexity results

Language pair	3-gram LM	5-gram LM
KK-EN		
1. Baseline system	- Incl OOVs: 829.59 - Excl OOVs: 77.79	- Incl OOVs: 617.36 - Excl OOVs: 45.51
2. Interpolation system	- Incl OOVs: 1034.50 - Excl OOVs: 94.72	- Incl OOVs: 762.79 - Excl OOVs: 50.93
EN-KK		
1. Baseline system	- Incl OOVs: 328.940 - Excl OOVs: 103.27	- Incl OOVs: 256.138 - Excl OOVs: 77.185
2. Interpolation system	- Incl OOVs: 256.13 - Excl OOVs: 79.34	- Incl OOVs: 276.85 - Excl OOVs: 85.40

Perplexity effects on the Interpolation system

Language model is one of the SMT components to ensure how good is the model by using perplexity as measurement. Lower perplexity score indicates better language models, while high perplexity score represents that the language model has poor quality. We show the perplexity score of the target language test dataset according to each n-gram language model trained on the respective training dataset in Table 3.3.

As shown in Table 3.3, the lowest perplexity score for Kk–En was obtained by the 5-gram Baseline system, i.e., 45.51. Thus, the best model for Kk–En was the 5-gram Baseline system. However, we found that the difference of perplexity score for the 5-gram model between Baseline and Interpolation systems was not quite significant, i.e., 5.42. Specifically, the perplexity of the 5-gram of Baseline was 45.51, while the perplexity of the 5-gram of Interpolation was 50.93. This finding might indicate that pivot language with the interpolation system could be a beneficial approach in the translation process.

In En–Kk, the lowest perplexity score was obtained by the 5-gram Baseline system, i.e., 77.18. Thus, the best model for En–Kk was the 5-gram Baseline system. However, we found that the difference of perplexity score between 5-gram Baseline and 3-gram Interpolation systems was not quite significant, i.e., 2.16. Specifically, the perplexity of the 5-gram of Baseline was 77.18, while the perplexity of 3-gram of Interpolation was 79.34. This finding might indicate that using the interpolation system with a 3-gram model could only reduce the perplexity score of En–Kk, than using the longer n-gram language model, i.e., 5-gram. Nevertheless, it would be better to study further the cause of this finding in the future.

3.2 Single and multiple pivots approach in Japanese to Indonesian

The pivot approach often uses English as pivot languages. However, Wu and Wang (2007) and Paul et al. (2013) showed that non-English as a pivot language can improve translation quality for specific language pairs. Wu and Wang (2007) showed that using Greek as a pivot language has improved the translation quality compared to English in French to Spanish language pair. Greek as pivot language obtained +5.00 points, meanwhile English obtained +2.00 points. Paul et al. (2013) showed that from 420 experiments language pair in Indo-European and Asian languages, 54.8% is preferable using non-English as the pivot language. Moreover, Wu and Wang (2007) and Dabre et al. (2015) showed promising results by using more than one non-English language. Wu and Wang (2007) showed that using four languages, namely Greek, Portuguese, English, and Finnish outperformed the baseline BLEU score with more than +5.00 points. Dabre et al. (2015) also showed that using seven non-English, namely Chinese, Korean, Marathi, Kannada, Telugu, Paite, and Esperanto pivot languages outperformed the baseline BLEU score with more than 3.00 points in Japanese to Hindi language pair.

In this preliminary study, we investigate single, and multiple pivots approach on Ja–Id, and vice versa, by using non-English as a pivot language. First, we construct a direct translation system as a Baseline system. Then, we build a single pivot system by exploiting the Cascade, Triangulation, Linear Interpolation (LI), and Fillup Interpolation (FI) techniques. Last, we construct multiple pivots system by combining four phrase tables using Linear Interpolation (LI) and Fillup Interpolation (FI) techniques.

In a single pivot system, we use four pivot languages, namely English, Myanmar, Malay, and Filipino. We measured the effect of single pivot by one parameter, i.e., dataset type. The dataset type divided into two categories, viz., sequential, and random. The sequential type means that the dataset remains unchanged. Meanwhile, random type means the dataset shuffled before processed into the SMT framework. In multiple pivots system, we examine the use of *with and without* source to target (src–trg) phrase table when we were exploiting the LI and FI techniques. We measured the effect of multiple pivots by two parameters, viz., dataset type, and phrase tables order. The dataset type is the same as well as in the single pivot. Whereas the phrase table order comprises of two, viz., descending and ascending. The descending order arranges the four phrase tables from highest to lowest according to their BLEU scores. Ascending order is the opposite.

3.2.1 Dataset and experimental setup

In this section, we first describe languages that involved in this work. Then, we explain how dataset is divided. Last, we describe the experimental setup.

Dataset

We use six datasets from ALT, i.e., Japanese, Indonesian, English, Myanmar, Malay, and Filipino. Japanese and Indonesian datasets were used to build the direct translation as the Baseline system. The Japanese language is an SOV language, while Indonesian is an SVO language. Therefore, we chose pivot languages based on the similarity of word order with Japanese or Indonesian. English and Malay have the same word order as Indonesia. Myanmar has the same word order as Japanese. Filipino was chosen to evaluate the effect of VOS language. Table 3.4 is shown the word order and language family.

Table 3.4: Language characteristics.

Languages	Word order	Language family
Japanese	SOV	Japonic
Indonesian	SVO	Austronesian
English	SVO	Indo-European
Myanmar	SOV	Sino-Tibetan
Malaysian	SVO	Austronesian
Filipino	VOS	Austronesian

We divide the dataset into two data types, viz., sequential (seq) and random (rnd). The sequential type means that the dataset remains unchanged. Meanwhile, random type means the dataset was shuffled before used in the SMT framework. We used `random.shuffle()` method from python library. We divide datasets into 8.5K for training (*train*), 2K for tuning (*dev*) and 1K for the evaluation *(eval)*. Overall, we conduct 132 experiments, i.e., four Baselines, 32 src-pvt and pvt-trg, 64 single pivots, and 32 multiple pivots.

Experimental setup for single pivot

We used Moses (Koehn et al., 2007) and Giza++ for word alignment process, phrase table extraction and decoding. We used 3-gram KenLM (Heafield, 2011b) for language model, MERT (Och, 2003) for tuning and BLEU (Papineni et al., 2002) for evaluation from Moses package.

In the single pivot, we implement four approaches, i.e., Cascade, Triangulation, Linear Interpolation (LI), and Fillup Interpolation (FI). In the Cascade approach, we construct two systems, viz., src–pvt, and pvt–trg systems. The src–pvt system translates the source language input into the pivot language. The pvt–trg system takes the translation result of src–pvt as input and translates into the target language. As a result, we construct 16 src–pvt and 16 pvt–trg systems.

In the Triangulation approach, we construct phrase tables as follows:

- Pruning the src-pvt and pvt-trg phrase table from the Cascade experiments using *filter-pt* (Johnson et al., 2007). The pruning activity intended to minimize the noise of the src–pvt and pvt–trg phrase table.

- Merging two pruning phrase tables using *TmTriangulate* (Hoang and Bojar, 2015). The result is denoted as `TmTriangulate` phrase table.

In the Linear Interpolation (LI) approach, we combine `TmTriangulate` and the src-trg phrase table with *dev* phrase table as a reference. The result is called `TmCombine` phrase table. In Fillup Interpolation (FI), we use *backoff* mode thus the result is called `TmCombine-Backoff` phrase table. We use *tmcombine* and *combine-ptables* tools to construct `TmCombine` and `TmCombine-Backoff` phrase tables.

Experimental setup for multiple pivots

In multiple pivots, we employ phrase tables from the best pivot approaches from a single pivot system. We identified that the LI and FI system obtained a high BLEU score among the Baseline, Cascade, and Triangulation system, as shown in Table 3.5 - Table 3.8. Therefore, we use the four phrase tables from LI and FI approaches, i.e., English phrase table (EnPT), Myanmar phrase table (MyPT), Malay phrase table (MsPT), and Filipino phrase table (FiPT).

Next, we combine the four phrase tables based on BLEU score order, viz., descending, and ascending orders. The descending order sorts the four phrase tables from highest to lowest according to their BLEU scores. The ascending order is the opposite. For example, the BLEU scores of the LI approach are 11.34 for EnPT, 12.21 for MyPT, 12.11 for MsPT, and 12.15 for FiPT. For descending order, we put the four phrase tables, i.e., MyPT, FiPT, MsPT, and EnPT, respectively. Meanwhile, for ascending order, we put the four phrase tables, i.e., EnPT, MsPT, FiPT, MyPT, respectively.

The combinations of multiple pivots phrase tables were examined *with and without* an src-trg phrase table, as follows:

- Merging of four phrase tables without an src-trg phrase table using the Linear Interpolation (LI) approach. The result denoted as All-LinearInterpolate `All-LI`.

- Merging of four phrase tables without an src-trg phrase table using the Fillup Interpolation (FI) approach. The result denoted as All-FillupInterpolation `All-FI`.

- Combining `All-LI` with an src-trg phrase table using the Linear Interpolation (LI) approach. The result denoted as `Base-LI`.

- Combining `All-FI` with an src-trg phrase table using the Fillup Interpolation (FI) approach. The result denoted as `Base-FI`.

3.2.2 Results and discussion

In this section, we will discuss results based on BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and perplexity scores. BLEU score is a metric for evaluating the generated sentence compared to the reference sentence. High BLEU scores indicate a better system. Perplexity score is frequently used as a quality measure for language models (Sennrich, 2012). Lower perplexity scores indicate that the language model is better compared to higher perplexity score. We used the query from KenLM (Heafield, 2011b) to get the perplexity including OOV (Out of Vocabulary). OOV is unknown words that do not appear in the training corpus. We show the perplexity scores of the target language test dataset according to the 3-gram language model trained on the respective training dataset.

Baseline translation results

The Baseline is a direct translation between languages pair, namely Ja–Id and Id–Ja. We construct two Baseline systems in each language pair, based on data types, i.e., sequential and random.

The Baseline BLEU scores of Ja–Id are given in Table 3.5 and Table 3.6. As shown in the tables, Baseline Random obtained higher BLEU score compared to Baseline Sequential. The BLEU score of Baseline Random Ja–Id is 12.17, +0.21 points higher compared to Baseline Sequential. Table 3.7 and Table 3.8 shown the BLEU scores of Id–Ja. The Baseline Random Id–Ja also obtained higher as much as 1.00, compare to Baseline Sequential.

Baseline perplexity scores are given in Table 3.5 - Table 3.8. As shown in the Tables, the Ja–Id and Id–Ja perplexity scores of Baseline Random obtained higher compared to the Baseline Sequential one. Take an example of perplexity score of Ja–Id in Baseline Random obtained 384.59, while Baseline Sequential obtained 291.51. Furthermore, the perplexity score of Id–Ja in Baseline Random also obtained higher as much as 81.58, while Baseline Sequential obtained 71.94.

The results denote that although the Random data type obtained higher BLEU score but it still has the OOV issue, compared to Sequential data type. In the next section, we showed our efforts to reduce the perplexity scores by using multiple pivots.

Table 3.5: Ja-Id BLEU score on sequential data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		11.96		
Single pivot system				
JaId (English)	10.89	9.71	11.97	12.07
JaId (Myanmar)	9.37	8.71	11.91	12.27
JaId (Malay)	12.01	8.37	11.71	12.09
JaId (Filipino)	9.95	9.41	12.23	12.19

Table 3.6: Ja-Id BLEU score on random data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		12.17		
Single pivot system				
JaId (English)	10.81	9.10	12.18	12.22
JaId (Myanmar)	9.60	8.60	11.91	12.29
JaId (Malay)	11.81	9.25	12.22	12.05
JaId (Filipino)	9.68	9.62	12.09	11.99

Table 3.7: Id-Ja BLEU score on sequence data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		11.00		
Single pivot system				
JaId (English)	12.07	8.26	12.65	12.05
JaId (Myanmar)	9.97	6.76	10.89	12.4
JaId (Malay)	12.18	6.76	12.2	11.87
JaId (Filipino)	10.36	7.28	12.06	12.2

Table 3.8: Id-Ja BLEU score on random data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		12.00		
Single pivot system				
JaId (English)	7.58	7.96	12.10	11.99
JaId (Myanmar)	10.32	6.51	12.84	12.88
JaId (Malay)	11.13	9.17	12.52	11.82
JaId (Filipino)	10.46	7.97	12.25	12.68

Single pivot results

Table 3.5 - Table 3.8 show the result of single pivot BLEU scores. From these tables, the Triangulation approach was the worst approach in Ja–Id and Id–Ja. All the results of Triangulation have smaller BLEU score compared to the Baseline. The Cascade approach also has lower scores compared to the Baseline, except three experiments in Sequential data type by using Malay and English as a pivot language. The three experiments outperformed the Baseline by range from 0.05 to 1.18 points. However, we didn't use the Cascade results because of its different technique compared to other approaches. The Cascade approach did not combine phrase tables such as LI and FI. The Cascade approach used two independently systems, i.e., src–pvt and pvt–trg. The src-p-vt system translates the Japanese text into the pivot language. The pvt–trg system takes the translation result as input and translates into Indonesian text.

The Linear Interpolation (LI) and Fillup Interpolation (FI) approaches show significant result in Ja–Id and Id–Ja. Both approaches have higher BLEU scores compared to Baseline, by more than 75% experiments. This was shown in Table 3.5 - Table 3.8.

In terms of language, Myanmar became a main option as pivot language in Ja–Id Sequential data type. Meanwhile, Ja–Id Random data type has two options of pivot language, i.e., Malay, and Myanmar. Surprisingly, Myanmar also became a main option as pivot language in Id–Ja Sequential and Random data types. As we look to the language characteristics in Table 3.4, Myanmar has the same word order as Japanese while Malay has the same word order as Indonesian. The results denote that word order closely related to the source or target language should be considered when choosing pivot language.

In terms of data type, Sequential or Random data types could be chosen in Ja–Id. Both data types have increased the BLEU scores by 75% of experiments. Random data type was preferable in Id–Ja because the highest improvement points were achieved by +1.84 compared to Baseline. The results denote that data type is an important parameter to consider to improve the BLEU score.

In terms of perplexity score, the LI and FI approaches in different data types are unable to reduce the scores. The single pivot language even increased the perplexity scores as shown in Figure 3.1 and Figure 3.2. We showed how to reduce the perplexity scores by using multiple pivots in the next section.

Multiple pivots results

From the single pivot, LI and FI become the best approach to improve the BLEU scores compared to the Baseline. Therefore, we use the phrase tables from both approaches and we did combinations of multiple pivots phrase tables, i.e., All-LI, All-FI, Base-LI, and Base-FI, as described in Section 3.2.1.

For example in Ja–Id of All-LI, we combine the four phrase tables from the single pivot LI approach by descending and ascending orders. First, we observe the BLEU scores of LI Sequential data type are 11.34 for EnPT, 12.21 for MyPT, 12.11 for MsPT, and 12.15 for FiPT. Next, we combine the four phrase tables according to their BLEU scores in descending order, i.e., MyPT, FiPT, MsPT, and EnPT, respectively. Last, we combine the four phrase tables according to their BLEU scores in ascending order, i.e., EnPT, MsPT, FiPT, MyPT, respectively. As a result, the BLEU scores have different scores for descending and ascending orders, i.e., 12.01 and 12.20, respectively. The results are shown in Figure 3.5.

We did not use src–trg phrase table in All-LI and All-FI approaches, and their BLEU scores outperformed Baseline. The results denote that the translation could be accomplished with multiple pivots and still produce high BLEU scores without using src-trg phrase table. Moreover, the translation results could have higher BLEU scores if there is a small src–trg phrase table, as shown in Base-LI and Base-FI results.

The combinations of multiple pivots phrase tables have different effects on the BLEU scores, when we used different order. In Ja-Id, the descending order was preferable because more than 87.5% experiments result outperformed the Baseline. In Id-Ja, the ascending order was preferable because all the experiments outperformed the Baseline. The results are shown in Figure 3.5 and Figure 3.6 for Ja–Id

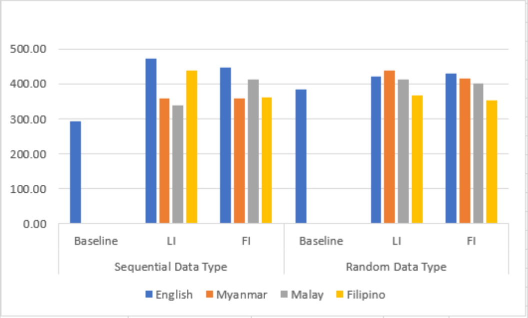


Figure 3.1: Perplexity Score of Ja-Id single pivot for LI and FI approaches.

