Preliminary Thesis Review

A Study of Machine Translation for Low-Resource Languages

by

Sari Dewi Budiwati

Submitted to the Graduate School of Science and Technology
KUMAMOTO UNIVERSITY

July 2020

Abstract

This thesis addresses a machine translation (MT) in two low-resource language pairs, namely, Kazakh to English (Kk-En) and Japanese to Indonesia (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 two 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 approaches, viz., extending source-pivot (src-pvt) phrase table and phrase table combination based on different symmetrization. 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 for 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 approach is a phrase table combination based on different

symmetrization. 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 first and second highest BLEU scores of direct translation. Our approach is competitive because it could improve the translation for Kk-En by more than 0.22 compared to the direct translation. Whereas the Ja-Id still obtained the best BLEU score by direct translation by more than 0.12 compared to our approach.

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, and Professor Masahiro Iida, 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 regarding the Statistical Machine Translation (SMT).

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
LI	Linear Interpolation
FI	Fillup Interpolation
MDP	Multiple Decoding Path
BLEU	The Billingual Evaluation Understudy
WMT	Workshop on Machine Translation
DBMS-KU	Database Management System - Kumamoto University
Kk-En	Kazakh to English
Ja-Id	Japanese to Indonesian
src-pvt	source to pivot
pvt-trg	pivot to target
src-trg	source to target
DSE	Direct System Experiments
ISE	Interpolation System Experiments
Std-ISE	Standard Interpolation System Experiments
F-ISE	First Interpolation System Experiments
S-ISE	Second Interpolation System Experiments
NE	Named Entity

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.

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, 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 the single pivot, we proposed a phrase table combination based on different symmetrization technique. These techniques come based on the fact that the pivot approach comprises three direct translations, viz., src-trg, src-pvt, and pvt-trg, that obtained different BLEU scores when different symmetrization employed. Therefore, we did phrase table combinations based on the first and second highest BLEU scores of direct translation. Our approach is competitive because it can improve the translation for Kk-En by more than 0.22 compared to the direct translation. Whereas the Ja-Id still obtained the best BLEU score by direct translation by more than 0.12 compared to our approach.

2. 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 src-pvt 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. As a result, our generated text could be more understand compared to the non pre-ordered Ja dataset.

1.3 Dissertation Outline

The rest of this thesis is organized and continued with Chapter 2, which describes the SMT model and pivot approaches that used in the experiments. Chapter 3 focuses on the single-pivot and multiple pivots approach for Kk-En, and Ja-Id, respectively. 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, and pivot approach. We also describe the current research result of Kazakh to English (Kk-En) and Japanese to Indonesian (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 (LM) and Translation Model (TM) are the two-part that form the basis of SMT. LM is used to determine the fluency of translation output or generated text. Whereas, TM 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 ($A_{\text{TS}} \cup A_{\text{ST}}$).
- 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):

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

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):

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

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`

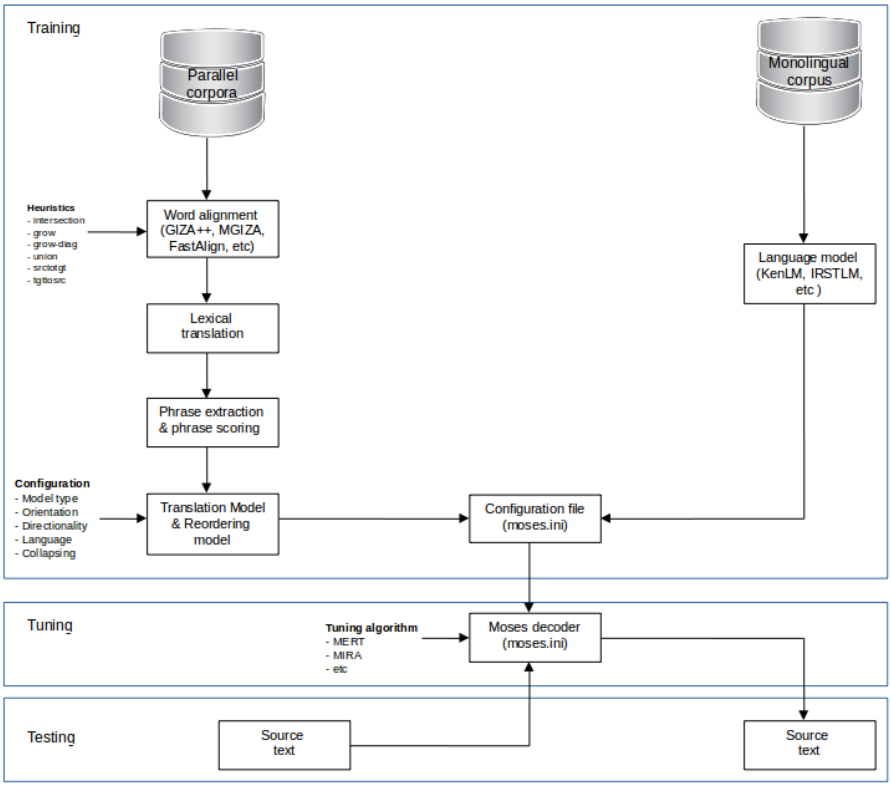

Figure 2.1: SMT architecture

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

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

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 (2003a), 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 source-pivot (src-pvt) and pivot-target (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^e_{i1}, h^e_{i2}, ..., h^e_{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^t_{ij1}, h^t_{ij2}, ..., h^t_{ij8}. The situation is as follows:

$$\begin{aligned} SRC\text{-}PVT &= p_i(h^e_{i1}, h^e_{i2}, ..., h^e_{i8}) \\ PVT\text{-}TRG &= t_{ij}(h^t_{ij1}, h^t_{ij2}, ..., h^t_{ij8}). \end{aligned} \tag{2.4}$$

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

$$S(t_{ij}) = \sum_{m=1}^{8} (\lambda^e_m h^e_{im} + \lambda^t_m h^t_{ijm}) \tag{2.5}$$

where λ^e_m and λ^t_m are weights set by performing minimum error rate training Och (2003a). Finally, t_{best} will be

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

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

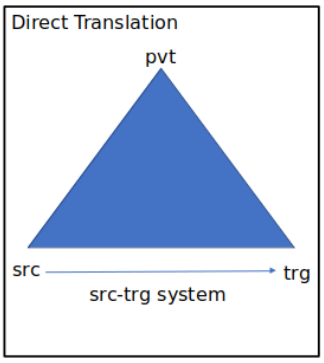

Figure 2.2: Direct translation approach

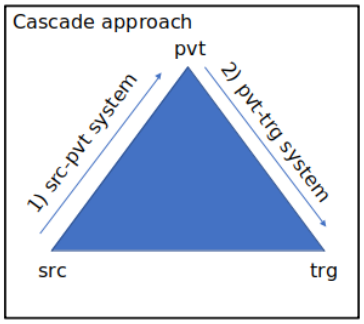

Figure 2.3: Cascade approach

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, $\sum_{i=1}^{n} \alpha_i = 1$, and $\sum_{i=1}^{n} \beta_i = 1$.

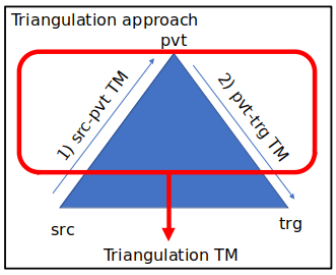

Figure 2.4: Triangulation approach

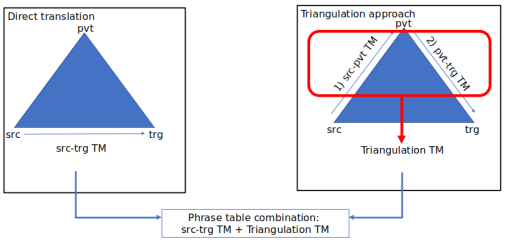

Figure 2.5: Phrase table combination approach

2.3 Previous work

2.3.1 Kazakh to English machine translation

The morphological segmentation approach has been shown a progressive improvement to the translation quality of Kk-En in SMT model (Assylbekov and Nurkas, 2014; Kartbayev, 2015b). The morphological segmentation is an approach that breaks words into morphemes. The approach implemented due to Kazakh is an agglutinative and highly inflected language. These characteristics lead to a different length of phrase when translating to English, as shown in Figure 2.6. Assylbekov and Nurkas (Assylbekov and Nurkas, 2014) have shown the improvement BLEU score from 17.64 to 18.74 in the small corpus, i.e., 636. Whereas Kartbayev (Kartbayev, 2015b) have shown the improvement BLEU score from 30.47 to 31.90 in the small corpus, i.e., 60K. Thus, the morphological segmentation is an important consideration in the SMT model for agglutinative language, i.e., 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 2.6: Example of Kazakh word and its suffixations (Assylbekov and Nurkas, 2014).

