Phrase Table Combination Based on Symmetrization of Word Alignment for Low-Resource Languages

Sari Dewi Budiwati 1,2,*, Al Hafiz Akbar Maulana Siagian 1,3, Tirana Noor Fatyanosa 1 and Masayoshi Aritsugi 4

Abstract: Phrase table combination in pivot approaches can be an effective method to deal with low-resource language pairs. The common practice to generate phrase tables in pivot approaches is to use standard symmetrization, i.e., grow-diag-final-and. Although some researchers found that the use of non-standard symmetrization could improve bilingual evaluation understudy (BLEU) scores, the use of non-standard symmetrization has not been commonly employed in pivot approaches. In this study, we propose a strategy that uses the non-standard symmetrization of word alignment in phrase table combination. The appropriate symmetrization is selected based on the highest BLEU scores in each direct translation of source–target, source–pivot, and pivot–target of Kazakh–English (Kk–En) and Japanese–Indonesian (Ja–Id). Our experiments show that our proposed strategy outperforms the direct translation in Kk–En with absolute improvements of 0.35 (a 11.3% relative improvement) and 0.22 (a 6.4% relative improvement) BLEU points for 3-gram and 5-gram, respectively. The proposed strategy shows an absolute gain of up to 0.11 (a 0.9% relative improvement) BLEU points compared to direct translation for 3-gram in Ja–Id. Our proposed strategy using a small phrase table obtains better BLEU scores than a strategy using a large phrase table. The size of the target monolingual and feature function weight of the language model (LM) could reduce perplexity scores.

Keywords: symmetrization of word alignment; phrase table combination; pivot approach; low-resource languages

1. Introduction

Low-resource languages suffer from data scarcity, which leads to poor translation quality. One of the common techniques to improve translation quality is using a phrase table combination in pivot approaches [1–6]. The phrase table or translation model comes from a word alignment model that uses a symmetrization technique to generate phrase pairs. The standard symmetrization for word alignment model is grow-diag-final-and (gdfand) [7]. Although prior studies have shown that non-standard symmetrization, i.e., intersection, could obtain higher bilingual evaluation understudy (BLEU) scores than the standard one [8–10], non-standard symmetrization has not been commonly used in pivot approaches. Thus, the appropriate symmetrization of word alignment model needs to be investigated to improve the performance of low-resource languages when using phrase table combination in pivot approaches.

Kholy and Habash [11] studied phrase table combinations based on the symmetrization of word alignment model in pivot approaches for the Hebrew–Arabic language pair. Their proposed approach was based on symmetrization relaxation, which extracted new
phrase pairs by removing a given word that was unlisted in the pivot phrase tables. They constructed two new symmetrizations, union_Relaxation (U_R) and grow-diag-final-Relaxation (GDFA_R), based on union and gdfand, respectively. The BLEU score from the combination of the GDFA and GDFA_R symmetrizations was 0.8 higher than that of the GDFA_R, which demonstrated the superiority of the phrase table combination.

Unlike Kholy and Habash [11], in this study, we propose a strategy, i.e., phrase table combination, that uses symmetrization of word alignment, which obtains the highest BLEU scores. Our strategy is based on previous research results that showed that symmetrization of word alignment is language-specific [8–10] and dataset-specific [12], as different language pairs with different datasets have different BLEU scores. In contrast to Kholy and Habash [11], who removed a given word, our strategy employs available parallel corpora of source-target (src–trg), source-pivot (src–pvt), and pivot-target (pvt–trg) without removing the phrase pair that could be a potential candidate in the translation process.

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

- We explore two language model (LM) orders, 3-gram and 5-gram, and five symmetrizations, gdfand, intersection, union, srctotgt, and tgttosrc, in a direct system approach (DSA). We find that 5-gram is a better choice for Kk–En and Ja–Id. We also find that non-standard symmetrization, i.e., tgttosrc, could improve the BLEU score compared to the standard one, i.e., gdfand.
- We propose a strategy in an interpolation system approach (ISA), named highest-ISA (H-ISA). The H-ISA uses the symmetrization of the word alignment that obtained the highest BLEU scores from the DSA for each phrase table of src–trg, src–pvt, and pvt–trg. We find that the H-ISA could be a competitive approach because it outperforms direct translation in Kk–En with absolute improvements of 0.35 and 0.22 BLEU points for 3-gram and 5-gram, respectively. Our proposed strategy also outperforms the direct translation of 3-gram in Ja–Id with an absolute improvement of 0.11 BLEU point.
- We found that using a smaller phrase table could also obtain higher BLEU scores than a larger phrase table because of higher phrase translation parameter scores, particularly in direct phrase translation probability ($p(s|t)$) and inverse lexical weight ($\text{lex}(t!s)$). Our findings contradict those of previous studies that used larger phrase tables to obtain higher BLEU scores [11].

The rest of the paper is organized as follows: Section 2 reviews related work on the phrase table combination and current research on the low-resource languages used in this work, i.e., Kk–En and Ja–Id. Section 3 explains our proposed strategy. Section 4 describes results and provides a discussion. Section 5 concludes with a summary and discusses future work.

2. Related Work

2.1. Phrase Table Combination

Phrase table combination has been used to improve the translation quality in direct translation and pivot approaches. In direct translation, researchers have combined two or more phrase tables of direct translation with various symmetrizations of word alignment [9,12]. Wu and Wang [12] used two approaches in their direct translation of Spanish–English: the direct translation of single symmetrization and phrase table combination. The first approach used three symmetrizations: gdfand, grow-diag, and intersection. The second approach combined three phrase tables produced by gdfand, grow-diag, and intersection. Their experimental results showed that phrase table combination outperformed
the direct translation of a single symmetrization approach. Singh [9] explored massive experiments, also arriving at the same conclusion. They combined three to eight phrase tables of direct translation in the Chinese–English language pair. The phrase tables were produced by several symmetrizations of word alignment: gdfand, intersection, union, grow-diag-final, grow-diag, grow, src tgt, tgt src, and grow-final. Their experimental results showed that the phrase table combination of several symmetrizations outperformed the Baseline.

Phrase table combination could also be used in pivot approaches to overcome the data scarcity between src–trg language pairs [1–4,6,13,14]. The phrase table combination outperformed direct translation or other pivot approaches, namely, sentence translation and triangulation. We identified that the phrase table combination in the pivot approaches uses the same symmetrization of word alignment, i.e., gdfand [1–4,6,13,14]. This is because the gdfand is a standard symmetrization of word alignment in statistical machine translation (SMT) [7]. Then, Kholy and Habash [11] proposed the phrase table combination that combined two symmetrization of word alignments: grow-diag-final-and (G DFA) and grow-diag-final-and_Relaxation (G DFA_R), for the Hebrew–Arabic language pair. G DFA is a symmetrization of word alignment that starts with an intersection of two alignments. Then, the symmetrization adds alignment points between two unaligned words [8]. G DFA is a common symmetrization of word alignment used in word alignment tools such as GIZA++.

