Multiple Pivots in Statistical Machine Translation for Low Resource Languages

Sari Dewi Budiwati[1,2], **Masayoshi Aritsugi**[3]
[1]Computer Science and Electrical Engineering
Graduate School of Science and Technology, Kumamoto University, Japan
[2]School of Applied Science, Telkom University, Indonesia
[3]Big Data Science and Technology
Faculty of Advanced Science and Technology, Kumamoto University, Japan
saridewi@st.cs.kumamoto-u.ac.jp, aritsugi@cs.kumamoto-u.ac.jp

Abstract

We investigate many combinations of multiple pivots of four phrase tables on a low resource language pair, i.e., Japanese to Indonesia, in phrase-based Statistical Machine Translation. English, Myanmar, Malay, and Filipino from Asian Language Treebank (ALT) were used as pivot languages. A combination of four phrase tables was examined with and without a source to target phrase table. Linear and Fillup Interpolation approaches were employed with two measurement parameters, namely, data types and phrase table orders. The dataset was divided into two data types, i.e., sequential and random. Furthermore, phrase table orders comprise of two, viz., descending and ascending. Experimental results show that the combination of multiple pivots outperformed the baseline. Moreover, the random type significantly improved BLEU scores, with the highest improvement by +0.23 and +1.02 for Japanese to Indonesia (Ja-Id) and Indonesia to Japanese (Id-Ja), respectively. Phrase tables order experiments show a different result for Ja-Id and Id-Ja. The descending order has a higher percentage as much as 87.5% compared to the ascending order in Ja-Id. Meanwhile, the ascending order obtained more than 90% in Id-Ja. Finally, the combination of multiple pivots attempt shows a significant effect to reduce perplexity score in Ja-Id and Id-Ja.

1 Introduction

Statistical Machine Translation (SMT) needs parallel corpora in order to learn translation rules. Parallel corpora are bilingual texts where one of the corpora is an exact translation of the other. Some European languages achieve high-quality translation results with BLEU score more than 40 (Koehn, 2005; Ziemski et al., 2016) by using millions of line parallel corpora and the availability of linguistic tools, e.g., morphological analyzer, POS (part of speech) taggers, and stemmer. Unfortunately, except for Chinese and Japanese, Asian languages have limited parallel corpora with few thousands of line sentences (Riza et al., 2016; Nomoto et al., 2018; Tiedemann, 2012). Moreover, most of the Asian languages still lack linguistic tools and it is thus difficult to achieve the same translation results as European.

With the limited parallel corpora, there are two strategies to achieve high-quality translations, namely building parallel corpora and utilizing existing corpora (Trieu, 2017). Building parallel corpora is difficult since it can be time-consuming and expensive, and needs experts. Therefore, many researchers have focused on utilizing existing corpora, i.e., using pivot approaches (Utiyama and Isahara, 2007; Paul et al., 2009; Habash and Hu, 2009; El Kholy et al., 2013; Dabre et al., 2015; Trieu and Nguyen, 2017; Ahmadnia et al., 2017; Budiwati et al., 2019). 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 preliminary assumption that there are enough parallel corpora between source-pivot and pivot-target languages.

In previous research, English has been the main choice of pivot languages. However Wu and Wang

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 345-355, Hakodate, Japan, September 13-15, 2019

(2007) and Paul et al., (2013) showed that non-English as a pivot language can improve translation quality for certain language pairs. Wu and Wang (2007) showed that using Greek as a pivot language has improved the translation quality compared to English in French to Spanish language pair. Greek as pivot language obtained +5.00 points, meanwhile English obtained +2.00 points. Paul et al., (2013) showed that from 420 experiments language pair in Indo-European and Asian languages, 54.8% is preferable using non-English as the pivot language. Moreover, Wu and Wang (2007) and Dabre et al., (2015) showed promising results by using more than one non-English language. Wu and Wang (2007) showed that using 4 languages, namely Greek, Portuguese, English, and Finnish outperformed the baseline BLEU score with more than +5.00 points. Dabre et al., (2015) also showed that using 7 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 paper, we investigate many combinations of multiple pivots of four phrase tables on low resource language pairs. To make the discussion of this paper concrete, we use Japanese to Indonesia (Ja-Id) and Indonesia to Japanese (Id-Ja) language pairs as an example of them. First, we generate single pivot phrase table by each pivot language, i.e., English, Myanmar, Malay, and Filipino from Asian Language Treebank (ALT). We generate phrase tables by using different approaches, namely Cascade, Triangulation, Linear Interpolation (LI), and Fillup Interpolation (FI). Second, we chose which single pivot approaches have the best result. Last, the combinations of multiple pivots phrase tables were examined with and without a source to target (src-trg) phrase table.