Then, the Kk-En machine translation has been introduced as a shared task of low resource language pair in the 2019 Workshop on Machine Translation (WMT 2019) (Barrault et al., 2019). Most participants used NMT model with several approaches, viz., back translation, transfer learning, multilingual transfer learning, and sequence-2-sequence. The transfer learning approach is a similar technique as the pivot approach in the SMT model. Transfer learning uses a high-resource language pair to train the parent model. Subsequently, the parent training data

are replaced with the training data of the low-resource language pair (Kocmi and Bojar, 2019). We identified three submission systems that uses transfer learning, viz., NICT system (Dabre et al., 2019), CUNI system (Kocmi and Bojar, 2019), and UMD system (Briakou and Carpuat, 2019). The NICT and CUNI systems use Russian-English as the parent model and obtain BLEU scores of 26.2 and 18.5, respectively. In comparison, the UMD system uses Turkish-English as a parent model and obtains BLEU scores of 9.2. The experimental results show that the third language still needed to improve the translation quality of Kk-En.

Different from the previous work, we implement a phrase table combination for Kk-En in this study. We proposed a phrase table combination that uses different symmetrizations. To the best of our knowledge, the phrase table combination never employed in the SMT model of Kk-En. The symmetrization technique consist of standard, i.e., *gdfand*, and non-standard, i.e., *union, intersection, srctotgt, tgttosrc,* as described in Section 2.1. Most researchers use standard symmetrization when combining phrase table (Ahmadnia et al., 2017; Dabre et al., 2015; Trieu, 2017; Wu and Wang, 2007). Additionally the standard symmetrization is a default word alignment in Moses [2]. In this study, we use different symmetrization because the translation quality of src-trg, src-pvt, and pvt-trg obtained different BLEU score when non-standard symmetrization employed. Thus, we combine these phrase tables and construct two proposed approach:

- First-Interpolation System Experiments (F-ISE) is our interpolation system that uses the first-best symmetrization obtained from the DSE (Direct System Experiment) for each phrase table of src-trg, src-pvt, and pvt-trg.
- Second-Interpolation System Experiments (S-ISE) is our interpolation system that uses the second-best symmetrization technique obtained from the DSE (Direct System Experiment) for each phrase table of src-trg, src-pvt, and pvt-trg.

Figure 2.7 shows our proposed approach in F-ISE that uses three symmetrization, i.e., *tgttosrc* in src-trg, *gdfand* in src-pvt, and *gdfand* in pvt-trg. Figure 2.8 shows our proposed approach in S-ISE that uses three symmetrization, i.e., *gdfand* in src-pvt, *srctotgt* in src-pvt, and *srctotgt* in pvt-trg. The choice of symmetrization and direct system experiment (DSE) described in Chapter 4 Section 4.2.

2.3.2 Japanese to Indonesian Machine Translation

The Ja-Id has been explored in SMT (Paul et al., 2009), (Simon and Purwarianti, 2013), (Sulaeman and Purwarianti, 2015) and NMT models (Adiputra and Arase,

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

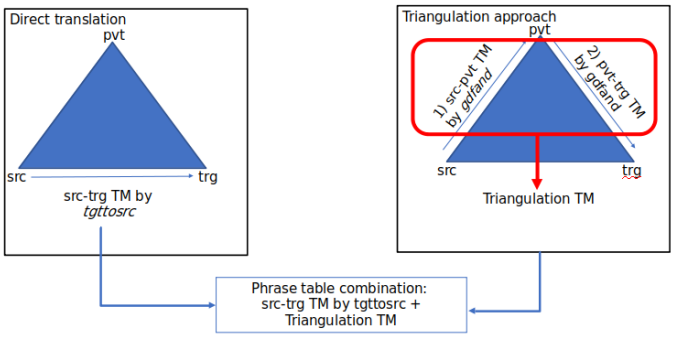

Figure 2.7: Proposed approach: First-Interpolation System Experiments(F-ISE) for Kk-En.

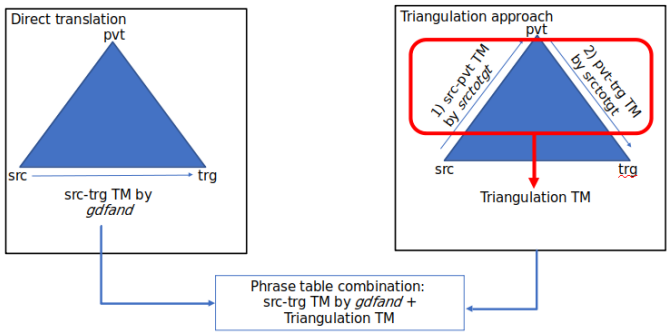

Figure 2.8: Proposed approach: Second-Interpolation System Experiments(S-ISE) for Kk-En.

2017). Paul et al. (Paul et al., 2009) and Cosmas et al. (Adiputra and Arase, 2017) focused on finding a technique that could improve the BLEU scores. Paul et al. find that Ja-Id BLEU scores outperformed the direct translation by using non-English as pivot languages. Whereas Cosmas et al. find that the SMT model outperformed the NMT model by more than 3.93.

Simon et al. (Simon and Purwarianti, 2013) and Sulaeman et al. (Sulaeman and Purwarianti, 2015) focused on finding a technique that could resolve the morphological issue, viz., word order problem, incorrect defined phrase, and words with affixes. Simon et al. proposed several techniques, viz., using pos-tag, increasing the LM dataset, stemming for the Indonesian dataset, removing Japanese particles for the Japanese dataset, and removing Named Entity (NE). The removing Japanese particle outperformed the Baseline by 0.02 compared to other techniques. Sulaeman et al. proposed several techniques, viz., the pos-tag model, the hierarchical model, lemmatizer, and post-processing. The pos tag and lemmatizer outperformed by 0.1 compared to the Baseline. Unfortunately, Simon et al. and Sulaeman et al. did not show the proposed technique generated text. Thus, it is hard to compare the pre-proposed and post-proposed generated text.

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

In this study, we employ *pre-ordering* technique in multiple pivots to overcome the morphological issue, i.e., different word order, from our preliminary work, described in Chapter 4 Section 4.1. The morphological issue also found in previous work (Simon and Purwarianti, 2013; Sulaeman and Purwarianti, 2015). Different from (Sulaeman and Purwarianti, 2015) that employ word order as a part of the decoding process, we employ *pre-ordering* as a stand-alone task before a translation process. Figure 2.9 illustrates word ordering of Japanese sentence into Indonesian sentence structure, using Lader (Neubig et al., 2012).

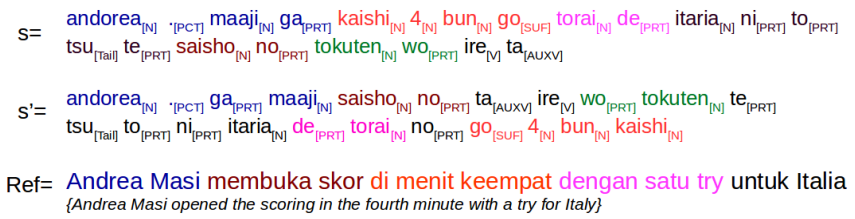

Figure 2.9: Word ordering of Ja sentence structure, i.e., SOV, into Indonesian sentence structure, i.e., SVO, using Lader.

We also proposed two techniques in pivot approach, namely an extended phrase table of src-pvt and phrase table combination based on different symmetrization. The extended phrase table is a merging phrase table from two symmetrization techniques, namely *gdfand* and *tgttosrc*. Our proposed come based on the result that the unknown words of *gdfand* were available in the *tgttosrc* phrase table. Firts, we construct two src-pvt systems, namely src-pvt *gdfand*, and src-pvt *tgttosrc* system. Then, we sorted the unknown words of generated text from the src-pvt *gdfand* system. Subsequently, we query the unknown words from the *tgttosrc* phrase table, identified as *filtered* phrase table. Last, we merged the src-pvt *gdfand* and src-pvt *filtered* phrase tables, identified as extended phrase table. The implementation of our proposed approach could be accessed in repositories [3].