GDF_R was constructed by Kholy and Habash [11] using the following steps:

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

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

From those studies, the application of the phrase table combination could improve the translation quality. Nevertheless, the symmetrization of word alignment can vary across language pairs and datasets (language- and dataset-specific). Koehn et al. [8] explored five symmetrizations (final, final-and, grow-diag, grow, and intersection) in five language pairs (Arabic–English, Japanese–English, Korean–English, Chinese–English, and English–Chinese). Given an example of Japanese–English, they obtained the highest BLEU score of 45.1 when using intersection, and the lowest BLEU score of 39.0 when using grow-diag. Stymne et al. [10] explored five symmetrizations (intersection, grow-diag, grow-diag-final-and, grow-diag-final, and union) in German–English. Their experimental result showed that the highest BLEU score was obtained by grow-diag with a score of 20.9, and the lowest BLEU score was obtained by intersection with a score of 19.1. Those investigations showed that the best symmetrization of word alignment can differ for each language pair. This conclusion confirmed the language-specific characteristic of the symmetrization of word alignment.

Symmetrization of word alignment is also dataset-specific because the same language pairs with different datasets can obtain different BLEU scores [9,12]. Wu and Wang [12] explored two types of test sets: in-domain and out-of-domain. The exploration was conducted using six symmetrizations: grow-diag-final, grow-final, union, grow-diag, grow, and intersection. In-domain means the test set type was the same domain as the training dataset. In contrast, out-of-domain uses a different type of test set from the training dataset. Their experiment was implemented in three language pairs: Spanish–English, French–English, and Dutch–English. In an example of Spanish–English, a higher BLEU score of 30.63 was obtained using grow-final in the in-domain test set. Using the same symmetrization, Spanish–English obtained a lower BLEU score of 25.00 in the out-of-domain test set.
Their results suggested that the symmetrization of word alignment can differ for each type of test set (dataset-specific). Singh [9] also suggested the same conclusion. They explored six datasets: NIST2002, NIST2003, NIST2004, NIST2005, NIST2006, and NIST2008. The test was conducted using nine symmetrizations (grow-diag-final-and, intersection, union, grow-diag-final, grow-diag, grow, src2tgt, tgt2src, and grow-final) in the Chinese–English language pair. Given an example of a baseline system that used the same symmetrization, i.e., grow-diag-final-and, in two datasets (NIST2002 and NIST2005), the BLEU scores obtained for both datasets were 31.56 and 25.82. Their results confirmed the dataset-specific characteristic, in agreement with Wu and Wang [12].

The symmetrization of word alignment can be applied either in direct translation [9,12] or pivot approaches [1–4,6,13,14]. No studies have compared the different strategies in direct translation and pivot approaches. Considering this gap, we applied a phrase table combination that uses different symmetrizations of word alignment in pivot approaches based on the highest BLEU scores. Our consideration was based on [8–10,12], which stated that the symmetrization of word alignment is language-specific and dataset-specific as different language pairs with different datasets have different BLEU scores. Lastly, our strategy employs the available parallel corpora of src–trg, src–pvt, and pvt–trg without removing the phrase pair that could be a potential candidate in the translation process.

2.2. Kk–En and Ja–Id as Low-Resource Language Pairs

Kk–En and Ja–Id are considered low-resource language pairs due to the scarcity of available parallel corpora. The available parallel corpora of Kk–En are open source parallel corpus (OPUS) [15] and news-commentary, which have 953,240 parallel sentences in total. Similarly, the available parallel corpora of Ja–Id are Asian language treebank (ALT) [16], TUFs Asian language parallel corpus (TALPCo) [17], and OPUS [15], which contain 1,468,155 parallel sentences in total. Due to this limitation, we propose an alternative approach to deal with the low-resource language pairs Kk–En and Ja–Id.

The morphological segmentation approach produced a progressive improvement in the translation quality of Kk–En in the SMT model [18,19]. The morphological segmentation is an approach that breaks words into morphemes. The approach was implemented because Kazakh is considered an agglutinative and highly inflected language. Figure 1 shows Kazakh words without morphological segmentation. Word “дос/friend” was joined by various suffixes and corresponded to English phrases of various lengths [18]. Assylbekov and Nurkas [18] showed the morphological segmentation approach could obtain an absolute gain of up to 1.05 BLEU points in the 636 K dataset. Kartbayev [19] also showed an absolute improvement of 1.43 BLEU points when using the morphological segmentation approach in 60 K dataset. Both research results [18,19] showed that the morphological segmentation is an important consideration in the SMT model for Kazakh.

![Figure 1](image.png)

**Figure 1.** Example of Kazakh suffixation. The left side is Kazakh words with various suffixes. The right side is English translation with various phrase lengths [18].

Kk–En machine translation was introduced as a shared task for low-resource language pairs in the Workshop on Machine Translation (WMT) 2019 [20]. Most participants used the NMT model with several approaches: back translation, transfer learning, multilingual transfer learning, and sequence-2-sequence. The transfer learning approach is a similar technique to the pivot approaches in the SMT model. Transfer learning uses a high-resource language pair to train the parent model, and then the parent training data are replaced with the training data of low-resource language pairs [21]. We identified three submission
systems that use transfer learning: Charles University (CUNI) [21], National Institute of Information and Communications Technology (NICT) [22], and University of Maryland (UMD) systems [23]. The NICT and CUNI systems use Russian–English as the parent model and obtained BLEU scores of 26.2 and 18.5, respectively. In comparison with both systems, the UMD system uses Turkish–English as the parent model and obtained a BLEU score of 9.2. Their experimental results [21–23] showed that the third language was still needed to improve the Kk–En translation quality.

Unlike Kk–En, the SMT model still outperformed NMT with an absolute improvement of 3.93 BLEU points in Ja–Id [24]. Moreover, the pivot approaches of the SMT model obtained a higher BLEU score than direct translation in Ja–Id [4, 25]. Paul et al. [25] showed that the single pivot approach produced an absolute gain of up to 0.23 BLEU points compared to direct translation. Budiwati and Aritsugi [4] showed that single and multiple pivot approaches obtained absolute improvements of 0.31 and 0.41 BLEU points, respectively, compared to direct translation. In this study, we focused on phrase table combination in the single pivot approach of the SMT model.

In addition to the model and techniques, Simon and Purwarianti [26] and Sulaeman and Purwarianti [27] found several morphological issues in Ja–Id: word order problems, incorrectly defined phrases, and words with affixes. In particular, Simon and Purwarianti [26] proposed several techniques: using pos-tag, increasing the LM dataset, stemming for the Indonesian dataset, removing Japanese particles for the Japanese dataset, and removing the named entity (NE). The BLEU score with the Japanese particles removal outperformed the baseline by 0.02. Then, Sulaeman and Purwarianti [27] proposed several techniques: the pos-tag model, the hierarchical model, lemmatizer, and post-processing. The pos-tag model and lemmatizer outperformed the baseline by 0.1 BLEU points. However, Simon and Purwarianti [26] and Sulaeman and Purwarianti [27] did not show their proposed generated text results. Consequently, it is hard to compare the pre- and post-proposed generated texts. In this paper, we show the generated texts to investigate whether morphological issues are still encountered when implementing the pivot approaches.