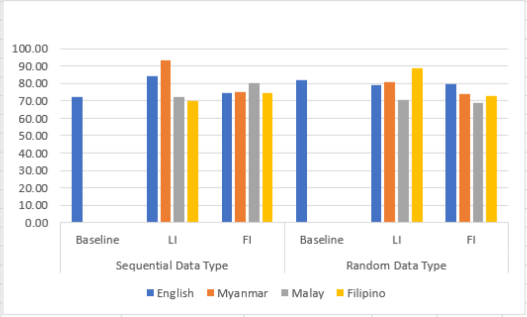


Figure 3.2: Perplexity Score of Id-Ja single pivot for LI and FI approaches.

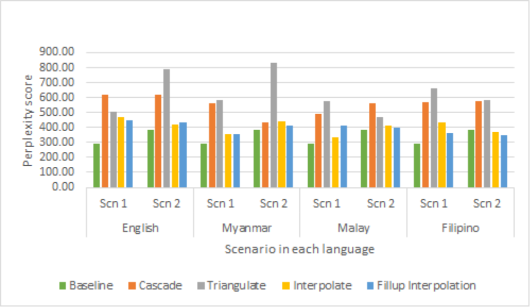


Figure 3.3: Perplexity score for Ja-Id in single pivot.

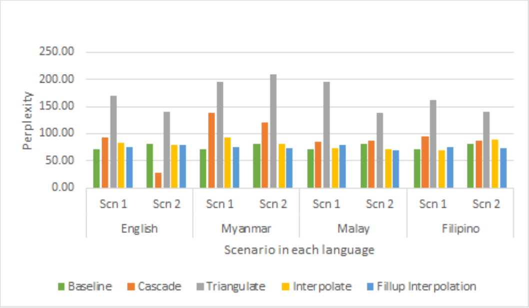


Figure 3.4: Perplexity score for Id-Ja in single pivot.

and Id–Ja, respectively.

In terms of data type, most of the results of Ja–Id outperformed the Baseline, excluding the Base-FI Random data type. Meanwhile, all the results of Id-Ja outperformed the Baseline. The highest improvement score was obtained by Base-LI Random data type in Ja–Id descending, by +0.23 points. Meanwhile, the highest improvement was obtained by ALL-FI Sequence data type in Id–Ja ascending, as much as +1.84 points. The results indicate that data types have a significant effect to improve the BLEU scores.

In terms of perplexity scores for Ja–Id, All-LI and All-FI show poor results. However, the perplexity scores could be reduced in Random data type of Base-LI and Base-FI. Both approaches use src–trg phrase table in the combination process. The results show that the src–trg phrase table has a significant impact on reducing the perplexity score. Meanwhile, the perplexity scores in Id–Ja could be reduced without using the src–trg phrase table. Moreover, the Base-LI and Base-FI results have lower perplexity scores compared to All-LI and All-FI. We show the perplexity scores in Figure 3.7 and Figure 3.8 for Ja–Id and Id–Ja, respectively. We summarize the results of single and multiple pivots in Table 3.9 and Table 3.10.

Table 3.9: Best BLEU score in Baseline, single and multiple pivots for Ja-Id

Scenario No	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulate	Interpolate	Fillup Interpolation	Desc	Asc
Scenario 1	11.96	12.01 (MS)	9.71 (EN)	12.21 (MY)	12.27 (MY)	12.23 (Cat 3)	12.37 (Cat 4)
Scenario 2	12.17	11.81 (MS)	9.62 (FI)	12.22 (MS)	12.29 (MY)	12.40 (Cat 3)	12.27 (Cat 2)

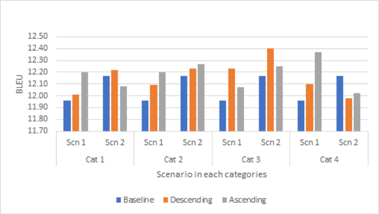


Figure 3.5: BLEU score for Ja-Id in multiple pivots.

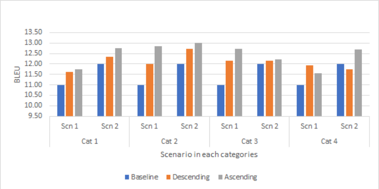


Figure 3.6: BLEU score for Id-Ja in multiple pivots.

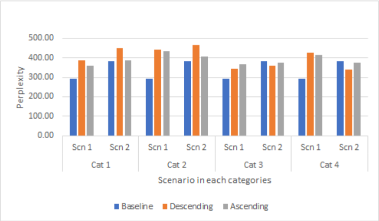


Figure 3.7: Perplexity score for Ja-Id in multiple pivots.

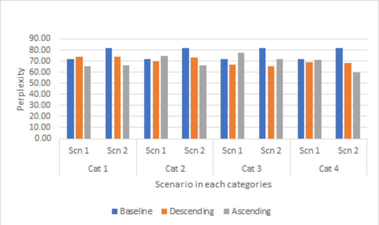


Figure 3.8: Perplexity score for Id-Ja in multiple pivots.

3.3 Summary

In this chapter, we applied single pivot in Kk–En (vice versa). We also applied single and multiple pivots in Ja–Id (vice versa). In the first section, we examined the effect of different LM orders with a linear interpolation method for participating in WMT19 shared task. Our Interpolation system utilized the combination of direct translation, i.e., Baseline, with Russian as our pivot language. We used 3-gram and 5-gram language model orders in our Baseline and Interpolation systems. The BLEU-cased score of using the Interpolation system could outperform that of utilizing Baseline one. This good performance of the Interpolation system was obtained by using 3-gram and 5-gram language model orders for both Kazakh to English (Kk–En) and English to Kazakh (En–Kk) translations. We found that the Interpolation system indicated a different effect on each of Kk–En and En–Kk in terms of the perplexity score. In Kk–En, the pivot language with the interpolation system could be an option in the translation process because the difference of perplexity score between Baseline and Interpolation was not quite significant. Interestingly, we found that the Interpolation system using a 3-gram language model order could reduce the perplexity score compared with utilizing longer n-gram one in En–Kk.

In the second section, we showed experiment results of single and multiple pivots in Ja-Id and Id-Ja. We used English, Myanmar, Malay, and Filipino as pivot languages in single pivot. We implemented four approaches, i.e., Cascade, Triangulation, Linear Interpolation (LI) and Fillup Interpolation (FI) in single pivot. We found that LI and FI approaches outperformed the Baseline. In multiple pivots, we implemented four approaches, i.e., All-LI, All-FI, Base-LI, and Base-FI. We found that most of all approaches in multiple pivots outperformed the Baseline.

We divided the dataset into two data types in single and multiple pivots, namely

Table 3.10: Best BLEU score in baseline, single and multiple pivots for Indonesia to Japanese

Scenario No	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulate	Interpolate	Fillup Interpolation	Desc	Asc
Scenario 1	11.00	12.18 (MS)	8.26 (EN)	12.03 (MY)	12.40 (MY)	12.15 (Cat 3)	12.84 (Cat 2)
Scenario 2	12.00	11.13 (MS)	9.17 (MS)	12.84 (MY)	12.88 (MY)	12.74 (Cat 2)	13.02 (Cat 2)

sequential and random. The data types showed different effects on the language pairs. In Ja–Id of single pivot, sequential or random could be chosen to improve the BLEU score. Both data types have increased the BLEU scores by 75% of experiments. However, random data type was preferable in Id–Ja because the highest improvement points were achieved by +1.84. Random data type was preferable for Ja–Id and Id–Ja in multiple pivots. The highest improvement points were achieved by +0.23 and 1.84 for Ja–Id and Id–Ja, respectively.

In multiple pivots, we combined the four phrase tables from the best single pivot approaches, i.e., Linear Interpolation (LI) and Fillup Interpolation (FI). The combinations of multiple pivots phrase tables were examined with and without src–trg phrase table. We measured the effect by phrase tables orders, i.e., descending and ascending. From the experiment results, the descending order was preferable in Ja–Id. Meanwhile, the ascending order was preferable in Id–Ja.

Chapter 4

Word reordering and phrase table combination

In this part, we expand our experiments from the previous one. This chapter divided into two discussions, viz., word reordering on multiple pivots, and phrase table combination. The word reordering on multiple pivots is our continuation experiment from the single and multiple pivots of Ja-Id. In this experiment, we applied pre-ordering of Japanese sentence and phrase table interpolation by gradually, viz., one pivot, two pivots, three pivots, and four pivots phrase table. Whereas phrase table combination is the experiment that compares standard and different symmetrization of word alignment. We describe our technique for two language pairs, viz., Kk-En and Ja-Id.

4.1 Word reordering on multiple pivots

The SMT is known as less-effective for language pairs with different word orders (Bisazza and Federico, 2016; Isozaki et al., 2012). This problem also arises with the Ja–Id language pair. Japanese uses subject-object-verb (SOV) word order, whereas Indonesian uses subject-verb-object (SVO). Simon and Purwarianti (2013) employed a pos-tagger approach, whereas Sulaeman and Purwarianti (2015) used hierarchical reordering to overcome the problem presented by the different word order in Ja–Id. Simon and Purwarianti (2013) stated that the pos-tagger approach was unable to effectively overcome the word order problem in their experiment. Sulaeman and Purwarianti (2015) improved the translation output by using hierarchical reordering, but the proposed solution did not report clearly whether their strategy is considered pre-ordering, post-ordering, or word ordering in the decoding process.

In this part, we investigated the reordering of Japanese words into the Indonesian word order. Changing the word ordering of Japanese (SOV) into the SVO word order proved to be an effective way to improve the translation quality (Hoshino et al., 2013; Isozaki et al., 2010; Neubig et al., 2012). However, this approach has never been employed for Ja–Id as a low-resource language pair. We implemented pre-ordering, that is, a stand-alone task to rearrange words in the target-like order before the translation process (Bisazza and Federico, 2016) by using Lader (Latent Derivation Reorderer) (Neubig et al., 2012). Lader is a reordering model based on context-free grammar with latent derivations using online discriminative learning.

We used two pivot approaches: single and multiple pivots. We chose pivot approaches because they have been proven to overcome the limitation of parallel corpora between the source-target (src–trg). Initially, researchers used English as a pivot language (Ahmadnia et al., 2017; Utiyama and Isahara, 2007); however, Paul et al. (2013) had an interesting result in that they showed that non-English is preferable to English as a pivot language for particular language pairs. For example, Paul et al. (2013) suggested that Malaysian is preferable as a pivot language for Ja-Id. Additionally, researchers suggested another strategy to improve the translation quality by using multiple pivots, that is, combining more than one pivot language (Budiwati and Aritsugi, 2019; Dabre et al., 2015; Wu and Wang, 2007). We used four languages from the ALT dataset, namely, English, Malaysian, Filipino, and the Myanmar language because of their characteristic languages. Malaysian and Filipino are considered to be closely related to Indonesian because they use the same word order and belong to the same language family. In comparison, the Myanmar language and English use the same word order as Japanese and Indonesian, respectively.

We conducted two experiments: without reordering (WoR) and with reordering (WR). WoR is our experiment that uses single and multiple pivot approaches without reordering the source language. In comparison, WR is our experiment that uses single and multiple pivot approaches to reorder the source language. When using multiple pivots for WoR, we directly combined four phrase tables. In contrast, in the WR experiment, we gradually combined the phrase tables in the multiple pivots, viz., two, three, and four pivot phrase tables. To the best of our knowledge, this study is the first work in which a pre-ordering technique is imposed upon Ja-Id language pairs while using multiple pivots.

In multiple pivots experiment, we propose a strategy: extending phrase tables. We merge two phrase tables of src–pvt, viz., src–pvt *gdfand* and src–pvt *tgttosrc*. Last, we conducted NMT experiments to compare with SMT. We used OpenNMT-py framework (Klein et al., 2017) with two pre-trained models, viz., encoder-decoder and transformer. The encoder-decoder model is a default pre-trained model of OpenNMT-py that uses multiple recurrent neural network (RNN) cells and attention types (Bahdanau et al., 2015; Luong et al., 2015). The transformer model is a pre-trained model of OpenNMT-py based on the Google transformer model (Vaswani et al., 2017).

4.1.1 Extending phrase table

The SMT model pipeline consists of training, tuning, and evaluation. The phrase table, or TM, contains the results of word alignment in the training process. Word alignment aligns the source word with the target word, which was stored in the lexical file and phrase table. By default, word alignments in Moses were obtained by using GIZA++ or MGIZA, which implement IBM models. The IBM models comprise five models, namely lexical translation, adds an absolute reordering model, adds a fertility model, a relative reordering model, and fixes deficiency. These five models estimate the alignments from corpora by using the expectation-maximization algorithms, and each model adds some complexity (Stymne et al., 2014). The models produce directional alignments where one word in the source is linked to many target words (1-m links), but not vice versa. Therefore, to link in both directions, symmetrization strategies were added, such as:

- Intersection preserves the word alignment points that occur in both alignments ($A_{\mathrm{TS}} \cap A_{\mathrm{ST}}$).
- Union joins the word alignment points from both alignments ($A_{\mathrm{TS}} \cup A_{\mathrm{ST}}$).
- Grow adds word alignment points from the neighborhood on the left, on the right, above or below.
- Grow diagonally (grow-diag) adds word alignment points from a diagonal neighborhood.
- Source to target (srctotgt) only considers word-to-word alignments from the src-trg GIZA++ alignment file.
- Target to source (tgttosrc) only considers word-to-word alignments from the trg-src GIZA++ alignment file.

The SMT model uses the default symmetrization, that is, *grow-diag-final-and*, to generate phrase pairs. Researchers found that some language pairs obtained a high BLEU score while using other types of symmetrization, that is, *intersection, tgttosrc* (Koehn et al., 2005; Singh, 2015; Stymne et al., 2014). In addition to a higher BLEU score, we found that other types of symmetrization, that is, *tgttosrc*, could generate phrases that were not available in the phrase table of *grow-diag-final-and.* Phrases that were not available in the phrase table could produce unknown words (UNK), whereas the translation system generates the translation output. As a result, UNK could reduce the BLEU score during the decoding process. We also found that the *tgttosrc* phrase table obtained a lower UNK value compared to *grow-diag-final-and*. Therefore, we decided to utilize our simple strategy in this study, that is, to extend the phrase table.

We employed an extending phrase table to minimize the Japanese UNK in the src-pvt system. The extending phrase table used two types of symmetrization: *grow-diag-final-and* and *tgttosrc*. We assumed that two phrase tables of Ja-En were available by using two types of symmetrization, namely, the Ja-En phrase table of *grow-diag-final-and* as T_{gdfand} and Ja-En phrase table of *tgttosrc* as $T_{tgttosrc}$. Each phrase table has four translation parameters, namely, the phrase translation probabilities for both directions, that is, $\phi(\bar{t}|\bar{s})$ and $\phi(\bar{s}|\bar{t})$, and lexical translation probabilities for both directions, that is, $p_{\mathrm{w}}(\bar{t}|\bar{s})$, and $p_{\mathrm{w}}(\bar{s}|\bar{t})$. First, we sorted the Japanese UNK of the src-pvt system as *condition (C)*. Then, we searched candidate pairs of C in the phrase table of *tgttosrc* as $T_{filtered}$. Finally, we merged the two phrase tables of T_{gdfand} and $T_{filtered}$. Our strategy is represented by the following equations:

$$T_{\text{filtered}} = \sigma_{\text{C}}(T_{\text{tgttosrc}}) \tag{4.1}$$

$$T_{\text{extend}} = T_{\text{filtered}} \cup T_{\text{gdfand}} \tag{4.2}$$

The implementation of our strategy can be accessed in online repositories[1].

4.1.2 Dataset and experimental setup

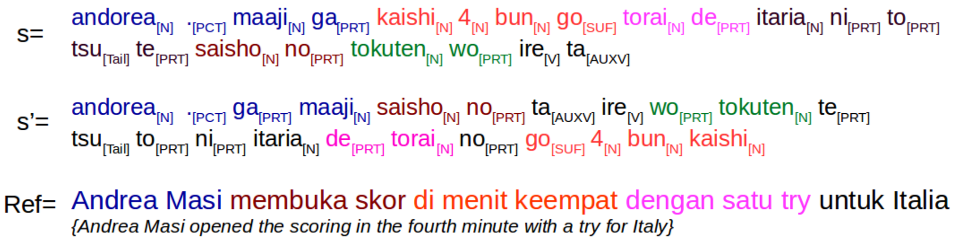
We used corpora for six different languages which are provided in the Asian language treebank (ALT), namely, Japanese (Ja), Indonesian (Id), English (En), Malaysian (Ms), Filipino (Fi), and the Myanmar language (My). We used Japanese and Indonesian (Ja-Id) as language pairs, and used the other languages as pivot languages. The ALT datasets consist of 20,106 sentences annotated with word segmentation, part-of-speech tags, and grammatical information, and are limited to a few datasets, viz., Ja, En, and My. We divide the ALT datasets into 8.5K sentences for training, 2K for tuning, and 1K for evaluation. The datasets we used as src-pvt language pairs were Ja-En, Ja-Ms, Ja-Fi, and Ja-My. At the same time, the pvt-trg language pairs consisted of En-Id, Ms-Id, Fi-Id, and My-Id.