[3]https://github.com/s4d3/Pivot-for-Low-Resource-Languages

We assume that we have a Ja-En *gdfand* phrase table as T_{gdfand} and a Ja-En *tgttosrc* phrase table as $T_{tgttosrc}$. From these tables, we construct a Ja-En extended phrase table as T_{extend}. Each phrase table has four feature functions, namely phrase translation probabilities for both directions, i.e., $\phi(\bar{t}|\bar{s})$ and $\phi(\bar{s}|\bar{t})$, and lexical translation probabilities for both directions, i.e., $p_{\mathrm{w}}(\bar{t}|\bar{s})$, and $p_{\mathrm{w}}(\bar{s}|\bar{t})$. First, we define the unknown words of *gdfand* as *condition(C)* which will be selected in $T_{tgttosrc}$. Subsequently we merged the $T_{filtered}$ and T_{gdfand}. The equation of phrase translation probabilities and lexical translation probabilities are as follows:

$$\phi(\bar{t}|\bar{s})_{\mathrm{filtered}} = \sigma_{\mathrm{C}}(\phi(\bar{t}|\bar{s})_{\mathrm{tgttosrc}}) \tag{2.10}$$

$$p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{filtered}} = \sigma_{\mathrm{C}}(p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{tgttosrc}}) \tag{2.11}$$

$$\phi(\bar{t}|\bar{s})_{\mathrm{extend}} = \phi(\bar{t}|\bar{s})_{\mathrm{filtered}} \cup \phi(\bar{t}|\bar{s})_{\mathrm{gdfand}} \tag{2.12}$$

$$p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{extend}} = (p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{filtered}}) \cup (p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{gdfand}}) \tag{2.13}$$

The second proposed approach is a phrase table combination that uses different symmetrizations. In this study, we use different symmetrization because the translation quality of src-trg, src-pvt, and pvt-trg obtained different BLEU score when non-standard symmetrization employed. Thus, we construct two proposed approaches, namely F-ISE and S-ISE. The proposed approach is the same as in Kk-En. However, the candidate for symmetrization is different. We discuss the differentiation in Chapter 4 Section 4.2.

Our two proposed approaches in Ja-Id are different from the previous work based on the characteristics, as follows:

- In the phrase table combination, most researchers use standard symmetrization, i.e., *gdfand*, when combining phrase table (Ahmadnia et al., 2017; Dabre et al., 2015; Trieu, 2017; Wu and Wang, 2007). In contrast, we use different symmetrization on phrase table combinations because the translation quality of src-trg, src-pvt, and pvt-trg obtained different BLEU scores when non-standard symmetrization employed. This is in line with the findings of several studies stating that the symetrization techniques are language-specific as different language pairs may have different best symmetrization techniques (Koehn et al., 2005; Singh, 2015; Stymne et al., 2014). We discussed the parameters that may caused the different BLEU score in Chapter 4 Section 4.2.

- In the extended phrase table, we employ our finding, i.e., the unknown words of *gdfand* were available in the *tgttosrc* phrase table. Then we merge these two phrase tables. The finding may caused by the different characteristics of *gdfand* and *tgttosrc* word alignment. The *gdfand* align the source and target words from left, right, top, bottom, diagonal neighborhood, and non-neighborhood. Whereas the *tgttosrc* align the source and target words from the target-source GIZA++ alignment file. Therefore, the *tgttosrc* word alignment could obtain a phrase pair that could not be achieved by *gdfand*. As a result, the phrase pair could replace the unknown word of *gdfand*. To the best of our knowledge, the merging process has not been employed in any other pivot research (Ahmadnia et al., 2017; Dabre et al., 2015; Trieu and Nguyen, 2017; Wu and Wang, 2007).

Chapter 3

Single and Multiple pivots in two low-resource languages

This chapter discusses our preliminary work in Kazakh to English (Kk-En) and Japanese to Indonesian (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 on 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 Kazakh-English (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.

Kazakh-English (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). As-

sylbekov and Nurkas (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 (LM) (Allauzen and Riley, 2011; Heafield et al., 2016; Liu et al., 2013) and Translation Model (TM) (Bisazza et al., 2011; Rosa et al., 2015; Sennrich, 2012a). 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 source-pivot (src-pvt) and pivot-target (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.,(Paul et al., 2013) and Dabre et al.,(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 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

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

Table 3.1.

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., 2007a).

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

Experimental setup

We used an open-source Moses(Koehn et al., 2007a) and Giza++ for word alignment, Ken-LM (Heafield, 2011a) for language model, and MERT (Och, 2003b) 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, 2015a). We built 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., 2007a). 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, 2015a). 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

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

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.

Perplexity effects on the Interpolation system

Language model (LM) 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, 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 (Wu and Wang, 2007) and Paul et al.,(Paul et al., 2013) showed that non-English as a pivot language can improve translation quality for specific language pairs. Wu and Wang (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.,(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 (Wu and Wang, 2007) and Dabre et al.,(Dabre et al., 2015) showed promising results by using more than one non-English language. Wu and Wang (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.,(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 part, we investigate single, and multiple pivots approach on Japanese to Indonesian (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 Indonesia 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. (2007b) and Giza++ for word alignment process, phrase table extraction and decoding. We used 3-gram KenLM Heafield (2011b) for language model, MERT Och (2003a) 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. (2007b). 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 (2015b). 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 (2012b). 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-Table3.8 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-pvt 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 Indonesia. 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 and Id-Ja, respectively.

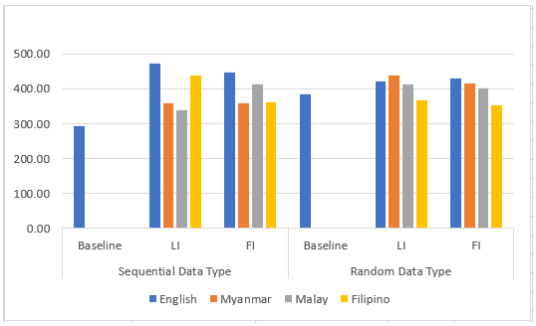

Figure 3.1: Perplexity Score of Ja-Id single pivot for LI and FI approaches.

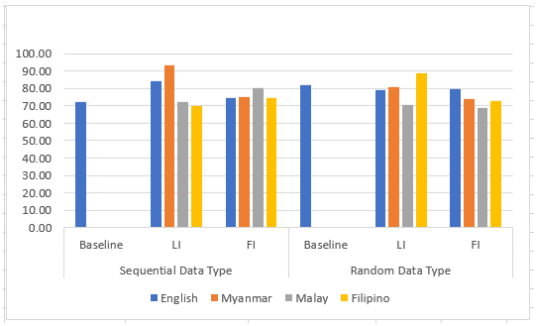

Figure 3.2: Perplexity Score of Id-Ja single pivot for LI and FI approaches.

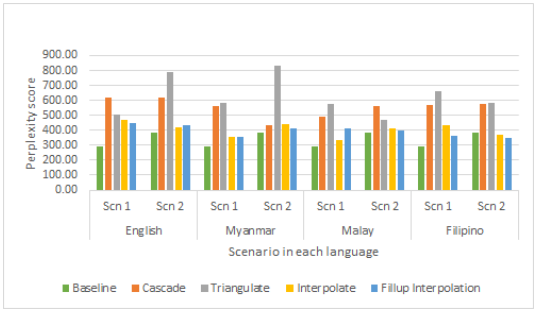

Figure 3.3: Perplexity score for Ja-Id in single pivot.

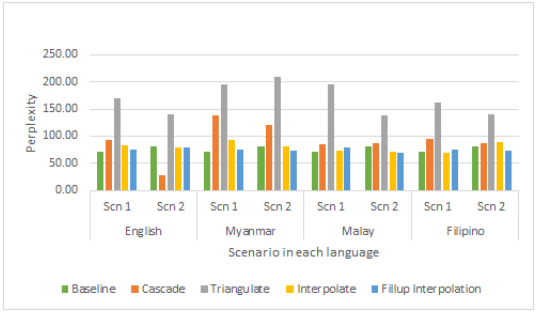

Figure 3.4: Perplexity score for Id-Ja in single pivot.

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)

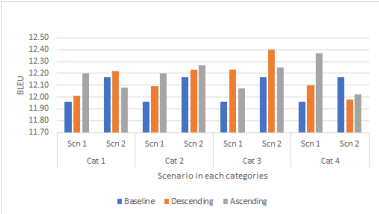

Figure 3.5: BLEU score for Ja-Id in multiple pivots.

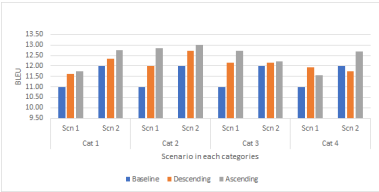

Figure 3.6: BLEU score for Id-Ja in multiple pivots.

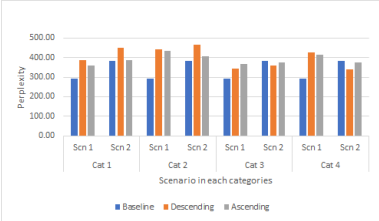

Figure 3.7: Perplexity score for Ja-Id in multiple pivots.