3. An Interpolation System Approach

This section provides information about the direct and interpolation approaches, the datasets and pre-processing, and an explanation about our proposed approach, that is, highest-interpolation system approach (H-ISA). In this study, we use the statistical machine translation (SMT) model as it outperformed the neural machine translation (NMT) for the Ja–Id language pair [24]. We used the same model, i.e., SMT model, for Kk–En to compare and investigate the obtained results between the two low-resource language pairs.

3.1. Direct and Pivot Translation

Direct translation uses an src–trg system to translate source to target sentences. Let \( s \) be a source sentence and \( t \) be a target sentence. Then, the SMT system outputs the best target translation \( t_{\text{best}} \) as follows:

\[
t_{\text{best}} = \arg \max \limits_{t} p(t|s) = \arg \max \limits_{t} \sum_{m=1}^{M} \lambda_{m} h_{m}(t|s)
\]

(1)

where \( h_{m}(t|s) \) represents feature function, and \( \lambda_{m} \) is the weight assigned to the corresponding feature function [1]. The feature function \( h_{m}(t|s) \) is the language model probability of the 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 \( (\lambda_{m}) \) can be set by minimum error rate training (MERT) [28].

Pivot translation, or pivot approaches, is a translation from a source to target using a bridge language [25]. Pivot approaches arise from the preliminary assumption that src–pvt and pvt–trg have enough parallel corpora to overcome the data scarcity of src–trg.
Triangulation is one of the strategies of pivot approaches that combines src–pvt and pvt–trg translation models. The result of triangulation strategies, i.e., the triangulation translation model, is often combined with src–trg translation models, called interpolation. There are three ways to combine triangulation with src–trg translation models: linear interpolation, fillup interpolation, and multiple decoding paths [2]. Linear interpolation merges two translation models and computes a weighted sum of phrase pair probabilities from each phrase table. Fillup interpolation does not modify phrase probabilities but selects phrase pair entries from the next phrase table if they are not present in the current phrase table. Multiple decoding paths simultaneously use all the phrase tables while the decoder ensures that each pivot table is kept separate and translation options are collected from all the phrase tables.

In this study, we used linear interpolation strategy with perplexity minimization to merge the src–trg with triangulation translation models [29]. Given n phrase tables, we are looking for a set of n weights \( \lambda_1, \ldots, \lambda_n \), such that \( \sum_{m=1}^{n} \lambda_i = 1 \), where \( \lambda_i \) is the interpolation weight of phrase table \( i \). Given a phrase pair \((x'|y)\), the linear interpolation of the n models is [30]:

\[
p(x'|y; \lambda) = \sum_{m=1}^{n} \lambda_i p_i(x'|y)
\]  

We used combine_given_tuning_set, which writes a new phrase table using the weights to minimize cross-entropy on a tuning (dev) set.

### 3.2. Datasets and Pre-Processing

In this study, we used the news-commentary [20] and ALT datasets [16] for Kazakh–English (Kk–En) and Japanese–Indonesian (Ja–Id), respectively. We considered the two news domain datasets because of the use of formal language. Some languages have flexible word order, such as Russian [31]. Flexible word order means sentences can use free word order of subject, verb, and object. Since a news domain uses formal language, we expected that the sentences would use one dominant word order. Thus, we could analyze the effects of different word orders in the pivot approaches. A news domain uses relatively long sentences, i.e., about more than 22 words on average, as shown in Tables 1 and 2. We expected that longer sentences would produce various phrase pairs from the symmetrization of word alignment. Additionally, news-commentary and ALT were used as training datasets for low-resource translation tasks in the Workshop on Machine Translation (WMT) 2019 and Workshop on Asian Translation (WAT) 2019.

We performed several pre-processing steps, i.e., tokenizing, normalizing punctuation, re-casing, and filtering sentences for both datasets. The tokenization step separates words and punctuation. We employed Moses [32] to tokenize the Kk, En, and Id, and MeCab [16] was used to tokenize Ja. We then normalized the punctuation so that the decoder system could recognize it. Next, the re-casing step reduced the data sparsity by converting the initial word in each sentence to its most probable casing. Lastly, we removed sentences that had a length of more than 80 words in the filtering step. We show the dataset statistics for Kk–En and Ja–Id in Tables 1 and 2, respectively.
Table 1. Dataset statistics of Kazakh–English (Kk–En).

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

Table 2. Dataset statistics of Japanese–Indonesian (Ja–Id).

| Dataset | #Sentences | Average Sentence Length | Vocabulary |
|---------|------------|-------------------------|------------|
| **Train** |            |                         |            |
| DataALT.01.jp-id.SP.true.jp | 8500 | 34.90 | 19,086 |
| DataALT.01.jp-id.SP.true.id | 8500 | 24.75 | 28,014 |
| **Dev** |            |                         |            |
| DataALT.02.jp-id.true.jp | 2000 | 33.88 | 6680 |
| DataALT.02.jp-id.true.id | 2000 | 24.26 | 10,201 |
| **Language Model** |            |                         |            |
| DataALT.01.jp-id.SP.true.id | 8500 | 24.75 | - |
| **Interpolation system** |            |                         |            |
| DataALT.01.jp-id.SP.true.jp | 8500 | 34.90 | 19,086 |
| DataALT.01.jp-id.SP.true.ms | 8500 | 25.11 | 26,835 |
| DataALT.01.jp-id.PT.true.ms | 8500 | 25.04 | 26,922 |
| DataALT.01.jp-id.PT.true.id | 8500 | 25.04 | 28,361 |
| **Dev** |            |                         |            |
| DataALT.02.jp-id.true.jp | 2000 | 33.88 | 6680 |
| DataALT.02.jp-id.true.ms | 2000 | 24.51 | 9888 |
| DataALT.02.jp-id.true.ms | 2000 | 24.51 | 9888 |
| DataALT.02.jp-id.true.id | 2000 | 24.26 | 10,201 |
| **Language Model** |            |                         |            |
| DataALT.01.jp-id.SP.true.ms | 8500 | 25.11 | - |
| DataALT.01.jp-id.SP.true.id | 8500 | 24.75 | - |
3.3. Highest-Interpolation System Approach (H-ISA)

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

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

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

- We prune the src–pvt and pvt–trg phrase tables using filter-pt [33].
- We then merge the two pruned phrase tables using the triangulation method [34]. We modified the triangulation method [34] for our H-ISA. Given the conditional probabilities of pruned src–pvt \( p(\bar{s}|\bar{p})^* \) and pvt–trg \( p(\bar{p}|\bar{t})^* \) phrase tables, we merge the two phrase tables as shown in Equation (3). The two pruned phrase tables were obtained from the DSA of src–pvt and pvt–trg with the highest BLEU scores.