We measured the effect of many combinations of multiple pivots by two parameters, namely data types and phrase table orders. The dataset was divided into two data types, i.e., sequential and random. Sequential type means that the dataset remains unchanged. Meanwhile, random type means the dataset was shuffled before being processed into SMT framework. Furthermore, phrase tables order comprises of two, viz., descending and ascending. Descending order arranges the four phrase tables from highest to lowest according to their BLEU scores. Ascending order is the opposite.

Our contributions are as follows:

- The use of *with and without* src-trg phrase table initiated by the fact that some language pairs have a small parallel corpus, while the others have none. We showed that for the language pair which does not have an src-trg parallel corpus, the translation could be accomplished with multiple pivots and produce high BLEU scores. Furthermore, employing the small src-trg parallel corpus could improve BLEU score more.

- The use of random data type became factors to make better translation results. We showed that the random data type has a significant improvement in translation results. The random data type could be applied in another language pair which has the same characteristics dataset as ALT, i.e., texts originating in English and translated into other languages.

- Phrase table orders can have some effect on perplexity scores. We showed that different phrase tables orders produced different perplexity scores in the experiments of this paper. We thus can say that the phrase tables order should be considered in the multiple pivots.

This paper is organized as follows. Section 2 discusses the availability of parallel corpora and efforts to improve the translation result in Ja-Id language pair. Sections 3 and 4 explain the SMT methodology and pivot approaches. Section 5 describes the experimental setup of many combinations of multiple pivots phrase tables. Section 6 discusses results. Section 7 concludes the paper.

2 Related Work

Current freely available Ja-Id parallel corpora are Asian Language Treebank (ALT) (Riza et al., 2016), TUFS Asian Language Parallel Corpus (TALPCo) (Nomoto et al., 2018), and OPUS (Tiedemann, 2012). ALT is a parallel treebank from English Wikinews to ten languages, i.e., English, Japanese, Indonesia, Khmer, Malay, Myanmar (Burmese), Filipino, Laotian, Thai and Vietnamese. ALT com-

prises of 20,106 sentences annotated with word segmentation, POS tags, and syntax information. The annotation information is limited to Japanese, English, Myanmar and Khmer languages. TALPCo is a parallel corpus of basic vocabulary words and example sentences in five languages, i.e., Japanese, English, Burmese (Myanmar), Indonesian and Malay. TALPCo comprises of 1,372 sentences and only the Burmese (Myanmar) data have been annotated for tokens and parts of speech (POS). OPUS is a collection of translated texts from movies subtitles, localization files (GNOME, Ubuntu, KDE4), Quran translations and a collection of translated sentences from Tatoeba. The parallel corpora of OPUS Ja-Id comprises of 2.9 M sentences from a different domain.

Several approaches have been done in Ja-Id machine translation as shown in Table 2, i.e., pivot languages (Paul et al., 2009), stemmer and removing particles (Simon and Purwarianti, 2013), lemmatization and reordering model (Sulaeman and Purwarianti, 2015), and neural machine translation (Adiputra and Arase, 2017). If we compare these approaches with their BLEU scores in Table 1, Paul et al., (2009) obtained the highest BLEU scores, i.e., 53.13 for Ja-Id and 55.52 for Id-Ja. This result denotes that high-quality translation results can be achieved with enough parallel corpora and certain strategy, e.g., pivot languages.