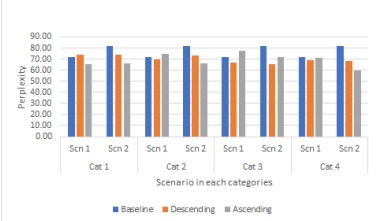

Figure 3.8: Perplexity score for Id-Ja in multiple pivots.

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 the comparison of phrase table combination

In this part, we expand our experiments from the previous one. This chapter divided into two discussions, viz., word reordering in multiple pivots, and comparison of phrase table combination based. The word reordering in 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 the comparison of phrase table combination is the experiment that compares standard and different symmetrization. We describe our technique for two language pairs, viz., Kk-En and Ja-Id.

4.1 Word Reordering in multiple pivots

The SMT model is known as it does not work for language pairs that have different word order. (Bisazza and Federico, 2016; Isozaki et al., 2012; Simon and Purwarianti, 2013). We also find this issue in our previous experiment, i.e., multiple pivots in Japanese to Indonesian (Ja-Id). The word ordering of Japanese is Subject-Object-Verb (SOV), whereas Indonesian is Subject-Verb-Object (SVO). We identified that the generated text of Indonesian from our multiple pivots experiment followed the Japanese sentence structure. As a result, our generated text was not comprehensible and hard to understand.

The word order issue in SMT generally solved by three techniques, e.g., *pre-ordering*, *post-ordering* and word ordering as a part of decoding process (Bisazza and Federico, 2016). The *pre-ordering* is a stand-alone task to rearrange words in a

target-like order before translating (Bisazza and Federico, 2016). The *post-ordering* also stand-alone task to rearrange words in a target-like order after translation process. The *pre-ordering* and *post-reordering* techniques mostly investigate in high-resource language such as Japanese to English (Ja-En) (Goto et al., 2012; Hoshino et al., 2013; Isozaki et al., 2010; Neubig et al., 2011). The lexicalized reordering models (Galley and Manning, 2008; Koehn et al., 2005), hierarchical and syntax-based SMT (Hoang et al., 2009) are approaches used in the decoding process.

In our previous experiment, we did word order as a part of the decoding process by using *lexicalized reordering models*. However, our generated text showed an incomprehensible sentence and hard to understand. The pre-ordering of Japanese into English word order has proven as an effective way to improve the translation quality (Hoshino et al., 2013; Isozaki et al., 2010; Neubig et al., 2012). Due to English has the same word order as Indonesian, i.e., SVO, then we applied the *pre-ordering* of the Japanese sentence in this work.

Differ with our previous multiple pivots experiment that combines four phrase table only. In this work, we combine the phrase table by gradually step, i.e., one pivot, two pivots, three pivots, and four pivots. We took our previous experimental result and identified it as Without Reordering (WoR), whereas our new experiment identified with Reordering (WR) in this work.

4.1.1 Dataset and Experimental Setup

Dataset

The experiments are investigated by the ALT dataset, which is a parallel corpus from English Wikinews to ten languages, namely English, Japanese, Indonesia, Khmer, Malay, Myanmar (Burmese), Filipino, Laotian, Thai and Vietnamese. We use six datasets from ALT, namely Japanese (Ja), Indonesia (Id), English (En), Malaysia (Ms), Filipino (Fi), and Myanmar (My) that consist of 20,106 sentences. We divide ALT datasets into 8.5K for *train*, 2K for *dev* and 1K for *eval*. These datasets used in src-pvt language pairs, namely Ja-En, Ja-Ms, Ja-Fi, and Ja-My. The pvt-trg language pairs consist of En-Id, Ms-Id, Fi-Id, and My-Id.

We used Lader (Neubig et al., 2012) to *pre-ordered* our Japanese dataset. Lader is a reordering model based on context-free-grammar with latent derivations using online discriminative learning. The Figure 4.1 is an example of pre-ordering result that consist of three sentences, viz., s, s', and Ref. The s is an original Japanese sentence that has not pre-ordered, whereas the s' is a result of pre-ordering by using Lader. The Ref is an Indonesian translation reference. Additionally, we used Kytea to identified the Japanese part of speech (POS)[1] and evaluate words

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

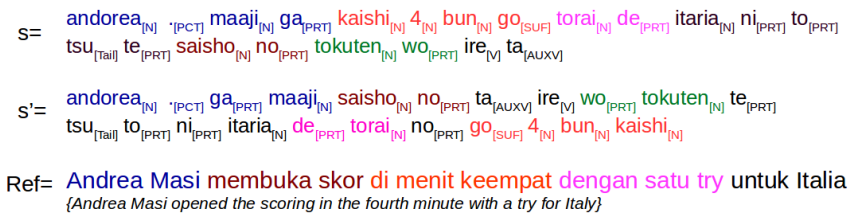

Figure 4.1: Word ordering of Ja sentence structure, i.e., SOV, into Indonesian sentence structure, i.e., SVO, using Lader.

movement. Figure 4.1 show a word movement between s and s'. We identified that several words moved from right to the left side, i.e., 最初 ($_{[N]}$), 得点 (tokuten$_{[N]}$), and 入れ (ire$_{[V]}$). We also identified that Japanese particles tends to move from right to left position before Noun words. For example 得点 を (tokuten$_{[N]}$ o$_{[PRT]}$) become を 得点 (o$_{[PRT]}$ tokuten$_{[N]}$), トライ で (torai$_{[N]}$ de$_{[PRT]}$) become で トライ (de$_{[PRT]}$ torai$_{[N]}$).

Experimental setup

In the *train* process, we used MGIZA as word alignment, 5-gram Language Model (LM) by Ken-LM (Heafield, 2011b) and MIRA to *tune* the 2K dataset. We built several systems namely direct translation, one pivot, two pivots, three pivots, and four pivots. Each system is explained as follows:

- Direct translation is a system between src-trg. This system used as Baseline and to acquire the feature functions, namely 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.

- One pivot system is a combination of src-pvt and pvt-trg phrase tables using Triangulation approach. We obtained four systems, namely *JaId-En*, *JaId-Ms*, *JaId-Fi*, and *JaId-My*.

- Two pivots system is a combination of two Triangulation phrase tables using LI approach. We obtained six systems, namely *JaId(EnMs)*, *JaId(EnFi)*, *JaId(EnMy)*, *JaId(MsFi)*, *JaId(MsMy)*, and *JaId (FiMy)*.

- Three pivots system is a combination of three Triangulation phrase tables using LI approach. We obtained four systems, namely *JaId(EnMsFi)*, *JaId(EnMsMy)*, *JaId(EnFiMy)*, and *JaId (MsFiMy)*.

- Four pivots system is a combination of four Triangulation phrase tables using LI approach. We obtained only only one system, i.e., *JaId(MsEnFiMy)*.

Before combining the src-pvt and pvt-trg phrase table, we did an extended phrase-table of src-pvt. The extended phrase-table is a merging phrase table from two symmetrization technique, namely *grow-diag-final-and(gdfand)*, and *tgttosrc*. Our initiative based on the fact that the unknown words (UNK) of *gdfand* have candidate phrase pairs in the *tgttosrc* phrase table. Therefore, first, we construct two src-pvt systems, namely src-pvt *gdfand*, and src-pvt *tgttosrc* system. Then, we sorted the UNK of generated text from the src-pvt *gdfand* system. Subsequently, we query the UNK from the *tgttosrc* phrase table, identified as *filtered* phrase table. Last, we merged the src-pvt *gdfand* and src-pvt *filtered* phrase tables, identified as extended phrase table.

We assume that we have a Ja-En *gdfand* phrase table as T_{gdfand} and a Ja-En *tgttosrc* phrase table as $T_{tgttosrc}$. From these tables, we construct a Ja-En extend phrase table as T_{extend}. Each phrase table has four feature functions, namely phrase translation probabilities for both directions, i.e., $\phi(\bar{t}|\bar{s})$ and $\phi(\bar{s}|\bar{t})$, and lexical translation probabilities for both directions, i.e., $p_{\mathrm{w}}(\bar{t}|\bar{s})$, and $p_{\mathrm{w}}(\bar{s}|\bar{t})$. First, we define the UNK of *gdfand* as *condition(C)* which will be selected in $T_{tgttosrc}$. Subsequently we merged the $T_{filtered}$ and T_{gdfand}. The equation of phrase translation probabilities and lexical translation probabilities are as follows:

$$\phi(\bar{t}|\bar{s})_{\mathrm{filtered}} = \sigma_{\mathrm{C}}(\phi(\bar{t}|\bar{s})_{\mathrm{tgttosrc}}) \tag{4.1}$$

$$p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{filtered}} = \sigma_{\mathrm{C}}(p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{tgttosrc}}) \tag{4.2}$$

$$\phi(\bar{t}|\bar{s})_{\mathrm{extend}} = \phi(\bar{t}|\bar{s})_{\mathrm{filtered}} \cup \phi(\bar{t}|\bar{s})_{\mathrm{gdfand}} \tag{4.3}$$

$$p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{extend}} = (p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{filtered}}) \cup (p_{\mathrm{w}}(\bar{t}|\bar{s})_{\mathrm{gdfand}}) \tag{4.4}$$

The implementation of our initiative approach could be accessed in repositories [2]. The BLEU score of the extended phrase-table and generated text examples showed in Table 4.3.