\[
p(\bar{s}|\bar{t})^* = \sum_{\bar{p}} p(\bar{s}|\bar{p},\bar{t})^* p(\bar{p}|\bar{t})^* \\
\approx \sum_{\bar{p}} p(\bar{s}|\bar{p})^* p(\bar{p}|\bar{t})^* \tag{3}
\]

- We merge the triangulation phrase table with the phrase table of src–trg according to Equation (2). We modified Equation (2) for our H-ISA. Given the conditional probabilities of src–trg phrase table \( p(s|t)^* \) and triangulation phrase table \( p(\bar{s}|\bar{t})^* \), we merge the two phrase tables as follows:

\[
p(s|t;\lambda) = \lambda_{\text{src-trg}} p(s|t)^* + \lambda_{\text{pvt}} p(\bar{s}|\bar{t})^* \tag{4}
\]

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

We use 3-gram and 5-gram orders as our language model in this ISA, as well as in the DSA. We chose 3-gram as the minimum order because it can predict a better probability of the next word than other lower \( n \)-grams, i.e., 2-gram. We used 5-gram as the maximum order because Liu et al. [35] showed that a system using more than 5-gram has relatively the
same BLEU score as 5-gram. Therefore, we used the two LM orders to identify which order produces better translation quality and perplexity scores in our low-resource language pairs.

We used minimum error rate training (MERT) [28] as a tuning algorithm in our experiments. We used bilingual evaluation understudy (BLEU) [36] and perplexity scores to measure the performance of our work in this study. The BLEU score is a metric used for evaluating the generated sentence compared to the reference sentence. The perplexity score is a metric that defines the performance of a language model order. We measured the perplexity score based on the LM order against the \text{eval} of the generated sentence. Ideally, a translation system that obtains a higher BLEU score has a lower perplexity score [35].

\textbf{Algorithm 1:} Highest-interpolation system approach (H-ISA).

\begin{verbatim}
Input:
Corpora of src, pvt, and trg
$\lambda_{src-trg, \lambda_{pvt}}$
Output: $p(s|t)$ of H-ISA

1 $symms \leftarrow \{gdfand, intersection, union, srctotgt, tgttosrc\}$
2 foreach $sym$ in $symms$ do
3 $SrcTrg$_Bleu$[sym] \leftarrow$ run Equation (1) with $sym$ and corpora of src-trg
4 $SrcPvt$_Bleu$[sym] \leftarrow$ run Equation (1) with $sym$ and corpora of src-pvt
5 $PvtTrg$_Bleu$[sym] \leftarrow$ run Equation (1) with $sym$ and corpora of pvt-trg
6 end
7 $SrcTrg$_Symm $\leftarrow$ Get symmetrization from $SrcTrg$_Bleu with the highest BLEU score
8 $SrcPvt$_Symm $\leftarrow$ Get symmetrization from $SrcPvt$_Bleu with the highest BLEU score
9 $PvtTrg$_Symm $\leftarrow$ Get symmetrization from $PvtTrg$_Bleu with the highest BLEU score
10 $p(s|t)* \leftarrow$ phrase table of $SrcTrg$_Symm
11 $p(\overline{s}|\overline{t})* \leftarrow$ phrase table of $SrcPvt$_Symm
12 $p(\overline{s}|\overline{t})* \leftarrow$ phrase table of $PvtTrg$_Symm
13 $p(s|t; \lambda) \leftarrow$ run Equation (4) //combine_given_tuning_set to setup the weights of $\lambda_{src-trg}$ and $\lambda_{pvt}$.
\end{verbatim}

4. Results and Discussion

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

4.1. Direct System Approach (DSA)

Table 3 shows the results of the DSA for each language model. Of 30 experiments, 20 showed that translation systems using LM05 produced higher BLEU scores than those using LM03. Table 3 also shows the BLEU scores of each symmetrization of word alignment. We found that the highest BLEU scores were not always generated by gdfand. For example, Kk–En LM05 tgttosrc obtained a higher BLEU score than Kk–En LM05 gdfand, that is, 3.56, showing that non-standard symmetrization could be an alternative option to improving the BLEU scores of the pivot approaches. Our results confirmed the language-specific [8–10] and dataset-specific [12] characteristics of the symmetrization of word alignment.
Table 3. The obtained bilingual evaluation understudy (BLEU) scores of direct system approach (DSA). Results in bold indicate the highest translation quality.

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

We identified the reasons for the different BLEU scores in the same LM, despite using the same automatic word alignment and decoder weights, i.e., GIZA++ and moses.ini, respectively. We compared the phrase translation parameter scores between two phrase tables of the highest and second-highest BLEU scores in the same LM. Phrase translation parameter scores were computed from the co-occurrence of aligned phrases in the training corpora, then stored in the phrase table along with the phrase pair. The scores consisted of inverse phrase translation probability \( p(t|s) \), inverse lexical weighting \( \text{lex}(t|s) \), direct phrase translation probability \( p(s|t) \), and direct lexical weight \( \text{lex}(s|t) \). First, we collected 2000 phrase pairs and their phrase translation parameter scores from the phrase table with the highest BLEU score: phrase Table 1 (PT1). Subsequently, we collected 12,000 phrase pairs and their phrase translation parameter scores from phrase table with the second-highest BLEU score: phrase Table 2 (PT2). Last, we examined which component of the phrase translation parameter of PT1 obtained higher scores than the phrase translation parameter of PT2 in the same phrase pair, as shown in Table 4. We changed PT2 from srtctgt to tgttosrc since the results could not obtain the same phrase pairs in two language pairs, Kk–Ru and Ru–En, marked by an asterisk (*) in Table 4. The comparison algorithm of phrase translation parameter scores between the two phrase tables can be accessed in our repositories [37].

Table 4. Comparison result of phrase translation parameter scores between two phrase tables. Results in bold indicate the highest scores. PT2 was changed from srtctgt to tgttosrc, marked by an asterisk (*).

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

Table 4 shows that most language pairs obtained higher score in \( p(t|s) \) and \( \text{lex}(t|s) \), except for two language pairs: Kk–En and Ms–Id. The inverse phrase translation probability \( p(t|s) \) and inverse lexical weighting \( \text{lex}(t|s) \) were obtained from the target–source \( t|s \) parallel corpora. The results indicated that target-source \( t|s \) parallel corpora more strongly influence the phrase translation parameter score than source–target \( s|t \) parallel corpora.
Table 5 shows the phrase pair and their phrase translation parameter score examples from the two phrase tables.