3 Statistical Machine Translation

Statistical Machine Translation (SMT) is based on a log-linear formulation (Och and Ney, 2002). Let *s* be a source sentence (e.g., Japanese) and *t* be a target sentence (e.g., Indonesia), SMT system outputs the best target translation t_{best} as follows

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

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 $h_m(t|s)$ is a 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 minimum error rate training (Och, 2003).

4 Pivot Methods

Pivot translation is a translation from a source language (SRC) to a target language (TRG) through an intermediate pivot (or bridging) language (PVT) (Paul et al., 2009). Several pivot approaches are sentence translation, triangulation and synthetic corpus.

4.1 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 $h^{\text{t}}_{\text{ij1}}$, $h^{\text{t}}_{\text{ij2}}$, ..., $h^{\text{t}}_{\text{ij8}}$. The situation is as follows:

$$\begin{aligned} SRC\text{-}PVT &= p_{\text{i}}(h^{\text{e}}_{\text{i1}}, h^{\text{e}}_{\text{i2}}, ..., h^{\text{e}}_{\text{i8}}) \\ PVT\text{-}TRG &= t_{\text{ij}}(h^{\text{t}}_{\text{ij1}}, h^{\text{t}}_{\text{ij2}}, ..., h^{\text{t}}_{\text{ij8}}). \end{aligned} \tag{2}$$

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

$$S(t_{\text{ij}}) = \sum_{m=1}^{8} (\lambda^{\text{e}}{}_{\text{m}} h^{\text{e}}{}_{\text{im}} + \lambda^{\text{t}}{}_{\text{m}} h^{\text{t}}{}_{\text{ijm}}) \tag{3}$$

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

$$t_{\text{best}} = \arg\max_{t_{\text{ij}}} S(t_{\text{ij}}). \tag{4}$$

4.2 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, 2016). First, we

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 1: BLEU score comparison of related work.

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

Table 2: Proposed approaches and dataset of the related works.

train two translation models for SRC-PVT and PVT-TRG, respectively. Second, we build an SRC-TRG translation model with **p** as a pivot language.

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

$$\begin{aligned} p(\mathbf{s}|\mathbf{t}) &= \sum_{p}(p(\mathbf{s}|\mathbf{t},\mathbf{p}))p(\mathbf{p}|\mathbf{t}) \\ &\approx \sum_{p}(p(\mathbf{s}|\mathbf{p}))p(\mathbf{p}|\mathbf{t}) \end{aligned} \tag{5}$$

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

The triangulation translation model is often combined with SRC-TRG translation model, called phrase table combination. There are 3 ways to combine triangulation with SRC-TRG translation model, namely Linear Interpolation (LI), Fillup Interpolation (FI), and Multiple Decoding Paths (MDP). Linear Interpolation 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. Multiple Decoding Paths (MDP) method of Moses which uses all the tables simultaneously while decoding ensures that each pivot table is kept separate and translation options are collected from all the tables (Dabre et al., 2015).

More than one pivot language can be used to improve the quality of the translation performance, this is 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):

$$P(s|t) = \sum_{i=1}^{n} \alpha_{\mathrm{i}} P_{\mathrm{i}}(s|t) \tag{6}$$

$$P(s|t,\alpha) = \sum_{i=1}^{n} \beta_{\mathrm{i}} P_{\mathrm{i}}(s|t,\alpha) \tag{7}$$

where $P(s|t)$ and $P(s|t,\ \alpha)$ are the phrase translation probability and the lexical weight trained with SRC-TRG corpus estimated by using pivot language, while α_{i} and β_{i} are interpolation coefficients. Last, $\sum_{i=1}^{n} \alpha_{\mathrm{i}} = 1$, and $\sum_{i=1}^{n} \beta_{\mathrm{i}} = 1$.

5 Description of Languages, Dataset Scenarios and Experiments

In this section, we first describe the characteristics of pivot languages. Further, we explain how dataset is divided and used.