4.1.2 Results and Discussion

In this section, we discuss the result based on the BLEU score and generated text. First, we discuss the WoR experiment results. Subsequently, we discuss WR experiment results. We also discuss the src-pvt and pvt-trg results. Then, we discuss the single pivot and multiple pivots. The discussion evaluates the effect of word

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

reordering in each experiment based on the generated text identified by Indonesian pos-tagger (Rashel et al., 2014).

WoR experiment results

In this section, we took the result from our previous research result in Chapter 3.2. We construct eight systems by using Triangulation and Linear Interpolation (LI) techniques in a single pivot. The multiple pivots consist of two systems, namely the JaId(EnMsFiMy) and Baseline(EnMsFiMy). The JaId(EnMsFiMy) system is a combination of four phrase tables, whereas the Baseline (EnMsFiMy is a combination of src-trg with four phrase tables.

Table 4.1 show the WoR experiment BLEU score. The multiple pivots system outperformed the Baseline by 0.24 and 0.11 for JaId(EnMsFiMy) and Baseline(EnMsFiMy), respectively. The JaId(EnMsFiMy) system generally outperformed all single pivot system, except JaId-My, by the difference score 0.01. The Baseline(EnMsFiMy) outperformed the Triangulation system, but poorly compare to LI system.

We evaluate the decline of BLEU score of Baseline(EnMsFiMy). We use two parameters, namely phrase table and feature functions. In the phrase table evaluation, we exchange the phrase table between JaId(EnMsFiMy) and Baseline(EnMsFiMy) system. However, since the amount of phrase pair in each phrase table relatively the same, as much as 1,041,599 then the BLEU score still has the same result. Then, we use second parameter to evaluate the BLEU score, i.e., feature functions. We compare the feature functions from both system and obtained that some feature functions in Baseline(EnMsFiMy) has lower score compare to JaId(EnMsFiMy). Those feature functions are distortion weight by 0.013, language model weight by 0.074, and translation model weight by 0.038. From both parameters, we argue that the feature functions has more effects on the BLEU score decline compare to the phrase table one.

Table 4.2 shows the generated text from the highest BLEU score in each system. Most of the generated text was hard to understand. We take an example in Figure 4.2 that the word *"melanda (struck)"* as verb followed by *"besar (big)"* as adjective does not have a meaning. Usually, in the Indonesian language, the word *"melanda (struck)"* as a verb always followed by a noun, for example, *"melanda Asia (struck Asia)"*. We also identified that the generated text has the same sentence structure as Japanese word order.

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 4.2: Generated text example of Ja-Id in WoR experiment.

Table 4.1: BLEU scores of single and multiple pivots in WoR experiments

Single pivot			Multiple pivot	
JaId		11.96		
Language Pair	Triangulation	LI	Language Pair	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		

WR experiment results

In one pivot, we construct four systems using the Triangulation technique, as shown in Table 4.4. The BLEU score obtained the lowest among other systems, even with the one pivot of the WoR experiment. However, surprisingly the generated text of one pivot WR experiment significantly change by means that it becomes more understand compared to one pivot WoR experiment, as shown in Table 4.5.

Table 4.4 shows several systems of two pivots, three pivots, and four pivots. The result shows that by combining more numbers of pivot languages, then the BLEU score gradually improved. Take an example of the highest BLEU score of each system, viz., 6.30 of one pivot, 6.92 of two pivots, 6.98 of three pivots, and 7.15 of four pivots. However, we found an interesting result that whether the BLEU scores gradually improved, but the generated text of each system has the same result, as shown in Table 4.5.

Table 4.2: Generated text examples of single and multiple pivots in 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 。		
Language pair	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.

Table 4.3: BLEU scores of src-pvt extended phrase table in WR experiment

Language pair	BLEU scores	
	gdfand phrase table	extended phrase table
Ja-En	8.11	8.14
Ja-Ms	7.60	7.56
Ja-Fi	7.95	8.01
Ja-My	4.50	4.45

Table 4.4: BLEU scores of single and multiple pivots in WR experiment

One pivot language		Two pivot language		Three pivot language		Four pivot language	
Language pair	Triangulation	Language Pair	LI	Language pair	LI	Language pair	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				

4.2 The comparison of phrase table combination

4.2.1 Dataset and Experimental Setup

We use news-commentary (Barrault et al., 2019) and ALT datasets (Riza et al., 2016) for Kazakh-English (Kk-En) and Japanese-Indonesia (Ja-Id), respectively. We perform several pre-processing steps, i.e., tokenizing, normalizing punctuation, recasing, and filtering sentences, for both datasets. We employ Moses (Koehn et al., 2007a) to tokenize Kk, En, and Id languages. Meanwhile, MeCab (Riza et al., 2016) is used to tokenize the Ja language. This tokenizing step is to separate words and punctuation. We then normalize the punctuation to be recognized by the decoder system, such as removing extra spaces and normalizing the Unicode punctuation. The recasing step is to reduce the data sparsity by converting the initial word in each sentence to its most probable casing. Last, we remove sentences that have a length of more than 80 words in the filtering step. We show the dataset statistic for Kk-En and Ja-Id in Tables 4.6 and 4.7, respectively.

To our purpose in this study, we have two experiments, namely, Direct System Experiment (DSE) and Interpolation System Experiment (ISE). The detail of DSE and ISE is as follows:

1. The DSE is a direct translation experiment between source and target language (src-trg), i.e., Kk-En and Ja-Id, source and pivot language (src-pvt), i.e., Kk-Ru and Ja-Ms, and pivot and target language (pvt-trg), i.e., Ru-En and Ms-Id, for Kk-En and Ja-Id. To do this, we use 3-gram and 5-gram as our language model, while we use *gdfand, intersection, union, srctotgt,* and *tgt-tosrc* as our symmetrization techniques. The purpose of this DSE is twofold. First, this DSE is to explore the performance of the utilized language models,

Table 4.5: Generated text examples of single and multiple pivots in 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。		
Language pair	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.

i.e., 3-gram and 5-gram, and symmetrization techniques, i.e., *gdfand, intersection, union, srctotgt,* and *tgttosrc*, doing a direct translation for low-resource language pairs. Second, this DSE is to find symmetrization techniques that generate the highest BLEU score for each of src-trg, src-pvt, and pvt-trg in Kk-En and Ja-Id. Then, we use those selected symmetrization techniques in our ISE.

2. The ISE is an experiment that combine a triangulation and src-trg phrase table for translating Kk-En and Ja-Id. In this ISE, we construct three sub systems, namely Standard-ISE (Std-ISE), First-ISE (F-ISE) and Second-ISE (S-ISE). The Std-ISE is our interpolation system that uses *gdfand* as a standard symmetrization technique, obtained from the DSE for each phrase-table of src-trg, src-pvt, and pvt-trg. The F-ISE is our interpolation system that uses the first-best symmetrization technique obtained from the DSE for each phrase table of src-trg, src-pvt, and pvt-trg. On the other hand, the S-ISE is our interpolation system that uses the second-best symmetrization technique obtained from the DSE for each phrase table of src-trg, src-pvt, and pvt-trg. Note that we use a triangulation phrase tables constructed as follows:

   - We prune the src-pvt phrase table and pvt-trg one by using *filter-pt* Johnson et al. (2007a).
   - We then merge two pruned phrase tables by using the Triangulation method Hoang and Bojar (2015a).

   Also, we use 3-gram and 5-gram orders as our language model in this ISE, as well as in the DSE.

We used MERT Och (2003b) as our tuning algorithm in the experiments. We shows BLEU and perplexity scores in the experimental results. The perplexity score is a measurement to define how good is the performance of an LM in translation system. Lower perplexity score indicates better LM, while high perplexity score represent poor LM. We measured the perplexity score of translation system based on the LM order against *eval* generated text.