We performed several pre-processing steps on all datasets, including tokenization, lowercasing, and limiting the sentence length. Tokenization is the activity of separating the words from the punctuation. We employed two tokenizers for the

[1] https://github.com/s4d3/Pivot-for-Low-Resource-Languges

Japanese dataset, namely, Mecab IPA-DIC and Kytea, in the WoR and WR experiments, respectively. We used Moses to tokenize Id, En, Ms, and Fi. At the same time, the Myanmar tokenizer was used to tokenize the Myanmar dataset. Lowercase is the activity of converting the initial word in each sentence to its most probable casing. Finally, we deleted sentences that were longer than 80 words.

We added a reordering activity in the pre-processing step of the WR experiment. Reordering is an activity that changes the Japanese word order, that is, SOV, to the Indonesian word order, that is, SVO. We accomplished this by employing Lader. Initially, Lader was used to reorder Japanese into the English word order. Because the Indonesian word order is the same as that of English, we chose Lader to reorder Japanese into the Indonesian word order. Figure 4.1 shows the result of reordering Japanese into the Indonesian word order. Source *(s)* is the original Japanese word order, and *s'* is the reordering result. Reference *(Ref)* is an Indonesian translation reference. Figure 4.1 shows several words that have been moved from the original position of *s* to the left side in *s'*, viz., 最初 (saisho $_{[N]}$), 得点 (tokuten $_{[N]}$), and 入れ (ire $_{[V]}$). We used Kytea to identify Japanese part-of-speech [2].



s= andorea$_{[N]}$ ・$_{[PCT]}$ maaji$_{[N]}$ ga$_{[PRT]}$ kaishi$_{[N]}$ 4$_{[N]}$ bun$_{[N]}$ go$_{[SUF]}$ torai$_{[N]}$ de$_{[PRT]}$ itaria$_{[N]}$ ni$_{[PRT]}$ to$_{[PRT]}$ tsu$_{[Tail]}$ te$_{[PRT]}$ saisho$_{[N]}$ no$_{[PRT]}$ tokuten$_{[N]}$ wo$_{[PRT]}$ ire$_{[V]}$ ta$_{[AUXV]}$

s'= andorea$_{[N]}$ ・$_{[PCT]}$ ga$_{[PRT]}$ maaji$_{[N]}$ saisho$_{[N]}$ no$_{[PRT]}$ ta$_{[AUXV]}$ ire$_{[V]}$ wo$_{[PRT]}$ tokuten$_{[N]}$ te$_{[PRT]}$ tsu$_{[Tail]}$ to$_{[PRT]}$ ni$_{[PRT]}$ itaria$_{[N]}$ de$_{[PRT]}$ torai$_{[N]}$ no$_{[PRT]}$ go$_{[SUF]}$ 4$_{[N]}$ bun$_{[N]}$ kaishi$_{[N]}$

Ref= Andrea Masi membuka skor di menit keempat dengan satu try untuk Italia
{Andrea Masi opened the scoring in the fourth minute with a try for Italy}

Figure 4.1: Example of reordering of Japanese sentence into Indonesian word order. The same colors indicate the same word positions.

SMT reordering experiments

We divided the experiment into two parts, namely, without reordering (WoR) and with reordering (WR). WoR is our experiment that uses single and multiple pivot strategies without reordering the source language. In comparison, WR is our experiment that uses single and multiple pivot strategies to reorder the source language. We measured the translation quality of the two experiments based on the BLEU score and translation output. It was assumed that the MT system obtained a higher BLEU score, and the translation output was understandable. However, Callison-Burch et al. (2006) argued that, because of grammatical or syntactical variations, a higher BLEU score does not necessarily mean higher translation quality. Therefore, we added a translation output evaluation to the experimental work we carried

[2] https://gist.github.com/neubig/2555399

out in this study. Our translation output is the Indonesian text results obtained from the two experiments. We used the Indonesian Pos-tagger to evaluate whether our translation output complies with Indonesian word order. With low-resources, it is difficult for Ja-Id to obtain the same translation results as other high-resource language pairs, that is, Ja-En. Therefore, in this study, we investigated approaches that could be used to improve the BLEU scores while also producing comprehensible translation output.

The SMT pipeline consists of training, tuning, and evaluation. We used different training settings, that is, LM order and automatic word alignment, in our two experiments. We employed two LM orders, namely, 3-gram and 5-gram, in the WoR and WR experiments, respectively. We also utilized two automatic word alignments, GIZA++ and MGIZA, in the WoR and WR experiments, respectively. We used different tuning, that is, MERT and MIRA, to tune the 2K datasets in the WoR and WR experiments, respectively.

Without Reordering (WoR)

In this part, we constructed several systems as follows:

- The direct translation is a translation system between src-trg, i.e., Ja-Id. We used this system as a baseline for acquiring feature functions.
- A single pivot system is a translation system that combines two phrase tables. We employed two pivot approaches, namely triangulation and LI. In triangulation, we combine two phrase tables, namely, src-pvt and pvt-trg phrase tables. In LI, we also combine two phrase tables, namely, the triangulation and src-trg phrase tables. Therefore, we obtained four single pivot systems in each approach: JaId-En, JaId-My, JaId-Ms, and JaId-Fi, as listed in Table 4.3.
- A multiple pivots system is a translation system that combines four phrase tables while using LI approaches. We constructed two systems, namely, JaId(EnMsFiMy) and Baseline(EnMsFiMy). The JaId(EnMsFiMy) is a combination of four phrase tables, viz., the phrase tables of JaId-En, JaId-Ms, JaId-Fi, and JaId-My, respectively. These combinations are based on the BLEU scores obtained from the LI of the single pivot in ascending order, that is, 11.34 of JaId-En, 12.11 of JaId-Ms, 12.15 of JaId-Fi, and 12.21 of JaId-My. Finally, the Baseline(EnMsFiMy) is a combination of two phrase tables, viz., the phrase table of the baseline and the phrase table of JaId(EnMsFiMy). Table 4.3 presents results obtained with our multiple-pivot system.

With Reordering (WR)

Certain important differences exist between the WoR and WR experiments, namely, the source language, multiple pivot system, and src-pvt phrase table. In this WR experiment, we reordered the source language into the target language word order. Then, in contrast to the multiple pivot system of WoR, which directly combines four phrase tables, we constructed three sub-systems of multiple pivots of WR: two pivot, three pivot, and four pivot systems. Additionally, we did not combine our multiple-pivot system with the baseline phrase table. This is based on the result of the WoR experiment, which shows that the BLEU score tends to decrease when combined with the baseline phrase table. Finally, we applied our simple strategy to the phrase table of src-pvt, that is, extending the phrase table, to minimize the Japanese UNK. The system was constructed as follows:

- Direct translation is a src-trg translation system, i.e., Ja-Id. We used this system as a baseline for acquiring feature functions.

- A one pivot system is a translation system that combines two phrase tables, viz., the src-pvt phrase table and the pvt-trg phrase table by using triangulation approaches. We obtained four systems: JaId-En, JaId-Ms, JaId-Fi, and JaId-My, as listed in Table 4.7.

- A two pivot system is a translation system that combines two phrase tables of triangulation by using LI approaches. We obtained six systems, namely, JaId(EnMs), JaId(EnFi), JaId(EnMy), JaId(MsFi), JaId(MsMy), and JaId(FiMy), as listed in Table 4.7.

- A three pivot system is a translation system that combines three phrase tables of triangulation by using LI approaches. We obtained four systems: JaId(EnMsFi), JaId(EnMsMy), JaId(EnFiMy), and JaId (MsFiMy), as listed in Table 4.7.

- A four pivot system is a translation system that combines four phrase tables of triangulation by using LI approaches. We obtained one system, namely, JaId(MsEnFiMy), as listed in Table 4.7.

NMT experiments

We experimented Ja-Id NMT using OpenNMT-py (Klein et al., 2017) framework. We used two pre-trained models, that is, the encoder-decoder and transformer. The encoder-decoder model is a default pre-trained model of OpenNMT-py that use multiple recurrent neural network (RNN) cells and attention types (Bahdanau et al., 2015; Luong et al., 2015) [3]. The model uses 2-layer long short-term memory (LSTM) with 500 hidden units on both the encoder and decoder and dropout 0.3. We constructed two systems of the encoder-decoder model, viz., LSTM-8.5K and LSTM-100K. The LSTM-8.5K system used 8.5K ALT dataset, the same as the previous experiments. We added another dataset, that is, 100K of the TEDTalk (Tiedemann, 2012), for the LSTM-100K system because the translation outputs of LSTM-8.5K were incomprehensible. We run the encoder-decoder model for 200,000 steps for both systems.

The transformer model is a pre-trained model of OpenNMT-py based on the Google Transformer model [4]. The transformer model eschewed recurrence and relied on an attention mechanism to draw global dependency between input and output (Vaswani et al., 2017). Due to long training times, we change the number of layers from 6 to 2. We followed Rubino et al. (2020) that suggests using the number of encoder and decoder layers options, that is, 1, 2, 4, and 6, in order to save computing time. Therefore, our transformer model used 2-layers, learning rate 2.0, batch size 4,096, and dropout 0.1. We constructed one system of the transformer model, that is, transformer-100K. We used the same dataset as LSTM-100K to compare the two models. We run the transformer-100K system for 100,000 steps.

4.1.3 Results and discussion

In this section, we discuss the experimental results based on the BLEU score and translation output. First, we discuss SMT experiments of the src-pvt and pvt-trg results as a building block for pivot approaches. Then, we discuss the results of single and multiple pivots. Last, we discuss NMT experiments and compare with SMT experiment results.

SMT WoR results

Source-pivot and pivot-target results
We performed a direct translation using src-pvt and pvt-trg as the building blocks of pivot approaches. We obtained four src-pvt systems: Ja-En, Ja-Ms, Ja-Fil, and Ja-My. We also obtained four pvt-trg systems: En-Id, Ms-Id, Fi-Id, and My-Id. The second and fifth columns of Table 4.1 show the BLEU scores of the src-pvt and pvt-trg systems, respectively. The results in Table 4.1 are interesting in that language pairs of the same word order, that is, Ja-My, obtained a lower BLEU score, that is, 9.75 in the src-pvt system. In contrast, language pairs of languages with a different word order, that is, Ja-En, obtained the highest BLEU scores, that is, 13.49

[3] https://github.com/OpenNMT/OpenNMT-py
[4] https://github.com/OpenNMT/OpenNMT-py/tree/master/config

in the src-pvt system. Our results contrast those of Bisazza and Federico (2016), who stated that SMT is a more appropriate model for language pairs of the same word order.

We evaluated the translation output of the src-pvt system. We found that most of the translation output of the src-pvt system was incomprehensible. We determined that the translation output follows the word order of the source language, that is, SOV. The second column of Table 4.2 presents an example of the Ja-En translation output. Although Ja-En obtained the highest BLEU score, that is, 13.49, the English translation output is still not well understood because the word order is incorrect. Our result is aligned with that of Callison-Burch et al. (2006) who mentioned that a higher BLEU score was not indicative of higher translation quality.

Table 4.1: BLEU scores of src-pvt and pvt-trg in WoR and WR experiments

Language pairs	src-pvt of WoR experiments	src-pvt of WR experiments	Language pairs	pvt-trg of WoR experiments	pvt-trg of WR experiments
Ja-En	13.49	Preliminary: 8.11 Extend PT: 8.14	En-Id	30.99	27.34
Ja-Ms	12.95	Preliminary: 7.60 Extend PT: 7.56	Ms-Id	35.07	30.27
Ja-Fil	11.22	Preliminary: 7.95 Extend PT: 8.01	Fi-Id	22.57	19.27
Ja-My	9.75	Preliminary: 4.50 Extend PT: 4.45	My-Id	10.02	5.43

Table 4.2: Examples of the translation output of the JaEn src-pvt system

	JaEn of src-pvt system in WoR experiment	JaEn of src-pvt system in WR experiment
Source	地震 は、南東 アジア を 壊滅 さ せ た 2004 年 の インド洋 大 地震 が 襲っ た 日 から ちょうど 二 年 後 に 起き た。	地震 は、ちょうど た 起き 二 年 に 後 から 日 たっ 襲 インド 洋 大 地震 が の 年 2004 た せ さ 壊滅 南東 アジア を 。
Translation results	The earthquake, southeast Asian has devastated 2004, a of the Indian Ocean, the quake struck just two years after the.	the earthquake occurred just two years after hit on from the Indian Ocean massive earthquake of the 2004 Asian of devastating.

The fifth column of Table 4.1 indicates that language pairs with the same word order obtained a high BLEU score, that is, 30.99 and 35.07, for En-Id and Ms-Id, respectively, in the pvt-trg system. Surprisingly, we obtained a relatively high BLEU score, that is, 22.57, for language pairs with a different word order, that is, Fi-Id. The results may be attributable to Filipino belonging to the same language family as Indonesian, that is, Malayo-Polynesian. Our results indicate that, generally, SMT models for language pairs with same word order can produce much higher BLEU score. But even for language pairs with different word order, if they belong to same language family, the model performance is still considerable due to the fact that these languages may potentially share some underline grammatical and syntactic linguistic properties.

Contrary to the translation output of src-pvt, which follows the word order of the source language, the translation output of the pvt-trg system generated better translation and could be understood well, particularly the Ms-Id language pair.

This result was not surprising because Malaysian and Indonesian have the same root language family, that is, Malay. Therefore, both languages have many mutually intelligible morphological properties. We also obtained interesting translation output results for the two language pairs of pvt-trg, that is, En-Id and Fi-Id. We found that certain phrases still follow the source language word order. For example, the phrase *"southeast asia"* is still translated according to the source word order, i.e., *"tenggara asia (southeast asia)"*, instead of according to target word order, i.e.,*"asia tenggara (asia southeast)"*.

Single and multiple pivots results

In this experiment, we constructed eight and two systems for single and multiple pivots, respectively. Table 4.3 shows that triangulation obtained lower BLEU scores than the LI approaches in a single pivot system. However, when the phrase table of triangulation combined with the phrase table of the baseline while using LI approaches, its BLEU score improves. Consider an example of JaId-My of LI that outperformed JaId-My of triangulation by 3.5 in a single pivot system. Moreover, JaId-My of LI outperformed the baseline BLEU score by 0.25. The results indicate that the single pivot system requires a baseline phrase table to obtain a higher BLEU score.

Table 4.3 indicates that multiple pivot systems outperformed the baseline and triangulation of the single pivot system. Consider the example of JaId (EnMsFiMy), which obtained a higher BLEU score than the baseline and JaId-En of triangulation, i.e., 0.24 and 2.49, respectively. The JaId(EnMsFiMy) also obtained a higher BLEU score compared with LI with a single pivot. However, our second multiple pivot system, i.e., Baseline(EnMsFiMy), obtained lower BLEU scores than our first multiple pivot system, i.e., JaId(EnMsFiMy). Table 4.3 indicates that the Baseline(EnMsFiMy) BLEU score decreased by -0.13. The results suggest that multiple pivot systems did not require the phrase table of the baseline because these systems tend to produce lower BLEU scores.

Table 4.3: BLEU scores of single and multiple pivots in without reordering (WoR) experiments

Single pivot			Multiple pivot	
JaId			11.96	
System	Triangulation	LI	System	LI
JaId-En	9.71	11.34	JaId(EnMsFiMy)	12.20
JaId-My	8.71	12.21	Baseline(EnMsFiMy)	12.07
JaId-Ms	8.37	12.11		
JaId-Fi	9.41	12.15		

We investigated the lower BLEU score of Baseline(EnMsFiMy) by using two parameters, viz., phrase table and feature functions. We compared 2,000 of the same phrase pairs from two phrase tables, viz., the phrase table of Baseline(EnMsFiMy) and that of JaId(EnMsFiMy). We found that more than 1,795 phrase pairs of Baseline(EnMsFiMy) obtain the same phrase translation parameter score with the phrase pairs of JaId(EnMsFiMy). Table 4.4 presents an example of the same phrase pair taken from two phrase tables that obtained the same phrase translation parameter scores. Additionally, we found that these two phrase tables have the same size, that is, they contain 1,041,599 phrases each. The Baseline(EnMsFiMy) is a combination of two phrase tables, that is, the phrase tables of the baseline and JaId(EnMsFiMy). The results indicate that the baseline phrase table does not significantly affect the combining process when using the LI approaches.