5.1 Languages involved

We use six datasets from ALT, i.e., Japanese, Indonesia, English, Myanmar, Malay and Filipino. Japanese and Indonesia datasets were used to build the direct translation as Baseline model. The Japanese language is an SOV language, while Indonesia is SVO language. Therefore, we chose pivot languages based on the similarity of a word order with Japanese or Indonesia. English and Malay have the same word order as Indonesia. Meanwhile, Myanmar has the same word order as Japanese. Filipino was chosen to evaluate the effect of VOS language. The word order and languages family can be seen in Table 3.

5.2 Dataset scenario

We divide the dataset into two data types, namely sequential (seq) and random (rnd). Sequential type

Languages	Word of order	Language Family
Japanese	SOV	Japonic
Indonesia	SVO	Austronesian
English	SVO	Indo-European
Myanmar	SOV	Sino-Tibetan
Malay	SVO	Austronesian
Filipino	VOS	Austronesian

Table 3: Language characteristics.

means that the dataset remains unchanged. Meanwhile, random type means the dataset was shuffled before used in 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., 4 Baselines, 32 SRC-PVT and PVT-TRG, 64 single pivots, and 32 multiple pivots.

5.3 Experimental setup

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

5.3.1 Single pivot

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 SRC-PVT and PVT-TRG systems, where the first system translates the source language input into the pivot language and the second system takes the translation result as input and translates into the target language. As a result, we construct 16 SRC-PVT and 16 PVT-TRG systems.

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

- Pruning the SRC-PVT and PVT-TRG phrase table from the Cascade experiments using *filter-pt* (Johnson et al., 2007). The pruning activity intended to minimize the noise of SRC-PVT and PVT-TRG phrase table.
- Merging two pruning phrase tables using *Tm-Triangulate* (Hoang and Bojar, 2015). The result is denoted as `TmTriangulate` phrase table.

In the Linear Interpolation approach, we combine `TmTriangulate` and SRC-TRG phrase table with *dev* phrase table as a reference. The result is called `TmCombine` phrase table. In Fillup interpolation, 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.

5.3.2 Multiple pivots

In multiple pivots, first, we observe BLEU scores result from each approach in a single pivot. Then, we employ phrase tables from the best pivot approaches into the next step, i.e., the combination of multiple pivots. As shown in Figure 1 and Figure 2, the Linear and Fillup Interpolation approaches have higher BLEU scores compared to Baseline. Therefore, we use the four phrase tables from Linear and Fillup Interpolation 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 the single pivot BLEU score, viz., descending and ascending orders. Descending order sorts the four phrase tables from highest to lowest according to their BLEU scores. Ascending order is the opposite. For example, the BLEU scores of Linear Interpolation 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 SRC-TRG phrase table using Linear Interpolation approach. The result is denoted as All-LinearInterpolate `All-LI`.

- Merging of four phrase tables without SRC-TRG phrase table using Fillup Interpolation approach. The result is denoted as All-FillupInterpolation `All-FI`.

- Combining `All-LI` with SRC-TRG phrase table using Linear Interpolation approach. The result is denoted as `Base-LI`.

- Combining `All-FI` with SRC-TRG phrase table using Fillup Interpolation approach. The result is denoted as `Base-FI`.

6 Result and Discussion

In this section, we will discuss results based on BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and perplexity scores. BLEU score is a metric for evaluating the generated sentence compared to the reference sentence. High BLEU scores indicate a better system. Perplexity score is frequently used as a quality measure for language models (Sennrich, 2012). Lower perplexity scores indicate that the language model is better compared to higher perplexity score. We used the query from KenLM (Heafield, 2011) 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.

6.1 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.

Baseline BLEU scores are given in column 2 of Table 4 and Table 5 for Ja-Id and Id-Ja, respectively. 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. Meanwhile, the BLEU score of Baseline Random Id-Ja is 12.00, +1.00 points higher compared to Baseline Sequential.

Baseline perplexity scores are given in Figure 3 and Figure 4 for Ja-Id and Id-Ja, respectively. As shown in the figures, the Ja-Id and Id-Ja perplexity scores of Random data type obtained higher point compared to the Sequential data type. Perplexity score of Ja-Id in Random data type has 384.59, while Sequential data type has 291.51. Furthermore, perplexity score of Id-Ja in Random data type has 81.58, while Sequential data type has 71.94.