Table 5. Example of phrase translation parameter scores in Kk–En LM03. Results in bold indicate the highest scores.

| Phrase Pair                  | Phrase Translation Parameters | Symmetrization |
|------------------------------|-------------------------------|----------------|
| 2007 жылдан бастап | Inverse phrase translation probability ($p(t|s)$) | 0.5 0.5 |
| | Inverse lexical weighting ($lex(t|s)$) | 9.30714e-4 4.8791e-5 |
| | Direct phrase translation probability ($p(s|t)$) | 0.5 3.33333e-1 |
| | Direct lexical weighting ($lex(s|t)$) | 5.96183e-3 1.28321e-2 |

4.2. Interpolation System Approach (ISA)

In the ISA, we constructed two subsystems: Std-ISA and H-ISA. Std-ISA was our interpolation system that uses gdfand, whereas H-ISA uses the symmetrization of word alignment that obtained the highest BLEU score. The choice of symmetrization of word alignment for H-ISA is shown in Table 6. Considering Kk–En LM05 H-ISA as an example, we employed tgttosrc in Kk–En as src–trg, whereas we used gdfand in Kk–Ru as src–pvt and Ru–En as pvt–trg.

Table 6. Candidates of symmetrization of word alignment for highest-interpolation system approach (H-ISA).

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

Table 7 shows the ISA result. We included the direct translation src–trg of Kk–En and Ja–Id as a baseline. We found that all the translation systems using LM05 obtained higher BLEU scores than those using LM03. For Kk–En, we found that H-ISA is a competitive approach because it provided absolute improvements of 0.35 and 0.22 BLEU points over baseline and Std-ISA in LM03 and LM05, respectively. Table 7 shows the different effect of H-ISA on Ja–Id. H-ISA obtained absolute improvements of 0.11 BLEU points over baseline in LM03. However, H-ISA obtained an absolute drop of −0.12 BLEU points compared to baseline in LM05. We compared 2000 of the same phrase pairs from two phrase tables: H-ISA and baseline. We provide an example of H-ISA and baseline phrase pairs and phrase translation parameter score in Table 8. We found that more than 1900 phrase pairs of H-ISA in LM05 obtained lower phrase translation parameter scores compared to baseline in LM05. Therefore, lower phrase translation parameter scores could be a reason for the lower BLEU score for Ja–Id using the H-ISA.
Table 7. The obtained BLEU scores of direct translation of src–trg (baseline) and interpolation system approach (ISA). Results in bold indicate the highest translation quality.

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

Table 8. Example of phrase translation parameter scores of Ja–Id LM05. Results in italic indicate the lowest score.

| Phrase-Pair                      | Phrase Translation Parameter                          | Phrase Table |
|----------------------------------|------------------------------------------------------|--------------|
| この建物の bangunan                | Inverse phrase translation probability \( p(t|s) \)    | 9.52381e-3   |
| This building bangunan           | Inverse lexical weighting \( lex(t|s) \)              | 2.89e-6      |
| (This building bangunan)         | Direct phrase translation probability \( p(s|t) \)    | 3.33333e-1   |
|                                  | Direct lexical weighting \( lex(s|t) \)              | 4.88889e-1   |

Additionally, we investigated why Ja–Id LM03 using the Std-ISA and Ja–Id LM03 using the H-ISA obtained same BLEU score: 12.07 and 12.07, respectively. We found that both systems used the same candidates of symmetrization of word alignment in three sides of the pivot approaches: gdfand in Ja–Id, gdfand in Ja–Ms, and gdfand in Ms-Id, as shown in Table 6. In contrast to Ja–Id LM03, Ja–Id LM05 with the Std-ISA and Ja–Id LM05 with the H-ISA obtained the same BLEU score when using different candidates of symmetrization of word alignment, as shown in Table 6. Ja–Id LM05 using the Std-ISA used gdfand in Ja–Id, gdfand in Ja–Ms and gdfand in Ms–Id. Ja–Id LM05 of H-ISA used gdfand in Ja–Id, gdfand in Ja–Ms and tgttosrc in Ms-Id. We compared 2000 of the same phrase pairs from two phrase tables: Std-ISA and H-ISA. Table 8 presents an example of Std-ISA and H-ISA phrase pairs and phrase translation parameter scores. We found that more than 1700 phrase pairs of Std-ISA and H-ISA obtained relatively similar phrase translation parameter scores. The result demonstrated that relatively similar phrase translation parameter scores using different symmetrization of word alignment could obtain the same BLEU score.

We investigated the relationship between BLEU score and phrase table size in each system, as shown in Tables 7 and 9. We found that systems with small phrase tables, i.e., Kk–En LM03 of H-ISA, Kk–En LM05 of H-ISA, and Ja–Id LM05 of baseline, obtained higher BLEU scores: 3.43, 3.64, and 12.20, respectively, as shown in Table 7. However, we also found that systems with large phrase tables, i.e., Ja–Id LM03 of Std-ISA and Ja–Id LM03 of H-ISA, obtained higher BLEU scores, that is, 12.07 and 12.07, respectively. Our first finding aligns with [38], who stated that higher BLEU scores could be obtained when using small phrase tables. Our second finding aligns with that of [11], who obtained higher BLEU scores when using large phrase tables. Our results demonstrated that higher BLEU scores can be obtained when using either small or large phrase tables.

Table 9. Phrase table size for direct translation of src–trg (baseline) and ISA.

| Language Model | Baseline | Std-ISA | H-ISA |
|----------------|----------|---------|-------|
| Kk–En          |          |         |       |
| LM03           | 723,960  | 742,948 | 323,850|
| LM05           | 723,960  | 742,948 | 323,850|
| Ja–Id          |          |         |       |
| LM03           | 875,038  | 935,717 | 935,717|
| LM05           | 875,038  | 935,717 | 925,732|
We identified why systems with small or large phrase tables could obtain higher BLEU scores when using the same LM order and decoder weights. We compared phrase translation parameter scores between the two phrase tables of the highest and second-highest BLEU score in the same LM. First, we collected 2000 phrase pairs and their phrase translation parameter scores from the phrase table of the highest BLEU score as phrase Table 1 (PT1). Subsequently, we collected 12,000 phrase pairs and their phrase translation parameter scores from the phrase table of second-highest BLEU score as phrase Table 2 (PT2). Lastly, we examined whether the phrase translation parameter of PT1 obtains higher scores than that of PT2 in the same phrase pair. We found that the phrase translation parameter of PT1 obtained higher scores than that of PT2 in a system with a small phrase table, as shown in Table 10. Consider an example of Kk–En LM03 of H-ISA that obtained higher scores, 514, 576, and 639, in $\text{lex}(t|s)$, $p(s|t)$ and $\text{lex}(s|t)$, respectively, compared to Kk–En LM03 using the Std-ISA. The result indicated that a system with a small phrase table could obtain a higher BLEU score because of the higher phrase translation parameter scores, particularly in $p(s|t)$ and $\text{lex}(s|t)$. In contrast to the system with a small phrase table, we found that the phrase translation parameter of PT1 had a lower score than that of PT2 in a system with a large phrase table. Table 10 shows that Ja–Id LM03 had lower scores, 13, 18, 17, and 34, in $p(t|s)$, $\text{lex}(t|s)$, $p(s|t)$, and $\text{lex}(s|t)$, respectively. The result demonstrated that a system with a large phrase table can obtain higher BLEU scores due to the lower phrase translation parameter scores. Table 11 shows the phrase pair and its phrase translation parameter score examples from Ja–Id LM05.