Table 4.4: Example of phrase pairs and their phrase translation parameter scores.

Phrase-pair	Phrase translation parameters	Score	
		JaId(EnMsFiMy)	Baseline(EnMsFiMy)
は 民主党 で |||nya adalah Demokrat	Inverse phrase translation probability (p(f|e))	0.886859	0.886859
	Inverse lexical weighting (lex(f|e))	0.00138704	0.00138704
	Direct phrase translation probability (p(e|f))	0.888217	0.888217
	Direct lexical weighting (lex(e|f))	0.000010435	0.000010435

The SMT model generated the best translation according to the product of translation and the language model probabilities with the feature function weights. Feature function weights are parameter settings consisting of lexical reordering, distortion, LM, word penalty, phrase penalty, and TM. The feature function weight is stored in the decoder configuration file, that is, moses.ini, during the tuning process. We found that certain feature functions of Baseline(EnMsFiMy) obtain lower weights, viz., distortion, LM, and TM, compared with the feature functions of JaId(EnMsFiMy), as shown in Figure 4.2. We used the same translation model probabilities, or phrase translation parameter score, as provided in Table 4.4 and the same LM order, that is, 3-gram, in Baseline(EnMsFiMy) and JaId(EnMsFiMy). We argue that a lower feature function weight might affect the final translation score. Therefore, we obtained lower BLEU scores with Baseline(EnMsFiMy) compared with JaId(EnMsFiMy).

We evaluated the translation output of multiple pivot systems. We found that most of the translation output of multiple pivot systems did not comply with the Indonesian word order. Figure 4.3 shows an example of the translation output of JaId(EnMsFiMy) in which the incorrect word order was obtained. Consider an example of the word *“melanda (struck)”* as a verb, followed by *“besar (big)”* as an adjective, which is meaningless. As a result, the translation output is incomprehensible. We also identified that, although most of the pvt-trg language pairs have the same word order, this does not significantly affect the translation output of multiple pivot systems. Moreover, we identified that the translation output of multiple pivot systems tends to be the same as that of src-pvt, for example, which follows the source language word order. Table 4.5 provides examples of the translation output

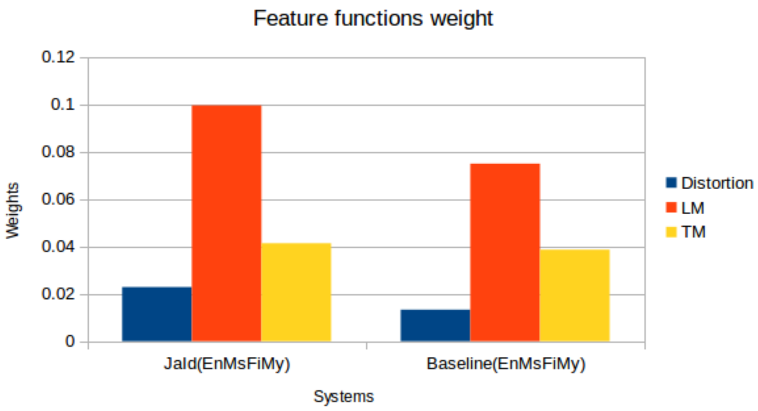
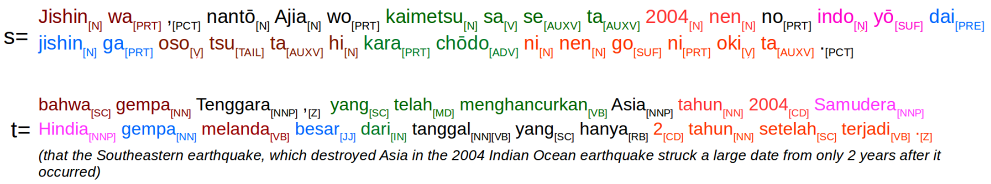


Figure 4.2: Feature functions weight of JaId(EnMsFiMy) and Baseline(EnMsFiMy).

of single and multiple pivot systems. The results indicate that the word order of src-pvt affects the translation output of multiple pivot systems compared with the word order of pvt-trg.



s= Jishin$_{[N]}$ wa$_{[PRT]}$ '$_{[PCT]}$ nantō$_{[N]}$ Ajia$_{[N]}$ wo$_{[PRT]}$ kaimetsu$_{[N]}$ sa$_{[V]}$ se$_{[AUXV]}$ ta$_{[AUXV]}$ 2004$_{[N]}$ nen$_{[N]}$ no$_{[PRT]}$ indo$_{[N]}$ yō$_{[SUF]}$ dai$_{[PRE]}$
jishin$_{[N]}$ ga$_{[PRT]}$ oso$_{[V]}$ tsu$_{[TAIL]}$ ta$_{[AUXV]}$ hi$_{[N]}$ kara$_{[PRT]}$ chōdo$_{[ADV]}$ ni$_{[N]}$ nen$_{[N]}$ go$_{[SUF]}$ ni$_{[PRT]}$ oki$_{[V]}$ ta$_{[AUXV]}$.$_{[PCT]}$

t= bahwa$_{[SC]}$ gempa$_{[NN]}$ Tenggara$_{[NNP]}$ '$_{[Z]}$ yang$_{[SC]}$ telah$_{[MD]}$ menghancurkan$_{[VB]}$ Asia$_{[NNP]}$ tahun$_{[NN]}$ 2004$_{[CD]}$ Samudera$_{[NNP]}$
Hindia$_{[NNP]}$ gempa$_{[NN]}$ melanda$_{[VB]}$ besar$_{[JJ]}$ dari$_{[IN]}$ tanggal$_{[NN][VB]}$ yang$_{[SC]}$ hanya$_{[RB]}$ 2$_{[CD]}$ tahun$_{[NN]}$ setelah$_{[SC]}$ terjadi$_{[VB]}$.$_{[Z]}$
(that the Southeastern earthquake, which destroyed Asia in the 2004 Indian Ocean earthquake struck a large date from only 2 years after it occurred)

Figure 4.3: Translation output example of Ja-Id in without reordering (WoR) experiment.

Table 4.5: Translation output examples of single and multiple pivots in without reordering (WoR) experiments

Source (Ja)	地震 は、南東 アジア を 壊滅 さ せ た 2004 年 の インド洋 大 地震 が 襲っ た 日 から ちょうど 二 年 後 に 起き た。		
System	Approach	BLEU score	Translation output
JaId(En)	Single pivot - Triangulation	9.71	gempa アジア tenggara, dan mereka untuk 壊滅 Samudra Hindia tahun 2004, menghantam gempa besar dari hanya dua tahun setelah 起き.
JaId(My)	Single pivot - LI	12.21	bahwa gempa Tenggara, yang telah menghancurkan Asia tahun 2004 Samudera Hindia gempa melanda besar dari tanggal yang hanya 2 tahun setelah terjadi.
JaId(EnMsFiMy)	Multiple pivots - LI	12.20	bahwa gempa Tenggara, yang telah menghancurkan Asia tahun 2004 Samudera Hindia gempa melanda besar dari tanggal yang hanya 2 tahun setelah terjadi.
Baseline(EnMsFiMy)	Multiple pivots - LI	12.07	gempa SNT.57162.18909 tenggara, yang telah menghancurkan Asia tahun 2004 gempa melanda besar Samudera Hindia, yang hanya dari hari kedkecualiua terjadi pada tahun.

SMT WR results

Source-pivot and pivot-target results

We performed a direct translation of the src-pvt and pvt-trg pivot approaches. We obtained four systems of src-pvt: Ja-En, Ja-Ms, Ja-Fil, and Ja-My. We also obtained four systems of pvt-trg: En-Id, Ms-Id, Fi-Id, and My-Id. In this experiment, we reordered the Japanese word into the Indonesian word order in the src-pvt system. We observed that the Japanese particles in the reordering result tend to move from the right to the left before the noun. Figure 4.1 shows an example of several Japanese particles, viz., を (o$_{\text{[PRT]}}$), で (de$_{\text{[PRT]}}$) that moved from the right in *s* to the left in *s'*. Consider the example of the Japanese particle を (o$_{\text{[PRT]}}$) in words 得点 を (tokuten$_{\text{[N]}}$ o$_{\text{[PRT]}}$ in *s* that become を 得点 (o$_{\text{[PRT]}}$ tokuten$_{\text{[N])}}$ in *s'*.

Unlike the src-pvt of WoR, we implemented our strategy by extending the src-pvt phrase table of WR. The third column of Table 4.1 lists the two BLEU scores, viz., the BLEU scores preliminary and BLEU scores extend. The BLEU score preliminary was obtained from the src-pvt system that uses the phrase table of *grow-diag-final-and*. The BLEU score extends was obtained from the src-pvt system that uses the extended phrase table. We obtained higher BLEU scores, that is, 8.14 and 8.01, for two language pairs, that is, Ja-En and Ja-Fil, respectively. However, two other language pairs, i.e., Ja-Ms and Ja-My, obtained a low BLEU score, i.e., 7.56 and 4.45, respectively.

The result in Table 4.1 is interesting in that src-pvt of WR obtained lower BLEU scores compared with src-pvt of WoR. The pvt-trg of WR also obtained lower BLEU scores compared with the pvt-trg of WoR. We identified the reasons that might cause the BLEU scores to decline, viz., the datasets and tokenizer. The ALT datasets contain line numbers in each sentence. We found that line numbers could interfere with the translation process. The Baseline(EnMsFiMy) of the WoR translation output in Table 4.5 shows that line numbers, that is, SNT.57162.18909, could move to the middle of a sentence during the translation process. These results made the translation output more incomprehensible. Therefore, we removed the line numbers from the WR datasets, as in Table 4.6.

Table 4.6: Example of WoR and WR experiments based on Japanese datasets.

Japanese sentence in WoR
SNT.80188.1 フランス の パリ、パルク・デ・プランス で 行わ れ た 2007 年 ラグビー ワールドカップ の プール C で、 イタリア は 31 対 5 で ポルトガル を 下し た。
Japanese sentence in WR
フランス の パリ、パルク・デ・年 2007 わ れ た 行 で プランス ラグビー ワールド カップ の プール で C 、 イタリア は 対 31 5 で し た 下 ポルトガル を。

Subsequently, we used different tokenizers for the Japanese dataset, viz., MeCab IPA-DIC and Kytea, for WoR and WR, respectively. We identified two effects of different tokenizers: character separation and sentence length. We found that the Kytea tokenizer tends to separate Japanese words with two characters initially by な, い, だ. Examples of Hiragana words, i.e., なく (naku), いる (iru), だが (daga), that become な く (na ku), い る (i ru), and だ が (da ga) when using the Kytea tokenizer. The Mecab IPA-DIC tokenizer stored the word なく as one phrase, whereas the Kytea tokenizer stored the word な く as two phrases, viz., な and く, in the phrase table. As a result, the separation effect could minimize the probability of the decoder finding the phrase なく while generating the translation output because it is not available in the phrase table. In addition, the Kytea tokenizer produced longer sentences. Figure 4.4 shows that the Japanese dataset acquired longer sentences when using the Kytea tokenizer in the WR experiment, whereas the other results were similar.

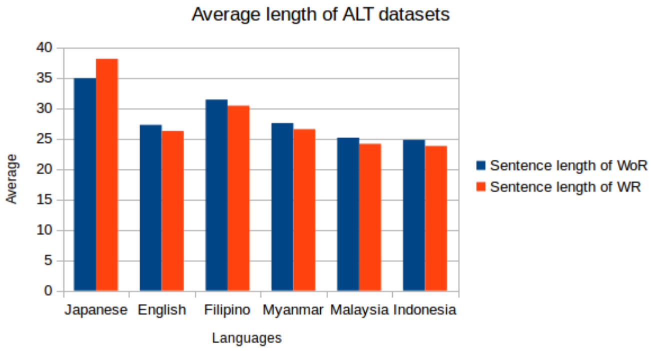


Figure 4.4: Average length of sentences in the ALT datasets in the WoR and WR experiments.

The third column of Table 4.2 shows an example of the translation output of Ja-En of WR. We determined that Ja-En of WR generates more accurate translation output compared to Ja-En of WoR. Consider an example of the sentences of Ja-En of WR that is more understandable, that is, *"the earthquake occurred just two years"*. In contrast, sentences of Ja-En of WoR were difficult to understand, that is, *"The earthquake, southeast Asia has devastated"*. Interestingly, Ja-En of WR obtained lower BLEU scores compared with Ja-En of WoR. The results indicate that reordering the Japanese word into the Indonesian word order positively affects the translation output, even though the system obtained a lower BLEU score.

Single and multiple pivot results

We constructed four systems of one pivot language in this experiment and used the triangulation approach, that is, JaId-En, JaId-Ms, JaId-Fi, and JaId-My. Column two of Table 4.7 shows the BLEU scores of one pivot language system. We compared the single pivot system that obtained the highest BLEU score between WoR and WR, viz., JaId-En, and JaId-Ms, respectively. The JaId-Ms of WR obtained lower BLEU scores, that is, 6.30, compared with JaId-En of WoR, that is, 9.71. However, the translation output of JaId-Ms of WR is more comprehensible than JaId-En of WoR. Table 4.8 presents the JaId-Ms of WR that generated a comprehensible sentence, i.e., *"gempa bumi tersebut terjadi hanya 2 tahun (the earthquake occurred only 2 years)"*. In contrast, JaId-En of WoR still generated UNK and the result was difficult to understand, i.e., *"gempa アジア tenggara, dan mereka untuk 壊滅 (earthquake アジア southeast, and they for 壊滅)"*, as shown in Table 4.5.

We constructed several sub-systems in multiple pivot systems, namely, two pivot,

Table 4.7: BLEU scores of single and multiple pivots in With Reordering (WR) experiment

One pivot language		Two pivot languages		Three pivot languages		Four pivot languages	
System	Triangulation	System	LI	System	LI	System	LI
JaId				6.75			
JaId-En	5.99	JaId(EnMs)	6.92	JaId(EnMsFi)	6.94	JaId(MsEnFiMy)	7.15
JaId-Ms	6.30	JaId(EnFi)	6.29	JaId(EnMsMy)	6.98		
JaId-Fi	5.05	JaId(EnMy)	6.49	JaId(EnFiMy)	6.59		
JaId-My	3.16	JaId(MsFi)	6.73	JaId(MsFiMy)	6.85		
		JaId(MsMy)	6.46				
		JaId(FiMy)	5.57				

Table 4.8: Example of translation output of single and multiple pivots in with reordering (WR) experiment.

Source (Ja)	地震 は、ちょうど た 起き 二 年 に 後 から 日 た っ 襲 インド 洋 大 地震 が の 年 2004 た せ さ 壊滅 南東 アジア を。		
System	Approaches	BLEU score	Translation Output
JaId-Ms	One pivot −Triangulation	6.30	gempa bumi tersebut terjadi hanya 2 tahun setelah dari hari melanda India besar gempa bumi pada tahun 2004 telah menghancurkan Tenggara Asia.
JaId(EnMs)	Two pivot −LI	6.92	gempa terjadi hanya 2 tahun setelah dari hari melanda India gempa besar pada tahun 2004 yang telah menghancurkan Asia selatan.
JaId(EnMsMy)	Three pivot −LI	6.98	gempa terjadi hanya 2 tahun setelah dari hari melanda India gempa besar pada tahun 2004 yang telah menghancurkan Asia Selatan.
JaId(MsEnFiMy)	Four pivot −LI	7.15	gempa terjadi hanya 2 tahun setelah dari hari melanda India gempa besar pada tahun 2004 yang telah menghancurkan Asia Selatan.

three pivot, and four pivot languages. Table 4.7 shows that the BLEU scores improve when a larger number of pivot languages is combined. Consider an example of the highest BLEU score from each system that gradually improves, viz., 6.92, 6.98, and 7.15 for JaId(EnMs), JaId(EnMsMy), and JaId(MsEnFiMy), respectively. Interestingly, even though these systems obtained progressively higher BLEU scores, the translation output was the same, as in Table 4.8. We identified that the improvement in the BLEU scores might be owing to the different number of translated lines from each of the multiple pivot systems. Consider an example of two pivot systems that obtained 210 translated lines, whereas four pivot systems obtained 301 translated lines. The results indicate that increasing the number of pivot languages only affects the BLEU scores, rather than the translation output.

NMT results

Figure 4.5 shows the BLEU scores of Ja-Id NMT experiments. The LSTM-8.5K system obtained a lower BLEU score among other NMT systems, that is, 1.0. In addition, the LSTM-8.5K system produced translation output unrelated to the input and did not match the reference file, as shown in Table 4.9. We found that such results could be caused by a small number of ALT datasets, that is, 8.5K dataset. The vocabulary source and target size were obtained only 24,035 and 24,767, respectively, with the 8.5K dataset. Our results align with Wu et al. (2016), who stated that NMT systems sometimes produce output sentences that fail to completely cover the input, which can result in surprising translations.