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

6.2 Multiple pivots translation results

6.2.1 Single pivot results

The Triangulation approach was the worst approach in Ja-Id and Id-Ja. All the results of Triangulation have smaller BLEU score compared to Baseline. The Cascade approach also has lower scores compared to 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 Linear and Fillup Interpolation. 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 Figure 1 and Figure 2 for Ja-Id and Id-Ja, respectively.

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, 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 and Figure 4. We showed how to reduce the perplexity scores by using multiple pivots in the next section.

6.2.2 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 5.3.

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 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 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 5 and Figure 6 for Ja-Id and Id-Ja, respectively.

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

In terms of perplexity scores for Ja-Id, All-LI and All-FI show poor results. However, the perplexity scores could be reduced in Random data type of Base-LI and Base-FI. Both approaches use SRC-TRG phrase table in the combination process. The results show that the SRC-TRG phrase table has a significant impact on reducing the perplexity score. Meanwhile, the perplexity scores in Id-Ja could be reduced without using the SRC-TRG phrase table. Moreover, the Base-LI and Base-FI results have lower perplexity scores compared to All-LI and All-FI. We show the perplexity scores in Figure 7 and Figure 8 for Ja-Id and Id-Ja, respectively.

We summarize the results of single and multiple pivots in Table 4 and Table 5. We show BLEU scores of best approaches in Figure 9 and Figure 10, and the perplexity scores of best approaches in Figure 11 and Figure 12.

7 Conclusion and Future Work

In this paper, 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 Inter-

Data Type	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulation	LI	FI	Desc	Asc
Sequential	11.96	12.01 (MS)	9.71 (EN)	12.21 (MY)	12.27 (MY)	12.23 (Base-LI)	12.37 (Base-FI)
Random	12.17	11.81 (MS)	9.62 (FI)	12.22 (MS)	12.29 (MY)	12.40 (Base-LI)	12.27 (All-FI)

Table 4: Best BLEU score in baseline, single and multiple pivots for Japanese to Indonesia

Data Type	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulation	LI	FI	Desc	Asc
Sequential	11.00	12.18 (MS)	8.26 (EN)	12.03 (MY)	12.40 (MY)	12.15 (Base-LI)	12.84 (ALL-FI)
Random	12.00	11.13 (MS)	9.17 (MS)	12.84 (MY)	12.88 (MY)	12.74 (All-FI)	13.02 (ALL-FI)

Table 5: Best BLEU score in baseline, single and multiple pivots for Indonesia to Japanese

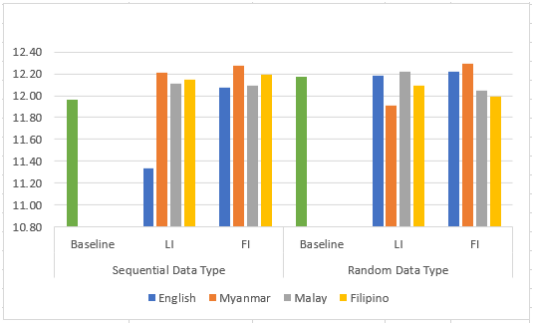

Figure 1: Single pivot BLEU scores of Ja-Id for LI and FI approaches.

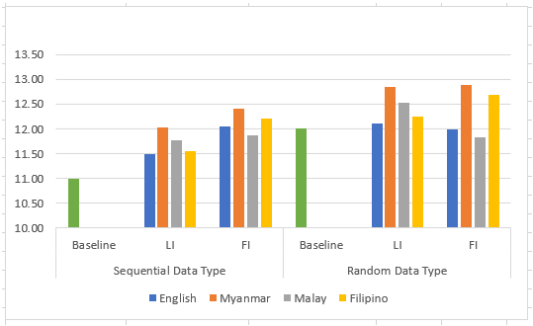

Figure 2: Single pivot BLEU scores of Id-Ja for LI and FI approaches.