Table 4.6: Dataset statistics Kk-En

Dataset	#Sentences	Average sentence length	Vocab
Baseline system			
Train			
news-commentary-v14.en-kk.kk	9,619	18.09	29,142
news-commentary-v14.en-kk.en	9,619	22.15	16,742
Dev			
newsdev2019-enkk.kk	2,068	18.02	11,389
newsdev2019-enkk.en	2,068	22.23	7,726
Language Model			
news-commentary-v14.en	532,560	21.58	-
Interpolation system			
Train			
news-commentary-v14.kk-ru.ru	7,230	23.68	27,819
news-commentary-v14.kk-ru.kk	7,230	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	2,000	20.88	11,841
news-commentary-v14.kk-ru.kk	2,000	18.05	10,561
newstest2018-ruen.dev.en	3,000	20.98	10,108
newstest2018-ruen.dev.ru	3,000	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 4.7: Dataset statistics Ja-Id

Dataset	#Sentences	Average sentence length	Vocab
Baseline system			
Train			
DataALT.01.jp-id.SP.true.jp	8,500	34.90	19,086
DataALT.01.jp-id.SP.true.id	8,500	24.75	28,014
Dev			
DataALT.02.jp-id.true.jp	2,000	33.88	6,680
DataALT.02.jp-id.true.id	2,000	24.26	10,201
Language Model			
DataALT.01.jp-id.SP.true.id	8,500	24.75	-
Interpolation system			
Train			
DataALT.01.jp-id.SP.true.jp	8,500	34.90	19,086
DataALT.01.jp-id.SP.true.ms	8,500	25.11	26,835
DataALT.01.jp-id.PT.true.ms	8,500	25.04	26,922
DataALT.01.jp-id.PT.true.id	8,500	25.04	28,361
Dev			
DataALT.02.jp-id-true.jp	2,000	33.88	6,680
DataALT.02.jp-id-true.ms	2,000	24.51	9,888
DataALT.02.jp-id-true.ms	2,000	24.51	9,888
DataALT.02.jp-id-true.id	2,000	24.26	10,201
Language Model			
DataALT.01.jp-id.SP.true.ms	8,500	25.11	-
DataALT.01.jp-id.SP.true.id	8,500	24.75	-

4.2.2 Results and Discussion

Direct System Experiment (DSE)

In this DSE, we obtained 30 results for a translation system that uses 3-gram (LM03) and 5-gram (LM05). Table 4.8 shows that overall the translation systems of using LM05 obtained better BLEU scores than those of using LM03. Additionally, we show that the translation system obtained different BLEU scores when different symmetrization employed.

We investigate the different BLEU scores between standard symmetrization, i.e.,*gdfand*, and non-standard one, i.e., *intersection, union, srctotgt, tgttosrc* in the same LM order. We identified that with the same LM order, i.e., 5-gram, the phrase translation parameters obtained different scores, despite we use the same automatic word alignment, i.e., MGIZA++. The phrase translation parameters is a scores obtained from scoring functions, consist 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 ($lex(s|t)$). The phrase translation parameters were stored in phrase table along with phrase pair. Take an example of Kk-En LM05 *tgttosrc* and *gdfand* that obtained BLEU score of 3.56 and 3.42. First, we took phrase pair and its phrase translation scores, as shown in Table 4.10. Then we compare both phrase translation scores. We identified from Table 4.10 that the inverse lexical weighting (lex(f|e)) and direct lexical weighting (lex(e|f)) scores of *tgttosrc* were higher compare to *gdfand*. We conclude that the parameters affect the BLEU score of *tgttosrc*, therefore its BLEU score were higher compare to the standard one, i.e., *gdfand*.

Table 4.8 shows the obtained BLEU scores of each symmetrization. Several results show that the non-standard symmetrization obtained a higher BLEU score

compare to the standard one. This result denotes that the non-standard one could be an alternative option to improve the BLEU score of the pivot approach. We listed the candidate of symmetrization Kk-En and Ja-Id in Table 4.9.

Table 4.8: The obtained BLEU scores of Direct System Experiments (DSE). Results in bold indicate the first highest translation quality, while those in italic indicate the second highest translation quality.

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	**12.2**	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 4.9: The symmetrization technique candidate for ISE. Results in (1) is a symmetrization technique for F-ISE, and (2) is for S-ISE, when doing phrase table combination

Kk-En				Ja-Id			
LM order	Lang pair	(1)	(2)	LM order	Lang pair	(1)	(2)
LM03	Kk-En	tgttosrc	gdfand	LM03	Ja-Id	gdfand	tgttosrc
	Kk-Ru	gdfand	srctotgt		Ja-Ms	gdfand	tgttosrc
	Ru-En	gdfand	srctotgt		Ms-Id	gdfand	tgttosrc
LM05	Kk-En	tgttosrc	gdfand	LM05	Ja-Id	gdfand	tgttosrc
	Kk-Ru	gdfand	srctotgt		Ja-Ms	gdfand	tgttosrc
	Ru-En	gdfand	srctotgt		Ms-Id	tgttosrc	gdfand

Table 4.10: Example of phrase translation parameter scores in Kk-En LM05. Results in bold indicates the score is higher.

Phrase-pair	Phrase translation parameters	Symmetrization	
		gdfand	tgttosrc
2007 жылдан бастап ||| since 2007,	Inverse phrase translation probability (p(f|e))	0.5	0.5
	Inverse lexical weighting (lex(f|e))	0.000930714	**4.8791e-05**
	Direct phrase translation probability (p(e|f))	0.5	0.333333
	Direct lexical weighting (lex(e|f))	0.00596183	**0.0128321**

Interpolation System Experiment (ISE)

In this ISE, we construct three experiments, viz., Std-ISE, F-ISE, and S-ISE. The Std-ISE is our interpolation system that uses *gdfand* as a standard symmetrization technique, obtained from the DSE for each phrase-table of src-trg, src-pvt, and pvt-trg. The F-ISE is our interpolation system that uses the first-best symmetrization technique obtained from the DSE for each phrase-table of src-trg, src-pvt, and pvt-trg. Last, the S-ISE is our interpolation system that the second-best symmetrization technique obtained from the DSE for each phrase-table of src-trg, src-pvt, and pvt-trg. The choice of symmetrization technique for F-ISE and S-ISE were shown in Table 4.9 with a sign (1) and (2), respectively. We employ *tgttosrc* for Ja-Id as

src-trg and Ja-Ms as src-pvt, then we use *gdfand* for Ms-Id as pvt-trg, when we construct Ja-Id LM05 S-ISE.

Table 4.11 shows the result of ISE. Additionally, we include the direct translation result of Kk-En and Ja-Id as a Baseline. We find that all the translation systems of using LM05 obtained a higher BLEU score than those of using LM03. We find that F-ISE is a competitive approach because it can improve the BLEU score of Kk-En. Take an example of Kk-En LM05 F-ISE that obtained BLEU score 0.22 higher compare to Kk-En LM05 Baseline and Std-ISE. We identified the different BLEU scores based on phrase translation parameters score, as shown in Table 4.13. First, we took the phrase pair and its phrase translation scores, as shown in Table 4.13. Then we compare phrase translation scores. We identified from Table 4.13 that direct phrase translation probability (p(e|f)) and direct lexical weighting (lex(e|f)) scores of *F-ISE* were higher compare to Baseline and Std-ISE. We conclude that the parameters affect the BLEU score of *F-ISE* in Kk-En.

Table 4.11 shows a different effect of F-ISE for Ja-Id. We identified that the BLEU score of Baseline Ja-Id was higher compare to F-ISE. We identified the different BLEU scores based on phrase translation parameters score, as shown in Table 4.12. First, we took the phrase pair and its phrase translation scores, as shown in Table 4.12. Then we compare phrase translation scores. We identified from Table 4.12 that all of Baseline phrase translation scores were higher compare to F-ISE. The result denotes that the parameters affect the BLEU score of *Baseline* in Ja-Id. Additionally, we investigate the same BLEU score of Std-ISE and F-ISE in LM05, i.e., 12.08. Table 4.12 shows that both translation parameters obtained same scores. Based on the results, we conclude that indeed the phrase translation parameters scores determine the improvement and decrement of the BLEU score in the translation system.

Based on Table 4.14, we find that the phrase table size does not directly affect the improvement of the BLEU score. Take an example of Kk-En LM05 F-ISE with smaller phrase table size, i.e., 323,850, and highest BLEU score, i.e., 3.64. In contrast, the Kk-En LM05 Std-ISE has a bigger phrase-table size, i.e., 742,948, but obtained a small BLEU score, i.e., 3.42. This phenomenon also happened on Ja-Id LM05 Baseline that has smaller phrase table size, i.e., 875,038 and highest BLEU score, i.e., 12.20, compare to other systems. This result denotes that the phrase table size does not determine the improvement of the BLEU score.

We also evaluate the perplexity score of each system, as shown in Table 4.15. We find that the longer LM order of Kk-En, i.e., LM05, obtained lower perplexity score in all systems. However, the longer LM order could not obtain a lower perplexity score in Ja-Id. Table 4.15 showed that Ja-Id LM05 obtained higher perplexity score compare to LM03, despite higher BLEU score. We argue this is because of the

monolingual target corpus size. We identified the English monolingual target corpus size in Kk-En has thirteen times bigger dataset, i.e., 114,375 compare to Indonesian monolingual target corpus size in Ja-Id, i.e., 8,500, as shown in Table 4.6 and Table 4.7. The monolingual target corpus size is a dataset used for language model. The bigger dataset then the list of *n*-gram probabilities score will varies. The n-gram probabilities score stored in *arpa* file that generated by Ken-LM (Heafield, 2011a).