We added 100K of the TEDTalk dataset for the LSTM-100K system. The

TEDTalk dataset contains text scripts of TED talks downloaded from OPUS (Tiedemann, 2012). Figure 4.5 shows that the LSTM-100K system obtained a higher BLEU score than the LSTM-8.5K system, that is, 7.8. The LSTM-100K system also obtained a better translation output, as shown in Table 4.9. We identified that word order was not an issue in the Ja-Id NMT experiment with the 100K dataset as the translation output was comprehensible. However, the LSTM-100K system was unabled to generate important words, viz., *"the year 2004"*, *"Indian Ocean"*, and thus could not maximizing BLEU script usage.

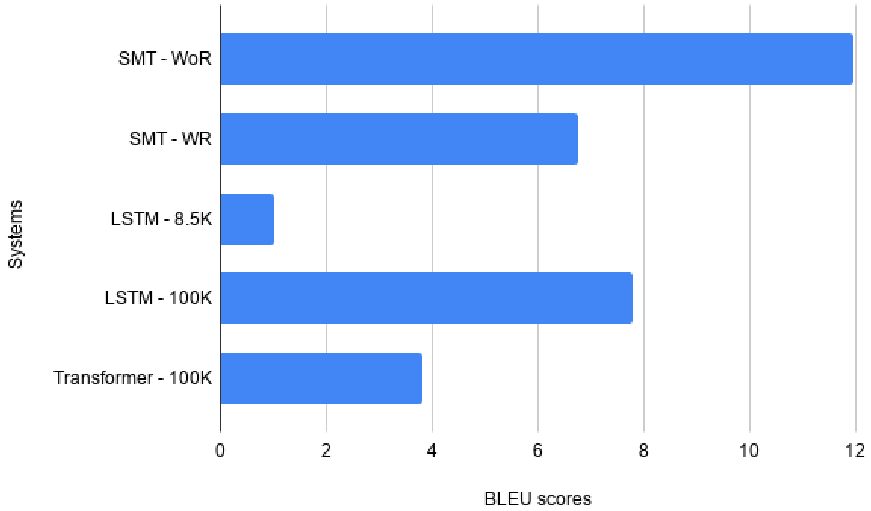


Figure 4.5: BLEU scores of the Ja-Id language pair in the SMT and NMT systems.

Figure 4.5 shows that the Transformer-100K system obtained a lower BLEU score than the LSTM-100K system, that is, 3.8. The lower BLEU score could be because the transformer model is very sensitive to hyper-parameters and the amounts of the pre-trained model datasets were different from our Transformer-100K system dataset. The pre-trained model was trained for English-German language pair with 4.5 million sentence pairs and used byte-pair encoding (Vaswani et al., 2017), while our dataset contains only 100K of the TEDTalk without byte-pair encoding.

We also show SMT BLEU scores of WoR and WR systems in Figure 4.5. The SMT-WoR is our direct translation system without reordering dataset, while the SMT-WR with reordering one. Figure 4.5 shows that the SMT-WoR system obtained the highest BLEU score, that is, 11.96, even with a small dataset, while the LSTM-100K system obtained a BLEU score of 7.8 with an additional dataset, that is, 100K of the TEDTalk. Our results denote that the SMT system with a small dataset outperformed the NMT systems for the Ja-Id language pair. Our results also align with Rubino et al. (2020), who stated that extremely low-resource NMT requires a large amount of synthetic parallel data obtained with back-translation in order to close the performance gap with the SMT approach.

Table 4.9: Example of translation output of the SMT and NMT systems.

Source	地震 は、南東 アジア を 壊滅 さ せ た 2004 年 の インド洋 大 地震 が 襲っ た 日 から ちょうど 二 年 後 に 起き た。 (The quake occurred just two years after the 2004 Indian Ocean quake that devastated Southeast Asia.)
Reference	Gempa itu terjadi tepat dua tahun sejak gempa tahun 2004 yang terjadi di Samudra Hindia, yang meluluhlantakkan Asia Tenggara. (The quake occurred exactly two years since the 2004 earthquake that struck the Indian Ocean, which devastated Southeast Asia.)
SMT-WoR	bahwa gempa tenggara, yang telah menghancurkan Asia tahun 2004 Samudera Hindia gempa melanda besar dari tanggal yang hanya 2 tahun setelah terjadi (that the Southeastern earthquake, which destroyed Asia in the 2004 Indian Ocean earthquake struck a large data from only 2 years after it occurred)
SMT-WR	gempa bumi tersebut terjadi hanya 2 tahun setelah dari hari melanda India besar gempa bumi pada tahun 2004 telah menghancurkan Tenggara Asia (The earthquake occurred just 2 years after from the day India struck the massive earthquake in 2004 that devastated Southeast Asia)
LSTM-8.5K	Gempa terjadi pada pukul 11:00 a.m. waktu setempat (03.00 UTC), dan saat berjalan lancar (The earthquake occurred at 11:00 a.m. local time (03.00 UTC), and as it went well)
LSTM-100K	Dua tahun kemudian, gempa bumi terjadi setelah gempa bumi yang menghancurkan gempa bumi barat dan menghancurkan Asia Tenggara (Two years later, the earthquake occurred after an earthquake that destroyed the western earthquake and devastated Southeast Asia)
Transformer-100K	Gempa bumi sungguh terjadi setelah gempa bumi dan tsunami yang saya rasakan pada hari 2004 setelah gempa tahun 2004 (The earthquake really happened after the earthquake and tsunami that I felt on the day of 2004 after the earthquake in 2004)

4.2 Phrase table combination based on symmetrization of word alignment

Low-resource languages suffer from data scarcity, which leads to poor translation quality. One of the common techniques to improve translation quality is using a phrase table combination in pivot approaches (Ahmadnia et al., 2017; Budiwati and Aritsugi, 2019; Dabre et al., 2015; Trieu and Nguyen, 2017; Utiyama and Isahara, 2007; Wu and Wang, 2007). The phrase table or translation model comes from a word alignment model that uses a symmetrization technique to generate phrase pairs. The standard symmetrization for word alignment model is grow-diag-final-and (gdfand) (Girgzdis et al., 2014). Although prior studies have shown that non-standard symmetrization, i.e., intersection, could obtain higher BLEU scores than the standard one (Koehn et al., 2005; Singh, 2015; Stymne et al., 2014), non-standard symmetrization has not been commonly used in pivot approaches. Thus, the appropriate symmetrization of word alignment model needs to be investigated to improve the performance of low-resource languages when using phrase table combination in pivot approaches.

Kholy and Habash (2014) studied phrase table combinations based on the symmetrization of word alignment model in pivot approaches for the Hebrew–Arabic language pair. Their proposed approach was based on symmetrization relaxation, which extracted new phrase pairs by removing a given word that was unlisted in the pivot phrase tables. They constructed two new symmetrizations, U_R (union_Relaxation) and GDFA_R (grow-diag-final-and_Relaxation), based on union and gdfand, respectively. The BLEU score from the combination of the GDFA and GDFA_R symmetrizations was 0.8 higher than that of the GDFA_R, which demonstrated the superiority of the phrase table combination.

Unlike Kholy and Habash (2014), in this part, we propose a strategy, i.e., phrase table combination, that uses symmetrization of word alignment, which obtains the highest BLEU scores. Our strategy is based on previous research results that showed that symmetrization of word alignment is language-specific (Koehn et al., 2005; Singh, 2015; Stymne et al., 2014) and dataset-specific (Wu Hua, 2007), as different language pairs with different datasets have different BLEU scores. In contrast to Kholy and Habash (2014), who removed a given word, our strategy employs available parallel corpora of src–trg, src–pvt, and pvt–trg without removing the phrase pair that could be a potential candidate in the translation process.

We applied our proposed strategy in pivot approaches to deal with two low-resource language pairs: Kazakh–English (Kk–En) and Japanese–Indonesian (Ja–Id). Kk–En and Ja–Id are considered low-resource language pairs because of the scarcity of their parallel corpora. In this study, we used Russian and Malaysian as pivot languages for Kk–En and Ja–Id, respectively. Our experimental results demonstrate that our proposed strategy obtains higher BLEU scores than the direct translation and system of standard symmetrization in Kk–En and Ja–Id.

4.2.1 Dataset and experimetal setup

Datasets and Pre-processing

In this study, we used the news-commentary (Barrault et al., 2019) and ALT datasets (Riza et al., 2016) for Kazakh–English (Kk–En) and Japanese–Indonesian (Ja–Id), respectively. We considered the two news domain datasets because of the use of formal language. Some languages have flexible word order, such as Russian (Dryer, 2013). Flexible word order means sentences can use free word order of subject, verb, and object. Since a news domain uses formal language, we expected that the sentences would use one dominant word order. Thus we could analyze the effects of different word orders in the pivot approaches. A news domain uses relatively long sentences, i.e., about more than 22 words on average, as shown in Tables 4.10 and 4.11. We expected that longer sentences would produce various phrase pairs from the symmetrization of word alignment. Additionally, news-commentary and ALT were used as training datasets for low-resource translation tasks in the Workshop on Machine Translation (WMT) 2019 and Workshop on Asian Translation (WAT) 2019.

We performed several pre-processing steps, i.e., tokenizing, normalizing punctuation, re-casing, and filtering sentences for both datasets. The tokenization step separates words and punctuation. We employed Moses (Koehn et al., 2007) to tokenize the Kk, En, and Id, and MeCab (Riza et al., 2016) was used to tokenize Ja. We then normalized the punctuation so that the decoder system could recognize it. Next, the re-casing step reduced the data sparsity by converting the initial word in

Dataset	#Sentences	Average Sentence Length	Vocabulary
Baseline system			
Train			
news-commentary-v14.en-kk.kk	9619	18.09	29,142
news-commentary-v14.en-kk.en	9619	22.15	16,742
Dev			
newsdev2019-enkk.kk	2068	18.02	11,389
newsdev2019-enkk.en	2068	22.23	7726
Language Model			
news-commentary-v14.en	532,560	21.58	-
Interpolation system			
Train			
news-commentary-v14.kk-ru.ru	7230	23.68	27,819
news-commentary-v14.kk-ru.kk	7230	20.12	24,627
news-commentary-v14.en-ru.en	97,652	23.04	51,566
news-commentary-v14.en-ru.ru	97,652	21.35	126,476
Dev			
news-commentary-v14.kk-ru.ru	2000	20.88	11,841
news-commentary-v14.kk-ru.kk	2000	18.05	10,561
newstest2018-ruen.dev.en	3000	20.98	10,108
newstest2018-ruen.dev.ru	3000	17.33	17,091
Language Model			
news-commentary-v14.en-ru.ru	114,375	21.27	-
news-commentary-v14.en-ru.en	114,375	22.98	-

Table 4.10: Dataset statistics of Kazakh-English (Kk-En)

Dataset	#Sentences	Average Sentence Length	Vocabulary
Baseline system			
Train			
DataALT.01.jp-id.SP.true.jp	8500	34.90	19,086
DataALT.01.jp-id.SP.true.id	8500	24.75	28,014
Dev			
DataALT.02.jp-id.true.jp	2000	33.88	6680
DataALT.02.jp-id.true.id	2000	24.26	10,201
Language Model			
DataALT.01.jp-id.SP.true.id	8500	24.75	-
Interpolation system			
Train			
DataALT.01.jp-id.SP.true.jp	8500	34.90	19,086
DataALT.01.jp-id.SP.true.ms	8500	25.11	26,835
DataALT.01.jp-id.PT.true.ms	8500	25.04	26,922
DataALT.01.jp-id.PT.true.id	8500	25.04	28,361
Dev			
DataALT.02.jp-id-true.jp	2000	33.88	6,680
DataALT.02.jp-id-true.ms	2000	24.51	9,888
DataALT.02.jp-id-true.ms	2000	24.51	9,888
DataALT.02.jp-id-true.id	2000	24.26	10,201
Language Model			
DataALT.01.jp-id.SP.true.ms	8500	25.11	-
DataALT.01.jp-id.SP.true.id	8500	24.75	-

Table 4.11: Dataset statistics of Japanese-Indonesian (Ja-Id).

each sentence to its most probable casing. Lastly, we removed sentences that had a length of more than 80 words in the filtering step. We show the dataset statistics for Kk–En and Ja–Id in Tables 4.10 and 4.11, respectively.

Highest-Interpolation System Approach (H-ISA)

Our proposed approach is based on two parts. The first part explores five symmetrizations of word alignment in three sides of the pivot approaches: src–trg, src–pvt, and pvt–trg. This part is named the direct system approach (DSA). The output of the first part is the candidate list of the symmetrization of word alignment. The second part is the phrase table combination using the symmetrization of word alignment that produced the highest BLEU scores. We call this part an interpolation system approach (ISA). The details of DSA and ISA are as follows:

1. The DSA is a direct translation between source–target languages (src–trg), i.e., Kk–En and Ja–Id; source–pivot languages (src–pvt), i.e., Kk–Ru and Ja–Ms; and pivot–target languages (pvt–trg), i.e., Ru–En and Ms–Id, for Kk–En and Ja–Id, respectively. To perform direct translation, we use 3-gram and 5-gram orders as our language model, while we use gdfand, intersection, union, srctotgt, and tgttosrc as symmetrization techniques. The purpose of this DSA is twofold: first, to explore the performance of the used language models and symmetrization techniques by performing a direct translation for the low-resource language pairs; second, to find symmetrization techniques that generate the highest BLEU scores for each of src–trg, src–pvt, and pvt–trg in Kk–En and Ja–Id. Then, we use those symmetrization techniques in our ISA.

2. The ISA is our proposed approach that combines the phrase tables of src–trg, src–pvt, and pvt–trg of Kk–En and Ja–Id. In this ISA, we construct two subsystems: standard-ISA (Std-ISA) and highest-ISA (H-ISA). Std-ISA is our interpolation system that uses gdfand symmetrization technique. H-ISA is our interpolation system that uses the symmetrization technique with the highest BLEU scores from the DSA for each phrase table of src–trg, src–pvt, and pvt–trg. Note that we use a triangulation approach for combining the src–pvt and pvt–trg phrase tables this ISA as follows:

    - We prune the src–pvt and pvt–trg phrase tables using *filter-pt* (Johnson et al., 2007).
    - We then merge the two pruned phrase tables using the triangulation method (Hoang and Bojar, 2015). We modified the triangulation method

(Hoang and Bojar, 2015) for our H-ISA. Given the conditional probabilities of pruned src–pvt ($p(\overline{s}|\overline{p})^*$) and pvt–trg ($p(\overline{p}|\overline{t})^*$) phrase tables, we merge the two phrase tables as shown in Equation 4.3. The two pruned phrase tables were obtained from the DSA of src–pvt and pvt–trg with the highest BLEU scores.

$$\begin{aligned} p(\overline{s}|\overline{t})^* &= \sum_{\overline{p}} p(\overline{s}|\overline{p}, \overline{t})^* p(\overline{p}|\overline{t})^* \\ &\approx \sum_{\overline{p}} p(\overline{s}|\overline{p})^* p(\overline{p}|\overline{t})^* \end{aligned} \tag{4.3}$$

- We merge the triangulation phrase table with the phrase table of src–trg according to Equation 4.4. We modified Equation 4.4 for our H-ISA. Given the conditional probabilities of src–trg phrase table ($p(s|t)^*$) and triangulation phrase table ($p(\overline{s}|\overline{t})^*$), we merge the two phrase tables as follows:

  $$p(s|t;\lambda) = \lambda_{\textit{src-trg}} p(s|t)^* + \lambda_{\textit{pvt}} p(\overline{s}|\overline{t})^* \tag{4.4}$$

  where $\lambda_{\textit{src-trg}}$ and $\lambda_{\textit{pvt}}$ are the interpolation weights of the phrase tables, and $\lambda_{\textit{src-trg}} + \lambda_{\textit{pvt}} = 1$. The src–trg phrase table ($p(s|t)^*$) were obtained from the DSA of src–trg with the highest BLEU scores. Algorithm 1 shows the strategy of our H-ISA.

We use 3-gram and 5-gram orders as our language model in this ISA, as well as in the DSA. We chose 3-gram as the minimum order because it can predict a better probability of the next word than other lower n-grams, i.e., 2-gram. We used 5-gram as the maximum order because Liu et al. (2014) showed that a system using more than 5-gram has relatively the same BLEU score as 5-gram. Therefore, we used the two LM orders to identify which order produces better translation quality and perplexity scores in our low-resource language pairs.