We evaluated the perplexity score of each system, as shown in Table 12. We found that the longer LM order of Kk–En, i.e., LM05, obtained a lower perplexity score than the shorter one, i.e., LM03. In contrast, the longer LM order could not obtain a lower perplexity score than the shorter one for Ja–Id. Table 12 shows that Ja–Id LM05 obtained higher perplexity scores than Ja–Id LM03, although Ja–Id LM05 obtained a higher BLEU score than Ja–Id LM03. We identified this higher perplexity score in Ja–Id using the target monolingual corpus size as the first parameter. The target monolingual corpus was trained by the LM toolkit, i.e., KenLM, and generated various lists of $n$-gram probabilities stored in arpa file. We compared the English and Indonesian target monolingual corpus in Kk–En and Ja–Id. English, with a larger target monolingual size, i.e., 532,560 and longer LM order, i.e., LM05, had larger various lists of $n$-gram probabilities, i.e., 29,181,816. In contrast, Indonesian with a smaller target monolingual size, i.e., 8500 and longer LM order, i.e., LM05, had smaller various lists of $n$-gram probabilities, i.e., 740,766. As a result, the choice of language model probabilities in the decoding process could be smaller and affected the perplexity score of Ja–Id.

Table 10. Comparison of phrase translation parameter scores between two phrase tables. Results in bold indicate the highest scores.

| Language Pair | PT1 | PT2          | #Same Phrase Pair | PT1   | PT2   |
|---------------|-----|--------------|-------------------|-------|-------|
|               |     |              |                   | p(ts) | lex(ts) | p(st) | lex(st) | p(ts) | lex(ts) | p(st) | lex(st) |
| Small phrase table |     |              |                   |       |        |       |        |       |        |       |        |
| Kk–En LM03    | H-ISA | Std-ISA     | 872               | 296  |  514  |  576  |  639  |  576  |  358  |  296  |  233  |
| Kk–En LM05    | H-ISA | Baseline    | 708               | 131  |  342  |  412  |  474  |  577  |  366  |  296  |  234  |
| Ja–Id LM05    | Baseline | Std-ISA | 2000             | 1984 | 1975  | 1981  | 1964  |   16  |   25  |    19 |    36 |
| Large phrase   |     |              |                   |       |        |       |        |       |        |       |        |
| Ja–Id LM03    | Std-ISA | Baseline | 1825             | 13   |  18   |  17   |  34   | 1812  | 1807  | 1808  | 1791  |
Table 11. Example of phrase translation parameters of Ja–Id LM05. Results in bold indicate the highest score.

| Phrase-Pair | Phrase Translation Parameters | Phrase Translation Scores |
|-------------|-------------------------------|--------------------------|
| から強制||| rumahnya digerebek ( Forced from ||| his house was raided) | Inverse phrase translation probability ($p(f|e)$) | 0.25 | 0.2209 | 0.2229 |
| から情報を受け取る||| menerima informasi dari (Receive information from ||| receive information from) | Inverse lexical weighting ($\text{lex}(f|e)$) | 0.0024 | 0.0018 | 0.0018 |
| から強制||| rumahnya digerebek ( Forced from ||| his house was raided) | Direct phrase translation probability ($p(s|t)$) | 1.00 | 0.8839 | 0.8847 |
| から情報を受け取る||| menerima informasi dari (Receive information from ||| receive information from) | Direct lexical weighting ($\text{lex}(e|f)$) | 0.1019 | 0.0899 | 0.0901 |

Table 12. Perplexity scores for the direct translation of src–trg (baseline) and ISA.

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

Additionally, we identified another parameter that could have influenced the increase in perplexity scores in Ja–Id LM05: the feature function weight of LM. This parameter weight was generated from the target monolingual corpus and then stored in the decoder, moses.ini. A decoder is an SMT component that finds the best translation according to the product of the translation and language model probabilities [39]. A good value for a feature function weight of LM is 0.1–1. We found that the feature function weight of LM for Kk–En LM05 was higher than for Kk–En LM03 (0.10 and 0.06, respectively) when using a larger target monolingual corpus, i.e., 532,560. In contrast to Kk–En, we found that the feature function weight of LM for Ja–Id LM05 was lower, 0.09, than for Ja–Id LM03, 0.11, when using a smaller target monolingual corpus, i.e., 8500. As a result, Kk–En LM05 could have obtained a lower perplexity score than Kk–En LM03, whereas Ja–Id LM05 obtained a higher perplexity score than Ja–Id LM03, as shown in Table 12. Figure 2 shows the feature function weight of LM for Kk–En and Ja–Id.

Figure 2. Feature function weight of the language model (LM) for Kk–En and Ja–Id.

We evaluated the generated text of the systems. Tables 13 and 14 show two sentence examples in each language pair, marked by (1) and (2). Sentence (1) is a long sentence,
whereas (2) is a short one. We added the English translation in the generated text of Ja–Id to better understand the translation results, marked in italics. Table 13 shows that the H-ISA generated a compact sentence compared to others in the short sentence. The compact sentence means that the generated text obtained the same keywords as a reference, i.e., without additional words. Consider an example of Kk–En LM03 H-ISA that generated compact keywords, i.e., that means ensuring that business investment, jobs. Kk–En LM05 baseline generated additional words, i.e., federal government provides, is to reverse the lost, which were not available in the reference. In contrast to Kk–En, all the systems of Ja–Id generated compact sentences, as shown in Table 14.

Table 13. Generated text examples of Kk–En.

| Source | Reference | Direct translation | Std-ISA | H-ISA |
|--------|-----------|--------------------|---------|-------|
| LM03   | the positive trends identified by the high rate of unemployment, just 2 years keep 4.8% runs into the limits to remain vigilant. | that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business. | that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business. | that means ensuring that—and business investment, jobs. |
| LM05   | the 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. | that means ensuring that the federal government provides the investment, is to reverse the loss of jobs and business. | that means ensuring that—and business investment. |

Tables 13 and 14 also show that the word order of the generated text was incorrect. We found that the generated text appeared to follow the source language’s sentence pattern, i.e., subject-object-verb (SOV), whereas the sentence pattern of the target language is subject-verb-object (SVO). We used the default reordering model, i.e., msd-bidirectional-fe; however, our generated text results still followed the source language’s pattern. The msd-bidirectional-fe is a default reordering model in SMT that considers the orientation of the model, directionality, and languages. The incorrect word order may also be the reason for the insignificant improvement in the BLEU score in our system, as shown in Table 7. BLEU is an evaluation metric that measures the similarity between two text strings and assigns too much weight to correct word order [40]. BLEU was used a reference file to evaluate the generated text. Our generated text results follow the source languages’ sentence pattern, i.e., SOV, whereas our reference files used the SOV pattern. Therefore, the evaluation process could not obtain the maximum results due to the differing word order between the generated text and the reference file. Figure 3 illustrates an example of the generated text from the system of Ja–Id LM03 H-ISA. The system translates a source sentence (s) into the target sentence (t). The result showed that the generated Indonesian text had the same word position as the Japanese text. Additionally, Figure 3 illustrates a comparison between t and r with a different phrase. Given an example of t, which used the phrase “3 orang tewas/3 people died”, r used the phrase “tiga orang yang tewas/three people who were killed”. Furthermore, t and r had different word positions, leading to the difficulty in maximizing BLEU script usage.
Table 14. Generated text examples of Ja–Id.