Additionally, we identified other parameters that improve the perplexity score based on the dataset. Table 4.7 shows that we use the same dataset for training and LM, i.e., *DataALT.01.jp-id.SP.true.id* in Ja-Id. In contrast, the Kk-En uses different datasets for training and LM, i.e., *news-commentary-v14.en-kk.en* and *news-commentary-v14.en-ru.en*. As a result, the Kk-En F-ISE LM05 obtains a lower perplexity score compare to the LM03. Whereas, the Kk-Id F-ISE LM05 obtain a higher perplexity score compare to the LM03 one. Thus, we argue that the parameter, i.e., training dataset, affects the improvement of the perplexity score of Ja-Id LM05.

Table 4.11: BLEU scores of the system

Language Model	Direct translation	Std-ISE	F-ISE	S-ISE
Kk-En				
LM03	3.08	3.08	3.43	3.09
LM05	3.42	3.42	3.64	3.42
Ja-Id				
LM03	11.96	12.07	12.07	11.16
LM05	12.20	12.08	12.08	11.26

Last, we also evaluate the generated text of the systems, as shown in Table 4.16 and Table 4.17. We show two-sentence examples in each language pair, marked by (1) and (2). The (1) sentence is long, whereas (2) is a short sentence. We add an English translation in Ja-Id generated text in order to better understand the result, marked as an italic sentence.

Table 4.16 shows that F-ISE generate a compact sentence compare to others in a short sentence. The compact sentence means the generated text obtain same keywords as reference, without additional words. Take an example of Kk-En F-ISE LM03 that generate compact keywords, i.e., *that means ensuring that business investment, jobs*. Whereas the Kk-En Baseline LM05 generate additional words, i.e., *that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business*, which is not available in the Reference. In

Table 4.12: Phrase translation scores of Ja-Id LM05. Results in bold indicates the score is higher.

Phrase-pair	Phrase translation parameters	Phrase translation scores		
		Baseline	Std-ISE	F-ISE
から 強制 ||| rumahnya digerebek	Inverse phrase translation probability (p(f|e))	**0.3333**	0.2946	0.2972
	Inverse lexical weighting (lex(f|e))	**0.0002**	0.0001	0.0001
	Direct phrase translation probability (p(e|f))	**0.3333**	0.2946	0.2949
	Direct lexical weighting (lex(e|f))	**1.23E-07**	1.09E-07	1.09E-07
から 情報 を 受け取る ||| menerima informasi dari	Inverse phrase translation probability (p(f|e))	**0.25**	0.2209	0.2229
	Inverse lexical weighting (lex(f|e))	**0.0024**	0.0018	0.0018
	Direct phrase translation probability (p(e|f))	**1**	0.8839	0.8847
	Direct lexical weighting (lex(e|f))	**0.1019**	0.0899	0.0901

Table 4.13: Phrase translation scores of Kk-En LM05. Results in bold indicates the score is higher.

Phrase-pair	Phrase translation parameters	Phrase translation scores		
		Baseline	Std-ISE	F-ISE
алу экономикалық ||| maintain economic	Inverse phrase translation probability (p(f|e))	0.25	0.2484	0.0014
	Inverse lexical weighting (lex(f|e))	0.0011	0.0011	**0.0013**
	Direct phrase translation probability (p(e|f))	0.5	0.4968	**0.9723**
	Direct lexical weighting (lex(e|f))	0.0002	0.0002	**0.7641**
жаңа бір маңызды саясатты ||| important new policies	Inverse phrase translation probability (p(f|e))	0.3333	0.3312	0.3241
	Inverse lexical weighting (lex(f|e))	0.0001	0.0001	0.0006
	Direct phrase translation probability (p(e|f))	0.25	0.2484	**0.972348**
	Direct lexical weighting (lex(e|f))	0.1046	0.1039	**0.1091**

Table 4.14: Phrase table size of the system

Language Model	Baseline	Standard approach (gdfand)	Initiative approach	
			F-ISE	S-ISE
Kk-En				
LM03	723,960	742,948	323,850	730,544
LM05	723,960	742,948	323,850	730,544
Ja-Id				
LM03	875,038	935,717	935,717	750,449
LM05	875,038	935,717	925,732	764,219

Table 4.15: Perplexity scores of the system

Language Model	Baseline	Standard approach (gdfand)	Initiative approach	
			F-ISE	S-ISE
Kk-En				
LM03	148.21	148.18	284.05	148.39
LM05	93.41	115.90	206.15	115.86
Ja-Id				
LM03	309.32	310.25	310.25	310.09
LM05	403.13	411.48	414.46	386.94

contrast, all the systems of Ja-Id generate compact sentence, as shown in Table 4.17.

Table 4.17 and Table 4.17 also show that the generate text obtain the wrong word position. Figure 4.3 illustrated an example of Ja-Id F-ISE LM05 generate text. We identified that the generated text appears to follow the source language's sentence pattern, i.e., SOV (Subject-Object-Verb). Whereas, the sentence pattern of the target language, i.e., Indonesian, is SVO (Subject-Verb-Object). The result showed an incomprehensible sentence and hard to understand.

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 exper-

Table 4.16: Generated text examples of Kk-En.

	(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.
	Baseline	
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-ISE	
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.
	F-ISE	
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.
	S-ISE	
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.

Table 4.17: Generated text examples of Ja-Id.

	(1)	(2)
Source	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。	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.)*
	Baseline	
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-ISE	
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.)*
	F-ISE	
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.)*
	S-ISE	
LM03	pada tahun 2004 tersebut gempa berkekuatan 9.1 dan 33 kaki, yang tinggi yang mencapai tsunami yang menyebabkan gelombang. *(in 2004 the earthquake was 9.1 and 33 feet high, which reached a tsunami that caused waves.)*	para pejabat 3 orang tewas identitas masih tidak bisa dikonfirmasi. *(officials 3 people died identity still could not be confirmed.)*
LM05	tersebut gempa berkekuatan tahun 2004, yang 9.1 dan 33 kaki mencapai tsunami yang menyebabkan gelombang. *(the 2004 magnitude earthquake, which 9.1 and 33 feet hit the tsunami which caused the waves.)*	orang yang berwenang 3 orang tewas identitas masih tidak bisa dikonfirmasi. *(the person authorized 3 people died identity still cannot be confirmed.)*

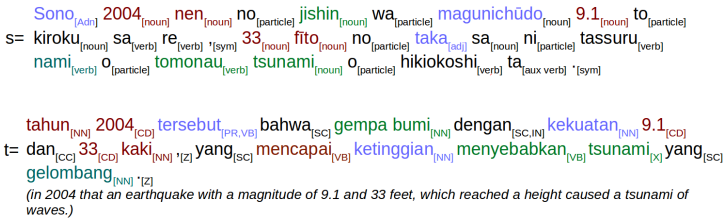

Figure 4.3: Sentence structure for Ja-Id, taken from LM05 F-ISE

iments 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.

In the word reordering multiple pivots, we combine the phrase table by gradually step, i.e., one pivot, two pivots, three pivots, and four pivots, to evaluate the improvement BLUE score that we did not applied in the previous experiment. The result shows that by combining more numbers of pivot languages, then the BLEU score gradually improved. The highest BLEU score of each system, viz., 6.30 of one pivot, 6.92 of two pivots, 6.98 of three pivots, and 7.15 of four pivots. However, we found an interesting result that whether the BLEU scores gradually improved, but the generated text of each system has the same result. As the generated text, the result become comprehensible and understandable compare to the previous experiment, i.e., Without Reordering (WR) experiment.

In the second section, we proposed an initiative approach in phrase table combi-

nation on different symmetrization techniques for Kazakh to English (Kk-En) and Japanese to Indonesian (Ja-Id). The initiative approach arise based on the fact that we find that non-standard symmetrization, i.e., *tgttosrc*, obtained higher BLEU score than the standard one, i.e., *gdfand* in direct translation. Subsequently, we argue that the different results of direct translation in src-trg, src-pvt, and pvt-trg could affect the pivot BLEU score. Thus, we proposed two initiative approach, i.e., First-Interpolation System Experiment (F-ISE) and Second-Interpolation System Experiment (S-ISE). The F-ISE and S-ISE are approaches obtained from the first and second highest BLEU score from the direct experiment.