We used MERT (Och, 2003) as a tuning algorithm in our experiments. We used bilingual evaluation understudy (BLEU) (Papineni et al., 2002) and perplexity scores to measure the performance of our work in this study. The BLEU score is a metric used for evaluating the generated sentence compared to the reference sentence. The perplexity score is a metric that defines the performance of a language model order. We measured the perplexity score based on the LM order against the *eval* of the generated sentence. Ideally, a translation system that obtains a higher BLEU score has a lower perplexity score (Liu et al., 2014).

Algorithm 1: Highest-interpolation system approach (H-ISA).

```
Input:
    Corpora of src, pvt, and trg
    $\lambda_{src\text{-}trg}$, $\lambda_{pvt}$
Output: $p(s|t)$ of H-ISA
symms ← [gdfand, intersection, union, srctotgt, tgttosrc]
SrcTrg_HighestBleu ← 0
SrcPvt_HighestBleu ← 0
PvtTrg_HighestBleu ← 0
foreach sym in symms do
    SymBleu ← run Equation 2.3 with sym
    if SymBleu > SrcTrg_HighestBleu then
        SrcTrg_HighestBleu ← SymBleu
    end
    if SymBleu > SrcPvt_HighestBleu then
        SrcPvt_HighestBleu ← SymBleu
    end
    if SymBleu > PvtTrg_HighestBleu then
        PvtTrg_HighestBleu ← SymBleu
    end
end
$p(s|t)*$ ← phrase table of SrcTrg_HighestBleu
$p(\overline{s}|\overline{p})*$ ← phrase table of SrcPvt_HighestBleu
$p(\overline{p}|\overline{t})*$ ← phrase table of PvtTrg_HighestBleu
$p(\overline{s}|\overline{t})^*$ ← run Equation 4.3
$p(s|t;\lambda)$ ← run Equation 4.4 //combine_given_tuning_set to setup the weights of $\lambda_{src\text{-}trg}$ and $\lambda_{pvt}$.
```

4.2.2 Results and discussion

In this section, we discuss the results based on the BLEU score. First, we discuss the DSA results as the basis for the ISA approaches. Then, we discuss the results of the ISA and show the generated text.

Direct System Approach (DSA)

Table 4.12 shows the results of the DSA for each language model. Of 30 experiments, 20 showed that translation systems using LM05 produced higher BLEU scores than those using LM03. Table 4.12 also shows the BLEU scores of each symmetrization of word alignment. We found that the highest BLEU scores were not always generated by gdfand. For example, Kk–En LM05 tgttosrc obtained a higher BLEU score than Kk–En LM05 gdfand, that is, 3.56, showing that non-standard symmetrization could be an alternative option to improving the BLEU scores of the pivot approaches. Our results confirmed the language-specific (Koehn et al., 2005; Singh, 2015; Stymne et al., 2014) and dataset-specific (Wu Hua, 2007) characteristics of the symmetrization of word alignment.

We identified the reasons for the different BLEU scores in the same LM, despite using the same automatic word alignment and decoder weights, i.e., GIZA++ and moses.ini, respectively. We compared the phrase translation parameter scores between two phrase tables of the highest and second-highest BLEU scores in the same LM. Phrase translation parameter scores were computed from the co-occurrence of aligned phrases in the training corpora, then stored in the phrase table along

Language pair system	BLEU scores gdfand	intersection	union	srctotgt	tgttosrc
Kk–En LM03	3.08	2.05	3.07	2.51	**3.36**
Kk–En LM05	3.42	2.26	3.28	2.77	**3.56**
Kk-Ru LM03	**6.22**	4.98	4.31	5.41	5.10
Kk-Ru LM05	**6.49**	5.17	4.35	5.64	5.56
Ru-En LM03	**4.77**	0	2.92	4.09	3.12
Ru-En LM05	**4.63**	0	2.73	3.80	2.85
Ja–Id LM03	**11.96**	10.54	9.55	9.79	11.63
Ja–Id LM05	**12.20**	10.47	9.43	9.82	12.04
Ja-Ms LM03	**12.95**	10.09	10.23	10.46	12.65
Ja-Ms LM05	**13.24**	11.06	10.17	10.54	12.93
Ms-Id LM03	**35.07**	34.66	34.90	34.52	34.99
Ms-Id LM05	35.04	34.75	34.89	34.62	**35.14**

Table 4.12: The obtained bilingual evaluation understudy (BLEU) scores of direct system approach (DSA). Results in bold indicate the highest translation quality.

Language Pair	PT1	PT2	$p(t/s)$	$lex(t/s)$	$p(s/t)$	$lex(s/t)$
Kk–En	tgttosrc	gdfand	37	68	**71**	**89**
Kk-Ru*	gdfand	tgttosrc	**68**	**65**	4	38
Ru-En*	gdfand	tgttosrc	**33**	**34**	0	34
Ja–Id	gdfand	tgttosrc	**338**	**269**	49	101
Ja-Ms	gdfand	tgttosrc	**390**	**300**	39	119
Ms-Id	tgttosrc	gdfand	18	894	**88**	**1060**

Table 4.13: Comparison result of phrase translation parameter scores between two phrase tables. Results in bold indicate the highest scores. PT2 was changed from srctotgt to tgttosrc, marked by an asterisk (*).

with the phrase pair. The scores consisted of inverse phrase translation probability ($p(t|s)$), inverse lexical weighting ($lex(t|s)$), direct phrase translation probability ($p(s|t)$), and direct lexical weight ($lex(s|t)$). First, we collected 2,000 phrase pairs and their phrase translation parameter scores from the phrase table with the highest BLEU score : phrase table 1 (PT1). Subsequently, we collected 12,000 phrase pairs and their phrase translation parameter scores from phrase table with the second-highest BLEU score: phrase table 2 (PT2). Last, we examined which component of the phrase translation parameter of PT1 obtained higher scores than the phrase translation parameter of PT2 in the same phrase pair, as shown in Table 4.13. We changed PT2 from srctotgt to tgttosrc since the results could not obtain the same phrase pairs in two language pairs, Kk–Ru and Ru–En, marked by an asterisk (*) in Table 4.13. The comparison algorithm of phrase translation parameter scores between the two phrase tables can be accessed in our repositories [5].

Table 4.13 shows that most language pairs obtained higher score in $p(t|s)$ and $lex(t|s)$, except for two language pairs: Kk–En and Ms–Id. The inverse phrase translation probability ($p(t|s)$) and inverse lexical weighting ($lex(t|s)$) were obtained from the target–source ($t|s$) parallel corpora. The results indicated that target-source ($t|s$) parallel corpora more strongly influence the phrase translation parameter score than source–target ($s|t$) parallel corpora. Table 4.14 shows the phrase pair and their phrase translation parameter score examples from the two phrase tables.

[5] `https://github.com/s4d3/PhraseTableCombination`

Phrase Pair	Phrase Translation Parameters	Symmetrization	
		gdfand	tgttosrc
2007 жылдан бастап||| since 2007	Inverse phrase translation probability ($p(t/s)$)	0.5	0.5
	Inverse lexical weighting ($lex(t/s)$)	0.000930714	**0.000048791**
	Direct phrase translation probability ($p(s/t)$)	0.5	0.333333
	Direct lexical weighting ($lex(s/t)$)	0.00596183	**0.0128321**

Table 4.14: Example of phrase translation parameter scores in Kk–En LM03. Results in bold indicate the highest scores.

Kk–En			Ja–Id		
LM Order	**Lang Pair**	**H-ISA**	**LM Order**	**Lang Pair**	**H-ISA**
LM03	Kk–En	tgttosrc	LM03	Ja–Id	gdfand
	Kk-Ru	gdfand		Ja-Ms	gdfand
	Ru-En	gdfand		Ms-Id	gdfand
LM05	Kk–En	tgttosrc	LM05	Ja–Id	gdfand
	Kk-Ru	gdfand		Ja-Ms	gdfand
	Ru-En	gdfand		Ms-Id	tgttosrc

Table 4.15: Candidates of symmetrization of word alignment for highest-interpolation system approach (H-ISA).

Interpolation System Approach (ISA)

In the ISA, we constructed two subsystems: Std-ISA and H-ISA. Std-ISA was our interpolation system that uses gdfand, whereas H-ISA uses the symmetrization of word alignment that obtained the highest BLEU score. The choice of symmetrization of word alignment for H-ISA is shown in Table 4.15. Considering Kk–En LM05 H-ISA as an example, we employed tgttosrc in Kk–En as src–trg, whereas we used gdfand in Kk–Ru as src–pvt and Ru–En as pvt–trg.

Table 4.16 shows the ISA result. We included the direct translation src–trg of Kk–En and Ja–Id as a baseline. We found that all the translation systems using LM05 obtained higher BLEU scores than those using LM03. For Kk–En, we found that H-ISA is a competitive approach because it provided absolute improvements of 0.35 and 0.22 BLEU points over baseline and Std-ISA in LM03 and LM05, respectively. Table 4.16 shows the different effect of H-ISA on Ja–Id. H-ISA obtained absolute improvements of 0.11 BLEU points over baseline in LM03. However, H-ISA obtained an absolute drop of –0.12 BLEU points compared to baseline in LM05. We compared 2,000 of the same phrase pairs from two phrase tables: H-ISA and baseline. We provide an example of H-ISA and baseline phrase pairs and phrase translation parameter score in Table 4.17. We found that more than 1,900 phrase pairs of H-ISA in LM05 obtained lower phrase translation parameter scores compared to baseline in LM05. Therefore, lower phrase translation parameter scores could be a reason for the lower BLEU score for Ja–Id using the H-ISA.

Additionally, we investigated why Ja–Id LM03 using the Std-ISA and Ja–Id LM03 using the H-ISA obtained same BLEU score: 12.07 and 12.07, respectively. We found that both systems used the same candidates of symmetrization of word alignment in three sides of the pivot approaches: gdfand in Ja–Id, gdfand in Ja–

Language Model	Baseline	Std-ISA	H-ISA
Kk–En			
LM03	3.08	3.08	**3.43**
LM05	3.42	3.42	**3.64**
Ja–Id			
LM03	11.96	**12.07**	**12.07**
LM05	**12.20**	12.08	12.08

Table 4.16: The obtained BLEU scores of direct translation of src–trg (baseline) and interpolation system approach (ISA). Results in bold indicate the highest translation quality.

Phrase-Pair	Phrase Translation Parameter	Phrase Table		
		Baseline	Std-ISA	H-ISA
この 建物 の ||| bangunan	Inverse phrase translation probability ($p(t/s)$)	0.00952381	*0.00841842*	*0.00849199*
	Inverse lexical weighting ($lex(t/s)$)	0.00000289	*0.00000024*	*0.00000024*
	Direct phrase translation probability ($p(s/t)$)	0.333333	*0.294644*	*0.294915*
	Direct lexical weighting ($lex(s/t)$)	0.488889	*0.431259*	*0.432099*

Table 4.17: Example of phrase translation parameter scores of Ja–Id LM05. Results in italic indicate the lowest score.

Ms, and gdfand in Ms-Id, as shown in Table 4.15. In contrast to Ja–Id LM03, Ja–Id LM05 with the Std-ISA and Ja–Id LM05 with the H-ISA obtained the same BLEU score when using different candidates of symmetrization of word alignment, as shown in Table 4.15. Ja–Id LM05 using the Std-ISA used gdfand in Ja–Id, gdfand in Ja–Ms and gdfand in Ms–Id. Ja–Id LM05 of H-ISA used gdfand in Ja–Id, gdfand in Ja–Ms and tgttosrc in Ms-Id. We compared 2,000 of the same phrase pairs from two phrase tables: Std-ISA and H-ISA. Table 4.17 presents an example of Std-ISA and H-ISA phrase pairs and phrase translation parameter scores. We found that more than 1,700 phrase pairs of Std-ISA and H-ISA obtained relatively similar phrase translation parameter scores. The result demonstrated that relatively similar phrase translation parameter scores using different symmetrization of word alignment could obtain the same BLEU score.

We investigated the relationship between BLEU score and phrase table size in each system, as shown in Tables 4.16 and 4.18. We found that systems with small phrase tables, i.e., Kk–En LM03 of H-ISA, Kk–En LM05 of H-ISA, and Ja–Id LM05 of baseline, obtained higher BLEU scores: 3.43, 3.64 and 12.20, respectively, as shown in Table 4.16. However, we also found that systems with large phrase tables, i.e., Ja–Id LM03 of Std-ISA and Ja–Id LM03 of H-ISA, obtained higher BLEU scores, that is, 12.07 and 12.07, respectively. Our first finding aligns with Tian et al. (2014), who stated that higher BLEU scores could be obtained when using small phrase tables. Our second finding aligns with that of Kholy and Habash (2014), who obtained higher BLEU scores when using large phrase tables. Our results demonstrated that higher BLEU scores can be obtained when using either small or large phrase tables.

We identified why systems with small or large phrase tables could obtain higher

Language Model	Baseline	Std-ISA	H-ISA
Kk–En			
LM03	723,960	742,948	323,850
LM05	723,960	742,948	323,850
Ja–Id			
LM03	875,038	935,717	935,717
LM05	875,038	935,717	925,732

Table 4.18: Phrase table size for direct translation of src–trg (baseline) and ISA.

Language Pair	PT1	PT2	#Same Phrase Pair	PT1				PT2			
				p(ts)	lex(ts)	p(st)	lex(st)	p(ts)	lex(ts)	p(st)	lex(st)
Small phrase table											
Kk–En LM03	H-ISA	Std-ISA	872	296	**514**	**576**	**639**	**576**	358	296	233
Kk–En LM05	H-ISA	Baseline	708	131	342	**412**	**474**	**577**	**366**	296	234
Ja–Id LM05	Baseline	Std-ISA	2000	**1984**	**1975**	**1981**	**1964**	16	25	19	36
Large phrase											
Ja–Id LM03	Std-ISA	Baseline	1825	13	18	17	34	**1812**	**1807**	**1808**	**1791**

Table 4.19: Comparison of phrase translation parameter scores between two phrase tables. Results in bold indicate the highest scores.

BLEU scores when using the same LM order and decoder weights. We compared phrase translation parameter scores between the two phrase tables of the highest and second-highest BLEU score in the same LM. First, we collected 2,000 phrase pairs and their phrase translation parameter scores from the phrase table of the highest BLEU score as phrase table 1 (PT1). Subsequently, we collected 12,000 phrase pairs and their phrase translation parameter scores from the phrase table of second-highest BLEU score as phrase table 2 (PT2). Lastly, we examined whether the phrase translation parameter of PT1 obtains higher scores than that of PT2 in the same phrase pair. We found that the phrase translation parameter of PT1 obtained higher scores than that of PT2 in a system with a small phrase table, as shown in Table 4.19. Consider an example of Kk–En LM03 of H-ISA that obtained higher scores, 514, 576, and 639, in $lex(t|s)$, $p(s|t)$ and $lex(s|t)$, respectively, compared to Kk–En LM03 using the Std-ISA. The result indicated that a system with a small phrase table could obtain a higher BLEU score because of the higher phrase translation parameter scores, particularly in $p(s|t)$ and $lex(s|t)$. In contrast to the system with a small phrase table, we found that the phrase translation parameter of PT1 had a lower score than that of PT2 in a system with a large phrase table. Table 4.19 shows that Ja–Id LM03 had lower scores, 13, 18, 17, and 34, in $p(t|s)$, $lex(t|s)$, $p(s|t)$, and $lex(s|t)$, respectively. The result demonstrated that a system with a large phrase table can obtain higher BLEU scores due to the lower phrase translation parameter scores.

We evaluated the perplexity score of each system, as shown in Table 4.20. We found that the longer LM order of Kk–En, i.e., LM05, obtained a lower perplexity score than the shorter one, i.e., LM03. In contrast, the longer LM order could not obtain a lower perplexity score than the shorter one for Ja–Id. Table 4.20 shows

Language Model	Direct Translation	Std-ISA	H-ISA
Kk–En			
LM03	148.21	148.18	284.05
LM05	93.41	115.90	206.15
Ja–Id			
LM03	309.32	310.25	310.25
LM05	403.13	411.48	414.46

Table 4.20: Perplexity scores for the direct translation of src–trg (baseline) and ISA.

that Ja–Id LM05 obtained higher perplexity scores than Ja–Id LM03, although Ja–Id LM05 obtained a higher BLEU score than Ja–Id LM03. We identified this higher perplexity score in Ja–Id using the target monolingual corpus size as the first parameter. The target monolingual corpus was trained by the LM toolkit, i.e., KenLM, and generated various lists of n-gram probabilities stored in *arpa* file. We compared the English and Indonesian target monolingual corpus in Kk–En and Ja–Id. English, with a larger target monolingual size, i.e., 532,560 and longer LM order, i.e., LM05, had larger various lists of n-gram probabilities, i.e., 29,181,816. In contrast, Indonesian with a smaller target monolingual size, i.e., 8,500 and longer LM order, i.e., LM05, had smaller various lists of n-gram probabilities, i.e., 740,766. As a result, the choice of language model probabilities in the decoding process could be smaller and affected the perplexity score of Ja–Id.