| (1)                                                    | (2)                                                    |
|--------------------------------------------------------|--------------------------------------------------------|
| Source                                                 | Reference                                              |
| その2004年の地震はマグニチュード9.1と記録され、33フィートの高さに達する波を伴う津波を引き起こした。| 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.) |

| Direct translation                                    |                          |
|--------------------------------------------------------|--------------------------|
| タイムオーダ3203 | タイムオーダ3203 |
| タイムオーダ3205 | タイムオーダ3205 |
| タイムオーダ3203 | タイムオーダ3205 |
| タイムオーダ3205 | タイムオーダ3205 |

| Ind-ISA                                                |                          |
|--------------------------------------------------------|--------------------------|
| タイムオーダ3203 | タイムオーダ3203 |
| タイムオーダ3205 | タイムオーダ3205 |
| タイムオーダ3203 | タイムオーダ3205 |
| タイムオーダ3205 | タイムオーダ3205 |

| H-ISA                                                 |                          |
|--------------------------------------------------------|--------------------------|
| タイムオーダ3203 | タイムオーダ3205 |
| タイムオーダ3205 | タイムオーダ3205 |
| タイムオーダ3203 | タイムオーダ3205 |
| タイムオーダ3205 | タイムオーダ3205 |

Figure 3. Sentence structure for Ja–Id, taken from LM03 H-ISA. Each color depicts the translation and word position of source–target and target–reference.

5. Conclusions and Future Work

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

Second, we proposed an approach to phrase table combination called H-ISA. The H-ISA is our interpolation system that uses the symmetrization of word alignment that obtained the highest BLEU scores from the DSA for each phrase table of src–trg, src–pvt, and pvt–trg. Our H-ISA is a competitive approach because it outperformed the direct translation of src–trg and standard-ISA (Std-ISA) by 0.38 and 0.22 in Kk–En LM03 and LM05, respectively. The H-ISA also outperformed the direct translation of src–trg by 0.11 in Ja–Id LM05. The direct translation of src–trg still outperformed the H-ISA by 0.12 in Ja–Id LM05. We found that the phrase translation parameter score of H-ISA in Ja–Id was
lower compared to that of baseline. We also found that the small phrase table could obtain higher BLEU scores, which contradicts the findings of a previous study [11].

Third, we evaluated the perplexity score of the ISA system. We found that the longer LM order, i.e., 5-gram, obtained lower perplexity scores than the shorter one, (3-gram) in Kk–En. However, the results for Ja–Id were the opposite to Kk–En. We found that this result could be caused by two parameters: the target monolingual corpus size and the feature function weight of the LM. For the first parameter, we identified that the English target monolingual corpus size in Kk–En was 62 times bigger, i.e., 532,560, than the Indonesian target monolingual corpus size in Ja–Id, i.e., 8500. Thus, the English LM obtained 29,181,816 various lists of \( n \)-gram probabilities, whereas the Indonesian LM obtained 740,766. We argued that smaller datasets could have limited the choice of LM probabilities in the decoding process and affected the perplexity scores of Ja–Id. For the second parameter, Ja–Id LM05 obtained a lower feature function weight of LM, i.e., 0.09, than Ja–Id LM03, i.e., 0.11. As a result, the perplexity scores of Ja–Id LM05 were higher than those of Ja–Id LM03. We argued that the lower feature function weight of LM could obtain lower translation score in the decoding process.

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

When we compared the perplexity scores of Ja–Id to those of Kk–En, we found that the target monolingual corpus size could be a factor for decreasing the perplexity scores. Thus, we will increase our Indonesian target monolingual corpus size in Ja–Id to be as large as Kk–En. Then, we will re-evaluate this parameter in our next work. Regarding the incorrect word order problem, we will implement the pre-ordering technique in the training dataset to improve the generated text and re-evaluate our system in our future work. Finally, the applicability of the proposed strategies was demonstrated on limited language pairs. Thus, another direction for future study is investigating other language pairs and datasets.

Author Contributions: Investigation, Conceptualization, S.D.B.; writing—original draft, S.D.B., A.H.A.M.S., and T.N.F.; writing—review and editing, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to restrictions e.g., privacy or ethical.

Acknowledgments: S.D.B. would like to thank Arie Ardiyanti Suryani, a member of InACL (Indonesia Association of Computational Linguistics) for useful discussions. A.H.A.M.S. acknowledges the scholarship from RISET-PRO (Research and Innovation in Science and Technology Project) KEMENRISTEKDIKTI (Ministry of Research, Technology and Higher Education of the Republic of Indonesia).

Conflicts of Interest: The authors declare no conflict of interest.
Abbreviations

The following abbreviations are used in this manuscript:

- SMT: Statistical Machine Translation
- NMT: Neural Machine Translation
- BLEU: Bilingual Evaluation Understudy
- MERT: Minimum Error Rate Training
- LM: Language Model
- DSA: Direct System Approach
- ISA: Interpolation System Approach
- H-ISA: Highest-Interpolation System Approach
- CUNI: Charles University
- NICT: National Institute of Information and Communications Technology
- UMD: University of Maryland
- Kk–En: Kazakh–English
- Ja–Id: Japanese–Indonesian
- src–trg: Source–target
- src–pvt: Source–pivot
- pvt–trg: Pivot–target
- gdfand: grow-diag-final-and
- srctotgt: source to target
- tgttosrc: target to source