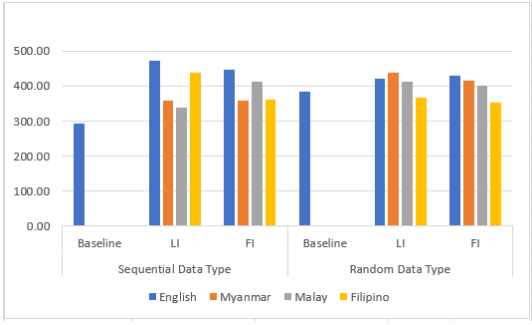

Figure 3: Perplexity Score of Ja-Id single pivot for LI and FI approaches.

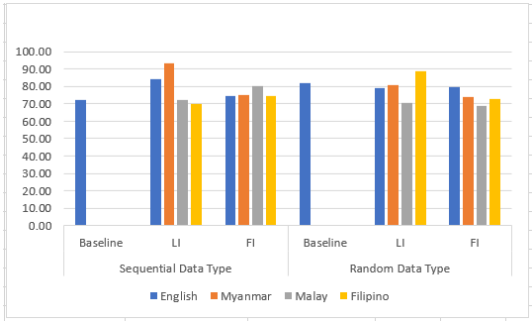

Figure 4: Perplexity Score of Id-Ja single pivot for LI and FI approaches.

polation (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 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.,

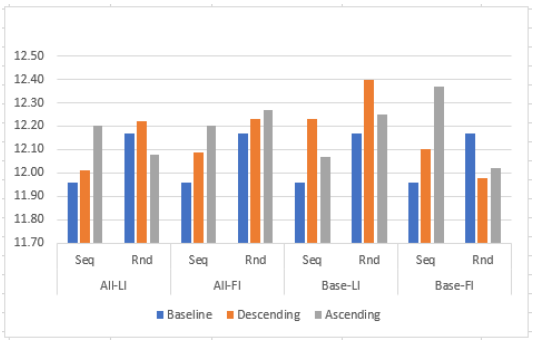

Figure 5: BLEU score for Ja-Id in multiple pivots.

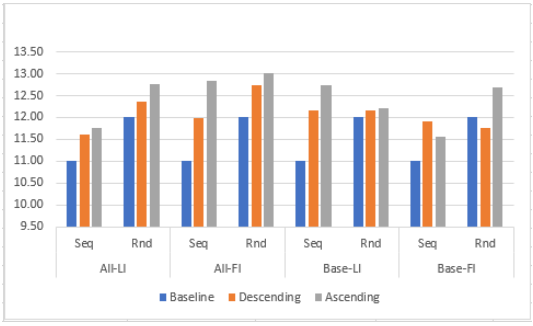

Figure 6: BLEU score for Id-Ja in multiple pivots.

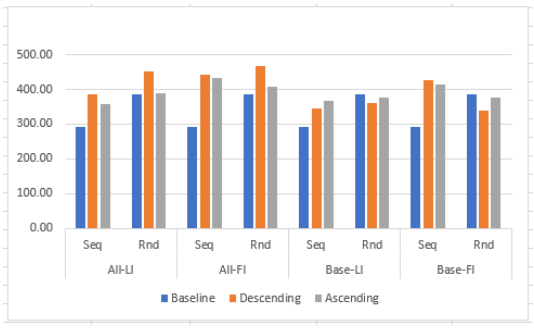

Figure 7: Perplexity score for Ja-Id in multiple pivots.

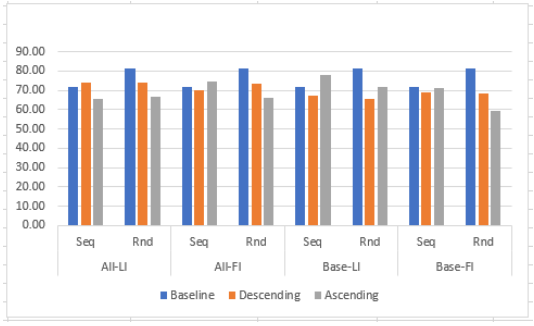

Figure 8: Perplexity score for Id-Ja in multiple pivots.

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.