Additionally, we identified another parameter that could have influenced the increase in perplexity scores in Ja–Id LM05: the feature function weight of LM. This parameter weight was generated from the target monolingual corpus and then stored in the decoder, moses.ini. A decoder is an SMT component that finds the best translation according to the product of the translation and language model probabilities (Jurafsky and Martin, 2009). A good value for a feature function weight of LM is 0.1–1. We found that the feature function weight of LM for Kk–En LM05 was higher than for Kk–En LM03 (0.10 and 0.06, respectively) when using a larger target monolingual corpus, i.e., 532,560. In contrast to Kk–En, we found that the feature function weight of LM for Ja–Id LM05 was lower, 0.09, than for Ja–Id LM03, 0.11, when using a smaller target monolingual corpus, i.e., 8,500. As a result, Kk–En LM05 could have obtained a lower perplexity score than Kk–En LM03, whereas Ja–Id LM05 obtained a higher perplexity score than Ja–Id LM03, as shown in Table 4.20. Figure 4.6 shows the feature function weight of LM for Kk–En and Ja–Id.

We evaluated the generated text of the systems. Table 4.21 and Table 4.22 show two sentence examples in each language pair, marked by (1) and (2). Sentence (1) is a long sentence, whereas (2) is a short one. We added the English translation in the generated text of Ja–Id to better understand the translation results, marked in italics. Table 4.21 shows that the H-ISA generated a compact sentence compared to

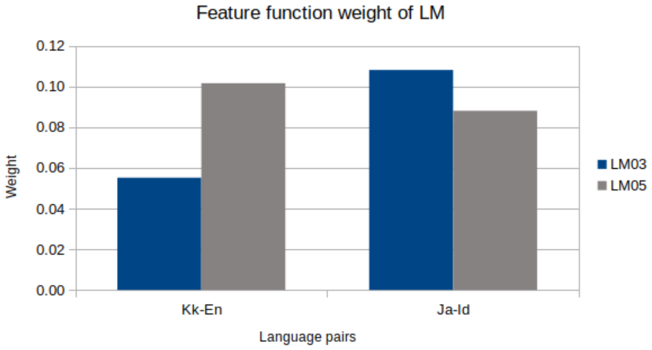


Figure 4.6: Feature function weight of LM for Kk–En and Ja–Id.

	(1)	(2)
Source	экономикадағы оң үрдістер 2 жыл бойы жұмыссыздық деңгейін 4.8% шегінде ұстап тұруға мүмкіндік берді	бұл дегеніміз---инвестиция, жұмыс орындары және бизнес
Reference	positive trends in the economy allowed to keep the unemployment rate within 4.8% for 2 years.	this means investment, jobs and business.
	Direct translation	
LM03	the positive trends identified by the high rate of unemployment, just 2 years keep 4.8% runs into the limits to remain vigilant.	that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business.
LM05	positive trends in the high rate of unemployment, just 2 years keep 4.8% runs into the limits to remain vigilant.	that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business.
	Std-ISA	
LM03	the positive trends identified by the high rate of unemployment, just 2 years keep 4.8% runs into the limits to remain vigilant.	that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business.
LM05	positive trends in the high rate of unemployment, just 2 years keep 4.8% runs into the limits to remain vigilant	that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business.
	H-ISA	
LM03	the positive trends 4.8% 2 years that they rate of unemployment has essentially reached full employment.	that means ensuring that—and business investment, jobs.
LM05	the positive trends 4.8% 2 years they has essentially reached full employment in the rate of unemployment.	this means, and business investment.

Table 4.21: Generated text examples of Kk–En.

others in the short sentence. The compact sentence means that the generated text obtained the same keywords as a reference, i.e., without additional words. Consider an example of Kk–En LM03 H-ISA that generated compact keywords, i.e., *that means ensuring that business investment, jobs*. Kk–En LM05 baseline generated additional words, i.e., *federal government provides, is to reverse the lost*, which were not available in the reference. In contrast to Kk–En, all the systems of Ja–Id generated compact sentences, as shown in Table 4.22.

Tables 4.21 and 4.22 also show that the word order of the generated text was incorrect. We found that the generated text appeared to follow the source language's sentence pattern, i.e., subject-object-verb (SOV), whereas the sentence pattern of the target language is subject-verb-object (SVO). We used the default reordering model, i.e., *msd-bidirectional-fe*; however, our generated text results still followed the source language's pattern. The *msd-bidirectional-fe* is a default reordering model in SMT that considers the orientation of the model, directionality, and languages. The incorrect word order may also be the reason for the insignificant improvement in the BLEU score in our system, as shown in Table 4.16. BLEU is an evaluation metric that measures the similarity between two text strings and assigns too much weight to correct word order (Zhang et al., 2004). BLEU was used a reference file to evaluate

	(1)	(2)
Source	その 2004 年 の 地震 は マグニチュード 9.1 と 記録 さ れ、33 フィート の 高 さ に 達する 波 を 伴う 津波 を 引き起こし た。	これ まで、当局 は 死亡 し た 人 3 人 の 身元 を まだ 確認 でき て い ない。
Reference	Gempa tahun 2004 itu mencatat level kekuatan 9,1 dan menciptakan sebuah Tsunami dengan ketinggian ombak yang mencapai 33 kaki. {That 2004 earthquake registered as a magnitude 9.1 and caused a tsunami with waves reaching as high as 33 feet.}	Sejauh ini pihak berwenang belum mengidentifikasi tiga orang yang tewas. {So far authorities have yet to identify the three people who were killed.}
	Direct translation	
LM03	tersebut pada tahun 2004, mengatakan bahwa gempa bumi dengan kekuatan 9.1 dan 33 kaki dari ketinggian tsunami yang mencapai gelombang. {said in 2004, that an earthquake with a magnitude of 9.1 and 33 feet from the height of the tsunami reached the waves.}	sejauh ini, orang yang tewas dari 3 orang masih belum dapat dikonfirmasi. {so far, people killed from 3 people still cannot be confirmed.}
LM05	tersebut pada tahun 2004 9.1 magnitude gempa dan tercatat, 33 kaki dari ketinggian mencapai menimbulkan gelombang yang menyebabkan tsunami. {in 2004 9.1 magnitude earthquake and recorded, 33 feet from the height reached the wave that caused the tsunami.}	sejauh ini, 3 orang tewas identitas orang yang masih belum dapat dikonfirmasi. {so far, 3 people have died identity of people who still cannot be confirmed.}
	Std-ISA	
LM03	tersebut pada tahun 2004, mengatakan bahwa gempa berkekuatan 9.1 dan 33 kaki dari ketinggian yang mencapai gelombang yang menyebabkan tsunami. {said in 2004, that the magnitude 9.1 and 33 feet from the height reached the waves that caused the tsunami.}	sejauh ini, 3 orang tewas identitas masih belum dapat dikonfirmasi. {so far, 3 people died identity still cannot be confirmed.}
LM05	tersebut pada tahun 2004, dan gempa berkekuatan 9.1 dan 33 kaki dari ketinggian mencapai menimbulkan gelombang yang menyebabkan tsunami. {in 2004, and earthquakes measuring 9.1 and 33 feet from the height reached the waves that caused the tsunami.}	sejauh ini, 3 orang tewas identitas masih belum dapat dikonfirmasi. {so far, 3 people died identity still cannot be confirmed.}
	H-ISA	
LM03	tersebut pada tahun 2004, mengatakan bahwa gempa berkekuatan 9.1 dan 33 kaki dari ketinggian yang mencapai gelombang yang menyebabkan tsunami. {said in 2004, that the magnitude 9.1 and 33 feet from the height reached the waves that caused the tsunami.}	sejauh ini, 3 orang tewas identitas masih belum dapat dikonfirmasi. {so far, 3 people died identity still cannot be confirmed.}
LM05	tersebut pada tahun 2004, mengatakan bahwa gempa berkekuatan 9.1 dan 33 kaki dari ketinggian mencapai menimbulkan gelombang yang menyebabkan tsunami. {said in 2004, that earthquakes measuring 9.1 and 33 feet from the height reached the waves that caused the tsunami}	sejauh ini, 3 orang tewas identitas masih belum dapat dikonfirmasi. {so far, 3 people died identity still cannot be confirmed.}

Table 4.22: Generated text examples of Ja–Id.

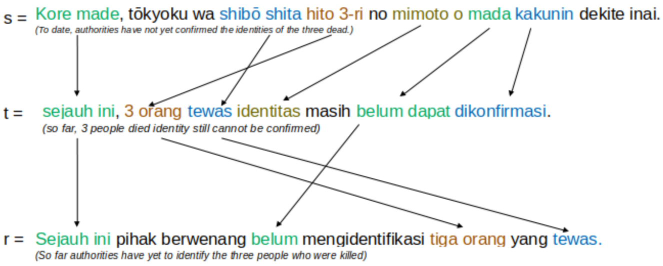

s = Kore made, tōkyoku wa shibō shita hito 3-ri no mimoto o mada kakunin dekite inai.
(To date, authorities have not yet confirmed the identities of the three dead.)
t = sejauh ini, 3 orang tewas identitas masih belum dapat dikonfirmasi.
(so far, 3 people died identity still cannot be confirmed)
r = Sejauh ini pihak berwenang belum mengidentifikasi tiga orang yang tewas.
(So far authorities have yet to identify the three people who were killed)

Figure 4.7: Sentence structure for Ja–Id, taken from LM03 H-ISA. Each color depicts the translation and word position of source–target and target–reference.

the generated text. Our generated text results follow the source languages' sentence pattern, i.e., SOV, whereas our reference files used the SOV pattern. Therefore, the evaluation process could not obtain the maximum results due to the differing word order between the generated text and the reference file. Figure 4.7 illustrates an example of the generated text from the system of Ja–Id LM03 H-ISA. The system translates a source sentence (s) into the target sentence (t). The result showed that the generated Indonesian text had the same word position as the Japanese text. Additionally, Figure 4.7 illustrates a comparison between t and r with a different phrase. Given an example of t, which used the phrase *"3 orang tewas/3 people died"*, r used the phrase *"tiga orang yang tewas/three people who were killed"*. Furthermore, t and r had different word positions, leading to the difficulty in maximizing BLEU script usage.

4.3 Summary

In this chapter, we applied pre-ordering the Japanese sentence for the experiment of word reordering in Japanese to Indonesian (Ja–Id). The *pre-ordering* is a stand-alone task to rearrange words in a target-like order before translating (Bisazza and Federico, 2016). The pre-ordering intend to solve the issue of the previous experiments in multiple pivots for Ja-Id, i.e., the generated text of Indonesian followed the Japanese sentence structure. Thus, our generated text was not comprehensible and hard to understand.

We experimented with single and multiple pivots for translation of Ja-Id by using English, Malay, Filipino, and the Myanmar language as pivot languages. We conducted two SMT experiments, viz., without reordering (WoR) and with reordering (WR) the source language to compare the effects of word reordering by way of two translation qualities, namely, the BLEU score and translation output.

In the WoR experiment, we constructed eight and two systems for single and multiple pivots, respectively. We used two approaches, namely triangulation and LI. The multiple pivot system outperformed the baseline and single pivot by 0.24 and 2.49, respectively. However, the translation output of multiple pivots was incomprehensible because the translation output obtained the same pattern as the src–pvt translation output, that is, it followed the Japanese word order. Although pvt–trg language pairs have the same word order, they do not affect the translation output of multiple pivots. The results indicate that the word order of src–pvt language pairs have a greater effect on the translation output of multiple pivots. Therefore, we argue that we need to improve the multiple pivot translation output, that is, word reordering.

In the WR experiment, we constructed four systems of one pivot language by using triangulation. Then, we constructed three sub-systems of multiple pivots of WR, viz., a two pivot, three pivot, and four pivot system by using LI. Additionally, we did not combine our multiple pivot system with the baseline phrase table. We found that the multiple pivot system produced a higher BLEU score when using more pivot languages than the baseline and one pivot. However, an interesting result was that although multiple pivots of four systems obtained a higher BLEU score, the translation output was the same with two pivot or three pivot systems. The results indicate that more pivot languages only affect the BLEU scores, rather than the translation output.

We also conducted NMT experiments to compare with SMT. We used two pretrained models, viz., encoder-decoder and transformer. We present empirical results to show that SMT outperformed the NMT for Ja-Id as low-resource language pair. SMT system obtained the highest BLEU score, that is, 11.96, even with a small dataset, that is, 8.5K ALT dataset, while the NMT system obtained a BLEU score of 7.8 with an additional dataset, that is, 100K of the TEDTalk. Our results indicates that the SMT obtained better results that NMT, even with a small dataset.

In the second section, we investigated the effect of the symmetrization of word alignment on the translation quality of Kk–En and Ja–Id language pairs in pivot approaches. First, we explored five symmetrization techniques, gdfand, intersection, union, srctotgt, and tgttosrc, in two LM orders, 3-gram and 5-gram, in the direct system approach (DSA). We found that non-standard symmetrization, i.e., tgttosrc, obtained a higher BLEU score than the standard one, i.e., gdfand. We identified that despite using the same automatic word alignment, i.e., GIZA++, the phrase translation parameter score of tgttosrc was higher than that of gdfand. Thus, the BLEU scores of tgttosrc could be much higher than those of gdfand. Additionally, we found that the longer LM order, i.e., 5-gram, obtained a higher BLEU score than the shorter one (3-gram).

Second, we proposed an approach to phrase table combination called H-ISA. The H-ISA is our interpolation system that uses the symmetrization of word alignment that obtained the highest BLEU scores from the DSA for each phrase table of src–trg, src–pvt, and pvt–trg. Our H-ISA is a competitive approach because it outperformed the direct translation of src–trg and standard-ISA (Std-ISA) by 0.38 and 0.22 in Kk–En LM03 and LM05, respectively. The H-ISA also outperformed the direct translation of src–trg by 0.11 in Ja–Id LM03. The direct translation of src–trg still outperformed the H-ISA by 0.12 in Ja–Id LM05. We found that the phrase translation parameter score of H-ISA in Ja–Id was lower compared to that of baseline. We also found that the small phrase table could obtain higher BLEU scores, which contradicts the findings of a previous study (Kholy and Habash, 2014).

Third, we evaluated the perplexity score of the ISA system. We found that the longer LM order, i.e., 5-gram, obtained lower perplexity scores than the shorter one, (3-gram) in Kk–En. However, the results for Ja–Id were the opposite to Kk–En. We found that this result could be caused by two parameters: the target monolingual corpus size and the feature function weight of the LM. For the first parameter, we identified that the English target monolingual corpus size in Kk–En was 62 times bigger, i.e., 532,560, than the Indonesian target monolingual corpus size in Ja–Id, i.e., 8,500. Thus, the English LM obtained 29,181,816 various lists of *n*-gram probabilities, whereas the Indonesian LM obtained 740,766. We argued that smaller datasets could have limited the choice of LM probabilities in the decoding process and affected the perplexity scores of Ja–Id. For the second parameter, Ja–Id LM05 obtained a lower feature function weight of LM, i.e., 0.09, than Ja–Id LM03, i.e., 0.11. As a result, the perplexity scores of Ja–Id LM05 were higher than those of Ja–Id LM03. We argued that the lower feature function weight of LM could obtain lower translation score in the decoding process.

Lastly, we evaluated the generated text from both Kk–En and Ja–Id language pairs. In the short sentence, the H-ISA could produce a more compact keyword compared to direct translation of src–trg and Std-ISA. However, the generated text of H-ISA had an incorrect word order. In the long sentence, the incorrect word order was more severe, resulting in ambiguous sentences in the target language. We found that the generated text tended to follow the source language's sentence pattern, i.e., SOV, but the sentence pattern of the target language is SVO.

Chapter 5

Conclusion and Future Work

In this study, we propose two strategies to improve the translation quality of Kk–En and Ja–Id, viz., extending phrase table, and phrase table combination based on symmetrization of word alignment. In the first strategy: extending phrase table, we merge two phrase tables of src–pvt, viz., src–pvt *gdfand* and src–pvt *tgttosrc* before the phrase table combination process. We employ this strategy in multiple pivots of Ja–Id. Our strategy produce an absolute gain of up to 0.06 BLEU points for Ja–En and Ja–Fil. However, our strategy obtain an absolute drop of -0.05 BLEU points for Ja–Ms and Ja–My. In multiple pivots of Ja–Id, we also employed the *pre-ordering* process for Ja dataset to overcome the issue of different word order between Japanese and Indonesian languages, i.e., SOV and SVO, respectively. As a result, our generated text could be more understand compared to the non pre-ordered Ja dataset.