We find that the F-ISE is a competitive approach since it outperformed the Baseline and Std-ISE of Kk-En by 0.22. However, our F-ISE could not improve the Ja-Id BLEU score due to its phrase translation parameters obtained lower scores than the Baseline and Std-ISE. We also found that the phrase table size does not directly affect the improvement of the BLEU score. We showed that the Kk-En LM05 F-ISE obtained a higher BLEU score, i.e., 3.64, despite its small-phrase table, i.e., 323,850, compared to other systems. This phenomenon also happened in Ja-Id LM05 Baseline that obtained a higher BLEU score, i.e., 12.20, despite its small phrase-table, i.e., 875,038, compared to other systems.

We also evaluate the perplexity score of the ISE system. We find that the longer LM order, i.e., 5-gram, obtained lower perplexity score than the shorter one, i.e., 3-gram, in Kk-En. However, this result differs from the Ja-Id longer LM order that achieved a higher perplexity score than the shorter one. We find that this phenomenon caused by the monolingual target corpus size and dataset training. We observe that the English monolingual target corpus size in Kk-En has a thirteen times bigger dataset, i.e., 114,375, than Indonesian, i.e., 8,500. We argue that with the bigger dataset, the list of n-gram will be much larger. Therefore the English LM probabilities score could be larger compared to the Indonesian. Additionally, we identified that the Ja-Id training process uses the same dataset as the language model, whereas in Kk-En training process uses two different datasets. Therefore, we argue that this parameter, i.e., training dataset, could be a reason why the Ja-Id LM05 obtained a higher perplexity score than the LM03 one.

Last, we also evaluate the generated text from both language pairs. In the short sentence, the F-ISE could produce a compact keyword, compare to Baseline and Std-ISE. However, its translation has the wrong word position. In the long sentence, this wrong word position more severe, particularly for Ja-Id, leading to ambiguous sentences in the target language. We identified that the generated text tends to follow the source language's sentence pattern, i.e., SOV. Meanwhile, the sentence pattern of the target language is SVO.

Chapter 5

Conclusion and Future Work

In this study, we explored the pivot approach on two low-resource languages, i.e., Kazakh to English (Kk-En) and Japanese to Indonesian (Ja-Id). We used two types of pivot approaches, viz., single, and multiple pivots. In the first work, we investigated a single pivot on Kk-En (vice versa) for participating in the Workshop on Machine Translation (WMT) 2019 task. We used Russian as a pivot language due to its similar writing system with Kazakh, i.e., Cyrillic. We examined the effect of different LM orders, viz., 3-gram, and 5-gram, with Linear Interpolation (LI) strategy. We conducted two experiments, viz., direct translation as a baseline system, and interpolation system. Our experimental results show that our interpolation system outperforms the Baseline by 0.5 and 0.1 in Kk-En and En-Kk. In particular, using the 5-gram LM order could obtain a better BLEU score than utilizing the 3-gram one.

We explored the single and multiple pivots for Ja-Id (vice versa) in the second work. We used six ALT datasets, namely Japanese, Indonesian, English, Myanmar, Malay, and Filipino. We used three approaches in the single-pivot, viz., cascade, triangulation, and interpolation. The interpolation comprises of two approaches, viz., linear and fill-up interpolation. We divided the dataset into two data types, namely, sequential and random. The data type showed different effects on language pairs. The Ja-Id of a single pivot could select the sequential or random data type to improve the BLEU score. The results showed that as much as 75% experiments obtained a better BLEU score for both data types. In contrast, the Id-Ja of a single pivot was more suitable for using random data types to improve the BLEU score. The experiment showed that the highest improvement points were achieved by 1.84 for Id-Ja when random data type employed.

In multiple pivots, we employed the interpolation approach by combining four phrase tables. The phrase table combination arranged based on descending and ascending of the BLEU score. Additionally, we also experimented with and without an src-trg phrase table. The results obtained from the experiment show that descend-

ing order more suitable for Ja-Id. In contrast, the ascending order more suitable for Id-Ja. In terms of data type, the Ja-Id and Id-Ja more suitable used random data type. The experiment shows that the highest improvement points achieved by 0.23 and 1.84 for Ja-Id and Id-Ja, respectively, compared to the sequential data type.

In the third work, we applied the *pre-ordering* of the Japanese dataset. Our aim was to overcome the issue from multiple pivots result, i.e., the generated text of Indonesian followed the Japanese sentence structure. Indonesian sentence structure is Subject-Verb-Object (SVO), whereas Japanese sentence structure is Subject-Object-Verb (SOV). Our generated text was not comprehensible, therefore it is hard to understand the sentence. The *pre-ordering* is a stand-alone task to rearrange words in a target-like order before translating. We reordered the Japanese sentence structure from SOV to SVO using Lader. Additionally, we did a phrase table combination gradually, i.e., one pivot, two pivots, three pivots, and four pivots, which we did not apply to the previous experiment. We find that by using more numbers of pivot languages, then the BLEU score gradually improved. However, we find an interesting result that the generated text has a similar sentence in all systems. In this experiment, we proposed an initiative approach on source-pivot (src-pvt) by extending the phrase table. The extending phrase table is a merging phrase table from two symmetrization technique, namely *gdfand* and *tgttosrc*. These techniques arise based on our finding that the *tgttosrc* has candidate phrase pair that could not be obtained by the *gdfand*. As a result, we could generate appropriate generate text as reference.

Last, we compared the phrase table combination between Kk-En and Ja-Id. We proposed an initiative approach in phrase table combination based on different symmetrization techniques. These approaches come based on the fact that the pivot approach comprises three direct translations, viz., src-trg, src-pvt, and pvt-trg, that obtained different BLEU scores when different symmetrization employed. Therefore, we did two phrase table combinations based on the first and second highest BLEU scores of direct translation, viz., F-ISE (First-Interpolation System Experiment), and S-ISE (Second-Interpolation System Experiment). We find that the F-ISE is a competitive approach since it outperformed the Baseline and the Std-ISE (Standard-Interpolation System Experiment) of Kk-En by 0.22. Whereas the Ja-Id still obtained the best BLEU score by Baseline by more than 0.12 compared to our approach. We also evaluate the generated text from both language pairs. In the short sentence, the F-ISE could produce a compact keyword, compare to Baseline and Std-ISE. However, its translation has the wrong word position due to we did not applied the *pre-ordering* as previous experiment. In the long sentence, this wrong word position more severe, particularly for Ja-Id, leading to ambiguous sentences in the target language. We identified that the generated text tends to follow the

source language's sentence pattern, i.e., SOV. Meanwhile, the sentence pattern of the target language is SVO.

In the next works, we will implement our initiative approach, i.e., the extending phrase table in pivot-target (pvt-trg). We expect that the extending phrase table would improve pvt-trg systems, affecting the performance of multiple pivots. We will also increase our Indonesian target monolingual corpus size in Ja-Id as much as Kk-En, i.e., thirteen times bigger. Then we will re-evaluate this parameter on our next works. As F-ISE shows a better result in a single pivot, we also will implement this approach in multiple pivots and compared it with the previous one.

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, 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

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 (Vol-*

ume 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

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

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

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. 848–856. https://www.aclweb.org/anthology/D08-1089/

Isao Goto, Masao Utiyama, and Eiichiro Sumita. 2012. Post-ordering by Parsing for Japanese-English Statistical Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 311–316. https://www.aclweb.org/anthology/P12-2061

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. 2015a. 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

Hieu Hoang, Philip Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *International Workshop on Spoken Language Translation (IWSLT)*. 152–159.

Tam Hoang and Ondřej Bojar. 2015b. TmTriangulate: A Tool for Phrase Table Triangulation. *The Prague Bulletin of Mathematical Linguistics* 104 (2015), 75–86.

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. 2007a. 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

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007b. 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

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, 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. 2007a. 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, 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. 2007b. 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.

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, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 529–533. https://www.aclweb.org/anthology/P11-2093

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

Franz Josef Och. 2003a. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on 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. 2003b. 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 Confer-*

ence 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

Fam Rashel, Andry Luthfi, Arawinda Dinakaramani, and Ruli Manurung. 2014. Building an Indonesian rule-based part-of-speech tagger. In *2014 International Conference on Asian Language Processing, IALP 2014, Kuching, Malaysia, October 20-22, 2014*. 70–73. https://doi.org/10.1109/IALP.2014.6973521

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.

Rico Sennrich. 2012a. 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

Rico Sennrich. 2012b. 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

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

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

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/

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 Japanese and Indonesia Language Pair. Machine Translation Journal. Springer. Submitted on February 2020.

2. Sari Dewi Budiwati, Al Hafiz Akbar Maulana Siagian, Tirana Noor Fatyanosa. Masayoshi Aritsugi. 2020. Investigating Phrase Table Combination based on Standard and Mixed Symmetrization Technique on Low Resource Languages. Draft manuscript.