In the experiments, we did not show the combinations of two or three phrase tables as in (Wu and Wang, 2007). This will be included in our future work to give a better explanation on whether the combinations of two or three phrase tables will give better improvement compared to four phrase tables. Furthermore, the combination of the best phrase tables from each data type should be taken into account for next future research.

References

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*. 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*, pages 24–30.

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*, pages 340–345, Florence, Italy. Association for Computational Linguistics.

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*, pages 1192–1202. Association for Computational Linguistics.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language Independent Connectivity Strength Features for Phrase

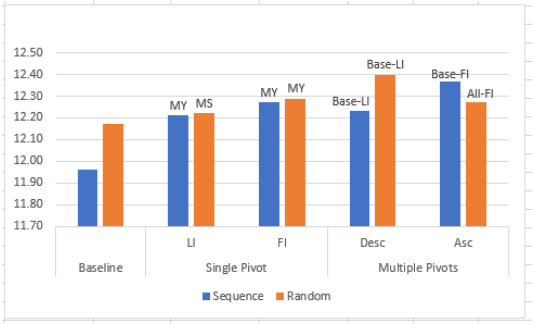

Figure 9: BLEU scores of Ja-Id best approach.

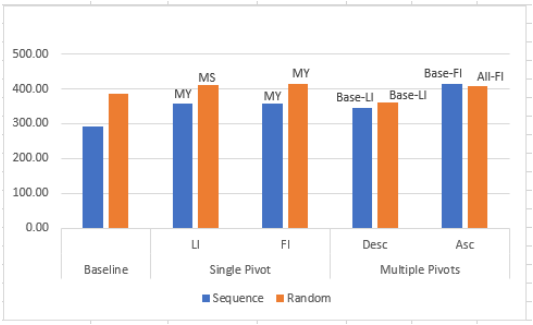

Figure 11: Perplexity scores of Ja-Id best approach.

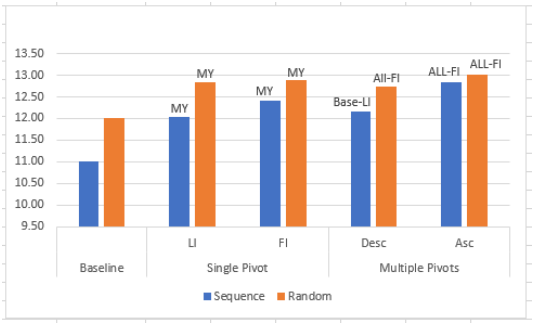

Figure 10: BLEU scores of Id-Ja best approach.

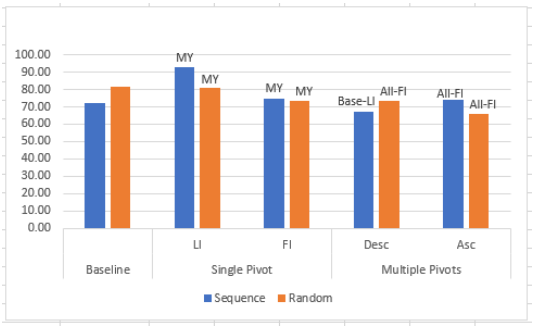

Figure 12: Perplexity scores of Id-Ja best approach.

Pivot Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418. Association for Computational Linguistics.

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*, StatMT '09, pages 173–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*.

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

Duc Tam Hoang and Ondrej Bojar. 2016. Pivoting Methods and Data for Czech-Vietnamese Translation via English. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 190–202.

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

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

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

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

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics*, ACL '02, pages 295–302, Strouds-

burg, PA, USA. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the Importance of Pivot Language Selection for Statistical Machine Translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 221–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

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):14:1–14:17, October.

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)*, pages 1–6, Oct.

Rico Sennrich. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

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)*, pages 80–84, Dec.

M. A. Sulaeman and A. Purwarianti. 2015. Development of Indonesian-Japanese Statistical Machine Translation Using Lemma Translation and Additional Postprocess. In *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 54–58, Aug.

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

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*.

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

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*, pages 484–491. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-based Statistical Machine Translation. *Machine Translation*, 21(3):165–181, September.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1. 0. European Language Resources Association (ELRA), May.