We employ our second strategy: phrase table combination based on symmetrization of word alignment, in single pivot of Kk–En and Ja–Id. We find that our strategy could be a competitive approach because it outperforms direct translation in Kk-En with absolute improvements of 0.35 and 0.22 BLEU points for 3-gram and 5-gram, respectively. Our strategy also outperforms the direct translation of 3-gram in Ja-Id with an absolute improvement of 0.11 BLEU point.

In this study, we also present empirical results to show that SMT outperformed the NMT for Ja-Id as low-resource language pair. SMT system obtained the highest BLEU score, that is, 11.96, even with a small dataset, that is, 8.5K ALT dataset, while the NMT system obtained a BLEU score of 7.8 with an additional dataset, that is, 100K of the TEDTalk. Our results indicates that the SMT obtained better results that NMT, even with a small dataset.

We listed another direction for future study based on experimental results, as follows:

- We consider implementing the extending phrase table of pvt–trg system, which did not employ in this study. We expect that our strategy would improve the pvt–trg systems, which would affect the performance of multiple pivots of Ja-Id.

- In a single pivot of the WR of multiple pivots of Ja–Id, we obtained a lower BLEU score using triangulation approaches. We plan to combine the src–pvt and pvt–trg phrase tables using LI to compare with triangulation approaches.

- We aim to investigate the cause of the similarity between two pivot, three pivot, and four pivots translation outputs in the WR of multiple pivots of Ja–Id.

- We plan to investigate the optimization of hyper-parameters, pre-processing of the NMT experiment. We also aim to apply other NMT models, i.e., transfer learning and multilingual, to compare with multiple pivots of SMT results.

- We will increase our Indonesian target monolingual corpus size of Ja–Id to be as large as Kk–En. Then, we will re-evaluate the parameter of target monolingual corpus as a factor for decreasing the perplexity scores.

- The applicability of our proposed strategies was demonstrated on limited language pairs. Thus, another direction is investigating other language pairs and datasets.

Bibliography

Cosmas Krisna Adiputra and Yuki Arase. 2017. Performance of Japanese-to-Indonesian Machine Translation on Different Models. In *The 23rd Annual Meeting of the Society of Language Processing* (University of Tsukuba). The Association for Natural Language Processing.

Benyamin Ahmadnia and Javier Serrano. 2017. Employing Pivot Language Technique Through Statistical and Neural Machine Translation Frameworks : The Case of Under-Resourced Persian-Spanish Language Pair. *International Journal on Natural Language Computing* 6, 5 (oct 2017), 37–47. `https://doi.org/10.5121/ijnlc.2017.6503`

Benyamin Ahmadnia, Javier Serrano, and Gholamreza Haffari. 2017. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*. 24–30. `https://doi.org/10.26615/978-954-452-049-6_004`

Benyamin Ahmadnia, Javier Serrano, Gholamreza Haffari, and Nik Mohammad Balouchzahi. 2018. Direct-bridge combination scenario for Persian-Spanish low-resource statistical machine translation. In *Communications in Computer and Information Science*. `https://doi.org/10.1007/978-3-030-01204-5_7`

Cyril Allauzen and Michael Riley. 2011. Bayesian Language Model Interpolation for Mobile Speech Input. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. 1429–1432. `http://www.isca-speech.org/archive/interspeech_2011/i11_1429.html`

Zhenisbek Assylbekov and Assulan Nurkas. 2014. Initial Explorations in Kazakh to English Statistical Machine Translation. In *Proceedings of the The First Italian Conference on Computational Linguistics CLiC-it 2014*. 12. `https://doi.org/10.12871/CLICIT201413`

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.0473

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, 1–61. http://www.aclweb.org/anthology/W19-5301

Arianna Bisazza and Marcello Federico. 2016. Surveys: A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena. *Computational Linguistics* 42, 2 (June 2016), 163–205. https://www.aclweb.org/anthology/J16-2001

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *2011 International Workshop on Spoken Language Translation, IWSLT 2011, San Francisco, CA, USA, December 8-9, 2011*. 136–143. http://www.isca-speech.org/archive/iwslt_11/sltb_136.html

Eleftheria Briakou and Marine Carpuat. 2019. The University of Maryland's Kazakh-English Neural Machine Translation System at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, 134–140. https://doi.org/10.18653/v1/W19-5308

Sari Dewi Budiwati and Masayoshi Aritsugi. 2019. Multiple Pivots in Statistical Machine Translation for Low Resource Languages. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*. Waseda Institute for the Study of Language and Information, Hakodate, Japan, 345–355.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Trento, Italy. https://www.aclweb.org/anthology/E06-1032

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 427–436. https://www.aclweb.org/anthology/N12-1047

David Chiang. 2012. Hope and Fear for Discriminative Training of Statistical Translation Models. *J. Mach. Learn. Res.* 13, null (April 2012), 1159‑1187.

Raj Dabre, Kehai Chen, Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's Supervised Neural Machine Translation Systems for the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*. 168–174. https://doi.org/10.18653/v1/w19-5313

Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2015. Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, Colorado). Association for Computational Linguistics, 1192–1202. https://doi.org/10.3115/v1/N15-1125

Matthew S. Dryer. 2013. Order of Subject, Object and Verb. In *The World Atlas of Language Structures Online*, Matthew S. Dryer and Martin Haspelmath (Eds.). Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info/chapter/81

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language Independent Connectivity Strength Features for Phrase Pivot Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Sofia, Bulgaria). Association for Computational Linguistics, 412–418. http://aclweb.org/anthology/P13-2073

Valdis Girgzdis, Maija Kale, Martins Vaicekauskis, Ieva Zarina, and Inguna Skadina. 2014. Tracing Mistakes and Finding Gaps in Automatic Word Alignments for Latvian-English Translation. In *Human Language Technologies - The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014, Kaunas, Lithuania, September 26-27*. 87–94. https://doi.org/10.3233/978-1-61499-442-8-87

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese Statistical Machine Translation Using English As Pivot Language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (Athens, Greece) *(StatMT '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 173–181. http://dl.acm.org/citation.cfm?id=1626431.1626467

Kenneth Heafield. 2011a. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, United Kingdom, 187–197. https://kheafield.com/papers/avenue/kenlm.pdf

Kenneth Heafield. 2011b. KenLM: Faster and smaller language model queries. In *Proc. of the Sixth Workshop on Statistical Machine Translation*.

Kenneth Heafield, Chase Geigle, Sean Massung, and Lane Schwartz. 2016. Normalized Log-Linear Interpolation of Backoff Language Models is Efficient. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. http://aclweb.org/anthology/P/P16/P16-1083.pdf

Duc Tam Hoang and Ondrej Bojar. 2015. TmTriangulate: A Tool for Phrase Table Triangulation. *Prague Bull. Math. Linguistics* 104 (2015), 75–86. http://ufal.mff.cuni.cz/pbml/104/art-hoang-bojar.pdf

Duc Tam Hoang and Ondrej Bojar. 2016a. Pivoting Methods and Data for Czech-Vietnamese Translation via English. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*. 190–202. https://www.aclweb.org/anthology/W16-3408

Duc Tam Hoang and Ondrej Bojar. 2016b. Pivoting Methods and Data for Czech-Vietnamese Translation via English. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation, EAMT 2017, Riga, Latvia, May 30 - June 1, 2016*. 190–202. https://aclanthology.info/papers/W16-3408/w16-3408

Mark Hopkins and Jonathan May. 2011. Tuning as Ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 1352–1362. https://www.aclweb.org/anthology/D11-1125

Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. 2013. Two-Stage Pre-ordering for Japanese-to-English Statistical Machine Translation. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP*

2013, Nagoya, Japan, October 14-18, 2013. 1062–1066. https://www.aclweb.org/anthology/I13-1147/

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010. Head Finalization: A Simple Reordering Rule for SOV Languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, WMT@ACL 2010, Uppsala, Sweden, July 15-16, 2010.* 244–251. https://www.aclweb.org/anthology/W10-1736/

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2012. HPSG-Based Preprocessing for English-to-Japanese Translation. 11, 3, Article 8 (Sept. 2012), 16 pages. https://doi.org/10.1145/2334801.2334802

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).* http://aclweb.org/anthology/D07-1103

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing, An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition (2Nd Edition).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Amandyk Kartbayev. 2015a. Learning Word Alignment Models for Kazakh-English Machine Translation. In *Integrated Uncertainty in Knowledge Modelling and Decision Making - 4th International Symposium, IUKM 2015, Nha Trang, Vietnam, October 15-17, 2015, Proceedings.* 326–335. https://doi.org/10.1007/978-3-319-25135-6_31

Amandyk Kartbayev. 2015b. SMT: A Case Study of Kazakh-English Word Alignment. In *Current Trends in Web Engineering - 15th International Conference, ICWE 2015 Workshops, NLPIT, PEWET, SoWEMine, Rotterdam, The Netherlands, June 23-26, 2015. Revised Selected Papers.* 40–49. https://doi.org/10.1007/978-3-319-24800-4_4

Ahmed El Kholy and Nizar Habash. 2014. Alignment symmetrization optimization targeting phrase pivot statistical machine translation. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation, EAMT 2014.* European Association for Machine Translation, 63–70.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations.* Association for Computational

Linguistics, Vancouver, Canada, 67–72. https://www.aclweb.org/anthology/P17-4012

Tom Kocmi and Ondrej Bojar. 2019. CUNI Submission for Low-Resource Languages in WMT News 2019. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*. 234–240. https://doi.org/10.18653/v1/w19-5322

Philipp Koehn. 2020. *Neural Machine Translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *2005 International Workshop on Spoken Language Translation, IWSLT 2005, Pittsburgh, PA, USA, October 24-25, 2005*. 68–75. http://www.isca-speech.org/archive/iwslt_05/slt5_068.html

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (Prague, Czech Republic). Association for Computational Linguistics, 177–180. http://aclweb.org/anthology/P07-2045

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, 28–39. https://doi.org/10.18653/v1/W17-3204

Ayana Kuandykova, Amandyk Kartbayev, and Tannur Kaldybekov. 2014. English-Kazakh Parallel Corpus For Statistical Machine Translation. In *International Journal on Natural Language Computing (IJNLC)*. 65. https://doi.org/10.5121/ijnlc.2014.3306

Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, and Pushpak Bhattacharyya. 2017. Utilizing Lexical Similarity between Related, Low-resource Languages for Pivot-based SMT. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 283–289. https://www.aclweb.org/anthology/I17-2048

Xunying Liu, Mark John Francis Gales, and Philip C. Woodland. 2013. Use of contexts in language model interpolation and adaptation. *Computer Speech & Language* 27, 1 (2013), 301–321. https://doi.org/10.1016/j.csl.2012.06.004

Yang Liu, Jiajun Zhang, Jie Hao, and Dakun Zhang. 2014. Making Language Model as Small as Possible in Statistical Machine Translation. In *Machine Translation*, Xiaodong Shi and Yidong Chen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 1412–1421. https://doi.org/10.18653/v1/d15-1166

Bagdat Myrzakhmetov and Zhanibek Kozhirbayev. 2018. Extended Language Modeling Experiments for Kazakh. In *Proceedings of 2018 International Workshop on Computational Models in Language and Speech, CMLS 2018*. CEUR-WS.

Graham Neubig, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a Discriminative Parser to Optimize Machine Translation Reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, 843–853. https://www.aclweb.org/anthology/D12-1077

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. TUFS Asian Language Parallel Corpus (TALPCo). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of Association for Computational Linguistics - Volume 1* (Sapporo, Japan) *(ACL '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 160–167. https://doi.org/10.3115/1075096.1075117

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29, 1 (2003), 19–51. https://doi.org/10.1162/089120103321337421

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics. http://aclweb.org/anthology/P02-1040

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2013. How to Choose the Best Pivot Language for Automatic Translation of Low-Resource Languages. *ACM Trans. Asian Lang. Inf. Process.* 12, 4, Article 14 (Oct. 2013), 17 pages. https://doi.org/10.1145/2505126

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the Importance of Pivot Language Selection for Statistical Machine Translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers* (Boulder, Colorado) *(NAACL-Short '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 221–224. http://dl.acm.org/citation.cfm?id=1620853.1620914

H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. M. Soe, K. T. Nwet, M. Utiyama, and C. Ding. 2016. Introduction of the Asian Language Treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. 1–6. https://doi.org/10.1109/ICSDA.2016.7918974

Rudolf Rosa, Ondrej Dusek, Michal Novak, and Martil Popel. 2015. Translation Model Interpolation for Domain Adaptation in TectoMT. In *Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015)* (Praha, Czech Republic), Vol. 27. 89–96.

Raphael Rubino, Benjamin Marie, Raj Dabre, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2020. Extremely low-resource neural machine translation for Asian languages. *Machine Translation* 34, 4 (2020), 347–382. https://doi.org/10.1007/s10590-020-09258-6

Rico Sennrich. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France) *(EACL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 539–549. http://dl.acm.org/citation.cfm?id=2380816.2380881

H. S. Simon and A. Purwarianti. 2013. Experiments on Indonesian-Japanese statistical machine translation. In *2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*. 80–84. https://doi.org/10.1109/CyberneticsCom.2013.6865786

Thoudam Doren Singh. 2015. An Empirical Study of Diversity of Word Alignment and its Symmetrization Techniques for System Combination. In *Proceedings of the 12th International Conference on Natural Language Processing*. NLP Association of India, Trivandrum, India, 124–129. `https://www.aclweb.org/anthology/W15-5919`

Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2014. Estimating Word Alignment Quality for SMT Reordering Tasks. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*. 275–286. `https://doi.org/10.3115/v1/w14-3334`

M. A. Sulaeman and A. Purwarianti. 2015. Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process. In *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*. 54–58. `https://doi.org/10.1109/ICEEI.2015.7352469`

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 3104–3112. `https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html`

Liang Tian, Derek F. Wong, Lidia S. Chao, and Francisco Oliveira. 2014. A relationship: Word alignment, phrase table, and translation quality. *The Scientific World Journal* 2014 (2014). `https://doi.org/10.1155/2014/438106`

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey.

Hai-Long Trieu. 2017. *A Study on Machine Translation for Low-Resource Languages.* Ph.D. Dissertation. Japan Advanced Institute of Science and Technology.

Hai-Long Trieu and Le-Minh Nguyen. 2017. A Multilingual Parallel Corpus for Improving Machine Translation on Southeast Asian Languages. In *Proceedings of MT Summit XVI, vol.1: Research Track.*

Masao Utiyama and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (Rochester, New York). Association for Computational Linguistics, 484–491. http://aclweb.org/anthology/N07-1061

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008.

Krzysztof Wołk and Agnieszka Wołk. 2018. Augmenting SMT with Semantically-Generated Virtual-Parallel Corpora from Monolingual Texts. In *Trends and Advances in Information Systems and Technologies*, Álvaro Rocha, Hojjat Adeli, Luís Paulo Reis, and Sandra Costanzo (Eds.). Springer International Publishing, Cham, 358–374.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-based Statistical Machine Translation. *Machine Translation* 21, 3 (Sept. 2007), 165–181. https://doi.org/10.1007/s10590-008-9041-6

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144 (2016). arXiv:1609.08144 http://arxiv.org/abs/1609.08144

Wang Haifeng Wu Hua. 2007. Comparative Study of Word Alignment Heuristics and Phrase-Based SMT. *In Proceedings of MT SUMMIT XI* 1 (2007), 507–514. http://www.fjoch.com/

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System?. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/755.pdf

Appendix A

List of Publication

Most of the content of this thesis are based on several published research papers as follows.

Chapter 3 is based on the following paper

1. Sari Dewi Budiwati, Al Hafiz Akbar Maulana Siagian, Tirana Noor Fatyanosa, and Masayoshi Aritsugi. 2019. DBMS-KU Interpolation for WMT19 News Translation Task. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Association for Computational Linguistics, Florence, Italy, 141-146.

2. Sari Dewi Budiwati and Masayoshi Aritsugi. 2019. Multiple Pivots in Statistical Machine Translation for Low Resource Languages. In Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation. Waseda Institute for the Study of Language and Information, Hakodate, Japan, 345-355.

Chapter 4 is based on the following paper

1. Sari Dewi Budiwati and Masayoshi Aritsugi. 2020. Word Reordering on Multiple Pivots for the Japanese and Indonesian Language Pair. Machine Translation Journal Special Issue on Machine Translation for Low-Resources Languages. Springer. Submitted on February 2020, second review on May 2021.

2. Sari Dewi Budiwati, Al Hafiz Akbar Maulana Siagian, Tirana Noor Fatyanosa. Masayoshi Aritsugi. 2020. Phrase Table Combination Based on Symmetrization of Word Alignment for Low-Resource Languages. MDPI Applied Science. 2021.