References

1. Utiyama, M.; Isahara, H. A comparison of pivot methods for phrase-based statistical machine translation. In Proceedings of the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, New York, NY, USA, 22–27 April 2007; pp. 484–491.
2. Dabre, R.; Cromieres, F.; Kurohashi, S.; Bhattacharyya, P. 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, CO, USA, 31 May–5 June 2015.
3. Ahmadnia, B.; Serrano, J.; Haffari, G. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. In Proceedings of the International Conference Recent Advances in Natural Language Processing, Varna, Bulgaria, 2–8 September 2017; pp. 24–30. [CrossRef]
4. Budiwati, S.D.; Aritsugi, M. Multiple Pivots in Statistical Machine Translation for Low Resource Languages. In Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation, Hakodate, Japan, 13–15 September 2019; pp. 345–355.
5. Trieu, H.L.; Nguyen, L.M. A Multilingual Parallel Corpus for Improving Machine Translation on Southeast Asian Languages. In Proceedings of the MT Summit XVI, Nagoya, Japan, 18–22 September 2017; pp. 268–281.
6. Wu, H.; Wang, H. Pivot Language Approach for Phrase-based Statistical Machine Translation. Mach. Trans. 2007, 21, 165–181. [CrossRef]
7. Girgzdis, V.; Kale, M.; Vaicekauskis, M.; Zarina, I.; Skadina, I. Tracing Mistakes and Finding Gaps in Automatic Word Alignments for Latvian-English Translation. In Proceedings of the Human Language Technologies, Kaunas, Lithuania, 26–27 September 2014; pp. 87–94. [CrossRef]
8. Koehn, P.; Axlerod, A.; Birch, A.; Callison-Burch, C.; Osborne, M.; Talbot, D. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In Proceedings of the 2005 International Workshop on Spoken Language Translation, IWSLT 2005, Pittsburgh, PA, USA, 24–25 October 2005; pp. 68–75.
9. Singh, T.D. 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, Trivandrum, India, 11–14 December 2015; pp. 124–129.
10. Stymne, S.; Tiedemann, J.; Nivre, J. Estimating Word Alignment Quality for SMT Reordering Tasks. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 275–286. [CrossRef]
11. Kholy, A.E.; Habash, N. Alignment symmetrization optimization targeting phrase pivot statistical machine translation. In Proceedings of the 17th Annual Conference of the European Association for Machine Translation, Dubrovnik, Croatia, 16–18 June 2014; pp. 63–70.
12. Wu, H.; Wang, H. Comparative Study of Word Alignment Heuristics and Phrase-Based SMT. In Proceedings of the Machine Translation Summit XI, Copenhagen, Denmark, 10–14 September 2007; pp. 507–514.
13. Ahmadnia, B.; Serrano, J. Employing Pivot Language Technique Through Statistical and Neural Machine Translation Frameworks: The Case of Under-Resourced Persian-Spanish Language Pair. Int. J. Natural Lang. Comput. 2017, 6, 37–47. [CrossRef]
14. Ahmadnia, B.; Serrano, J.; Haffari, G.; Balouchzahi, N.M. Direct-bridge combination scenario for Persian-Spanish low-resource statistical machine translation. *Commun. Comput. Inf. Sci.* 2018. [CrossRef]

15. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12);* Chair, N.C.C.; Choukri, K.; Declerck, T.; Doğan, M.U.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; Piperidis, S.; Eds.; European Language Resources Association (ELRA): Istanbul, Turkey, 2012.

16. Riza, H.; Purwoadi, M.; Gunarso; Uliniansyah, T.; Ti, A.A.; Aljunied, S.M.; Mai, L.C.; Thang, V.T.; Thai, N.P.; Chea, V.; et al. Introduction of the Asian Language Treebank. In *Proceedings of the 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA),* Bali, Indonesia, 26–28 October 2016; pp. 1–6. [CrossRef]

17. Nomoto, H.; Okano, K.; Moeljadi, D.; Sawada, H. TUFs Asian Language Parallel Corpus (TALPCo). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing,* Melbourne, Australia, 19 October 2018; pp. 436–439.

18. Assylbekov, Z.; Nurkas, A. Initial Explorations in Kazakh to English Statistical Machine Translation. In *Proceedings of The First Italian Conference on Computational Linguistics (CILC-it 2014) and the Fourth International Workshop (EVALITA 2014);* Pisa University Press: Pisa, Italy, 2014; pp. 12–16. [CrossRef]

19. Kartbayev, A. SMT: A Case Study of Kazakh-English Word Alignment. In *Current Trends in Web Engineering*; Springer: Cham, Switzerland, 2015; pp. 40–49. [CrossRef]

20. Barrault, L.; Bojar, O.; Costa-jussà, M.R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; et al. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1);* Association for Computational Linguistics: Florence, Italy, 2019; pp. 1–61. [CrossRef]

21. Kocmi, T.; Bojar, O. CUNI Submission for Low-Language Resources in WMT News 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1);* Association for Computational Linguistics: Florence, Italy, 2019; pp. 234–240. [CrossRef]

22. Dabre, R.; Chen, K.; Marie, B.; Wang, R.; Fujita, A.; Utiyama, M.; Sumita, E. NICT’s Supervised Neural Machine Translation Systems for the 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, 2019; pp. 168–174. [CrossRef]

23. Briakou, E.; Carpuat, M. 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, 2019; pp. 134–140. [CrossRef]

24. Adiputra, C.K.; Arase, Y. 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: Cambridge, MA, USA, 2017.

25. Paul, M.; Yamamoto, H.; Sumita, E.; Nakamura, S. On the Importance of Pivot Language Selection for Statistical Machine Translation. In *Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* Stroudsburg, PA, USA, 31 May–6 June 2009; pp. 221–224.

26. Simon, H.S.; Purwarianti, A. Experiments on Indonesian-Japanese statistical machine translation. In *Proceedings of the 2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM),* Yogyakarta, Indonesia, 3–4 December 2013; pp. 80–84. [CrossRef]

27. Sulaiman, M.; Purwarianti, A. Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process. In *Proceedings of the 2015 International Conference on Electrical Engineering and Informatics (ICEEI),* Bali, Indonesia, 10–11 August 2015; pp. 54–58. [CrossRef]

28. Och, F.J. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of Association for Computational Linguistics Stroudsburg, PA, USA, 7–12 July 2003;* pp. 160–167. [CrossRef]

29. Sennrich, R. 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,* Stroudsburg, PA, USA, 19–23 April 2012; pp. 539–549.

30. Lembersky, G.; Ordan, N.; Wintner, S. Improving Statistical Machine Translation by Adapting Translation Models to Translationese. *Comput. Lingu.* 2013, 39, 999–1023. [CrossRef]

31. Dryer, M.S. Order of Subject, Object and Verb. In *The World Atlas of Language Structures Online;* Dryer, M.S., Haspelmath, M., Eds.; Max Planck Institute for Evolutionary Anthropology: Leipzig, Germany, 2013.

32. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. 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, 27–30 June 2007; pp. 177–180.

33. Johnson, H.; Martin, J.; Foster, G.; Kuhn, R. 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),* Prague, Czech Republic, 27–30 June 2007.

34. Hoang, D.T.; Bojar, O. TmTriangulate: A Tool for Phrase Table Triangulation. *Prague Bull. Math. Ling.* 2015, 104, 75–86. [CrossRef]

35. Liu, Y.; Zhang, J.; Hao, J.; Zhang, D. Making Language Model as Small as Possible in Statistical Machine Translation; Machine Translation; Shi, X., Chen, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1–12.
36. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002.

37. Budiwati, S.D.; Siagian, A.H.A.M, Fatyanosa, T.N.; Aritsugi, M. The comparison algorithm of phrase translation parameter scores. Available online: https://github.com/s4d3/PhraseTableCombination (accessed on 20 February 2021).

38. Tian, L.; Wong, D.F.; Chao, L.S.; Oliveira, F. A relationship: Word alignment, phrase table, and translation quality. Sci. World J. 2014, 2014. [CrossRef] [PubMed]

39. Jurafsky, D.; Martin, J.H. Speech and Language Processing, 2nd ed.; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2009.

40. Zhang, Y.; Vogel, S.; Waibel, A. Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System? In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004.