Benefiting from Language Similarity in the Multilingual MT Training: Case Study of Indonesian and Malaysian

Alberto Poncelas and **Johanes Effendi**
Rakuten Institute of Technology
Rakuten Group, Inc.
{first.last}@rakuten.com

Abstract

The development of machine translation (MT) has been successful in breaking the language barrier of the world's top 10-20 languages. However, for the rest of it, delivering an acceptable translation quality is still a challenge due to the limited resource. To tackle this problem, most studies focus on augmenting data while overlooking the fact that we can "borrow" high-quality natural data from the closely-related language. In this work, we propose an MT model training strategy by increasing the language directions as a means of augmentation in a multilingual setting. Our experiment result using Indonesian and Malaysian on the state-of-the-art MT model showcases the effectiveness and robustness of our method.

1 Introduction

In machine translation (MT), the definition of "low-resource" is not always tied to the language itself, but also to the pair of which languages we want to translate. For example, although Japanese cannot be classified as a low-resource language, training a Japanese-to-Indonesian (JA→ID) or Japanese-to-Malaysian (JA→MS) MT system can be challenging given that there is a very small number of parallel data for that language pair.

Accordingly, researches on multilingual MT (Liu et al., 2020; Fan et al., 2021) focus on improving translation results in low-resource language with the help of other high-resource languages in a unified singular model. However, such improvements are just significant in the translation directions involving English (EN→XX and XX→EN), while on non-English translation pair (XX→YY) the translation performance significantly decreases (NLLB Team et al., 2022) because the parallel data for those language pairs are not available.

In this work, we propose an MT model training strategy that benefits from the similar language, even when the similar language does not include initially in the model. We showcased the effectiveness of our proposed method using a commonly known similar language family of JA→ID and JA→MS translation pairs.

We use Indonesian and Malaysian as examples of closely-related languages. Both languages are commonly known as such due to their similar geographical and historical contexts, where it was used as the *lingua franca* throughout the Malay archipelago for over a thousand years (Paauw, 2009).

Popular strategies to improve the models that use low-resourced languages are based on data augmentation or transfer learning (Zoph et al., 2016) from a richer-resourced language. However, in this work, we focus on investigating exclusively the impact of including additional language-direction in the training process of a multilingual MT. We present several alternatives in a multilingual context where translation directions of similar languages can be used in combination, and how they can benefit each other. As Malaysian and Indonesian are similar languages, we hypothesize that performing multilingual training from Japanese in these two directions could be mutually beneficial.

We detail our proposal in Section 3. In Section 4 we describe the settings of the data and the models built. The performance of the different models are displayed in Section 5 and an analysis of the outputs in Section 6. Finally, we conclude this paper in Section 7 and propose different experiments that could be carried out in the future.

2 Related Work

In a resource-rich condition, improving an MT model can be done by simply adding more parallel data. This is possible for some European language pairs such as EN-DE, FR-EN, EN-IT and others. However, for most of the languages in the world, such data is more limited, which also yields to a

Proceedings of the 29th International Conference on Computational Linguistic, pages 84–92
October 12–17, 2022.

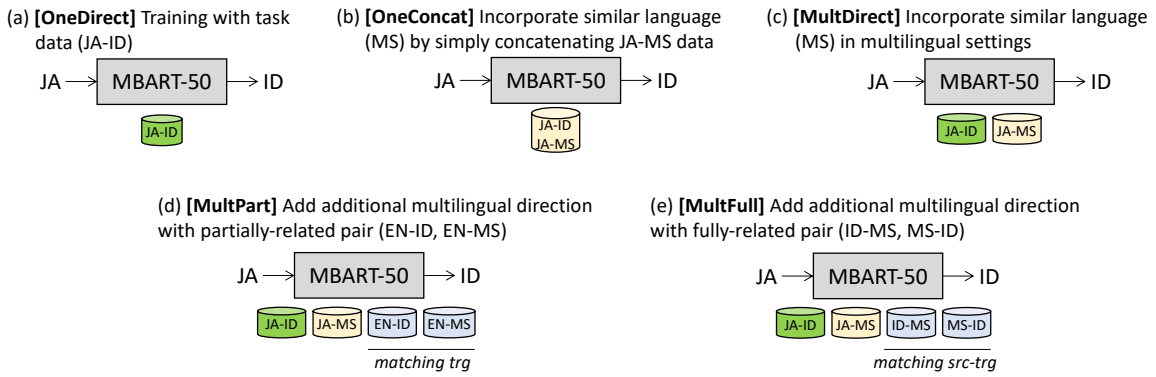

Figure 1: Overview of our proposed training strategy: (a) **[OneDirect]** is the single direction baseline training for JA→ID task, (b) **[OneConcat]** concatenate MS data with ID assuming that it is equivalent, (c) **[MultDirect]** adding JA-MS as additional multilingual direction, (d) **[MultPart]** add additional multilingual direction with partially-related pairs, and (e) **[MultFull]** add additional multilingual direction with fully-related pairs.

more limited performance.

Several method such as mixture-of-experts (Shazeer et al., 2017), data augmentation by paraphrasing (Mehdizadeh Seraj et al., 2015; Sekizawa et al., 2017; Effendi et al., 2018; Zhou et al., 2019) or by backtranslation (Sennrich et al., 2016; Hoang et al., 2018). Unfortunately, augmentation methods looking at the closeness of the language are sometimes overlooked. We argue that before applying an augmentation method that generates synthetic data, we should first focus on using the already available natural data from neighboring similar languages.

Previous studies, such as the work of Aw et al. (2009) and Susanto et al. (2012), focus on developing translation between Indonesian and Malaysian from the perspective of low-resource language. However, we observed that in practice, the translation demand actually comes from high-resource to low-resource language and vice-versa. Given that, the monolingual data will be imbalanced in either source and target of the translation, in addition to the difficulties of looking for parallel data.

Similarly, Zoph et al. (2016) studied transfer learning for NMT, in which a model built on high-resourced language data is used as an initialization for training on a similar low-resource language. While this was developed with data scarcity in mind, this is different from our work as we focus on the benefits of training simultaneously two low-resourced languages rather than transferring the knowledge from one language to another.

Furthermore, Nakov and Ng (2012) proposed a method to paraphrase between Indonesian and Malaysian through a confusion network. The paraphrase between both languages were then used to enrich the phrase table probability in the statistical machine translation (SMT) settings. Unfortunately, such methods are not compatible with the current state-of-the-art MT model, where probabilities are implicit and updated by backpropagation.

3 Extending the Bilingual MT with More Training Directions

In this study, we develop a training strategy that leverages the already available natural data as a means of augmentation, with the closeness of the language in mind, in a multilingual training context. In particular, we use a multilingual pretrained model such as the MBART-50 (Tang et al., 2020) (more information in Section 4). This configuration allows us to explore different approaches to build models. We take advantage of this multilingual setting to propose alternatives to improve the Japanese-to-Indonesian and Japanese-to-Malaysian translations.

The MBART-50 is a sequence-to-sequence model trained on 50 languages (Many-to-Many language directions, pivoting via English).[1] The model is built following the work of Liu et al. (2020). First, a denoising autoencoder on different languages is trained. Then, they performed multilingual training using parallel data where each sentence (both source and target) has a language identifier tag attached in the beginning.

In this work, we refer as "train" to the process of training the MBART-50 model, although in practice

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/multilingual#mbart50-models

it corresponds to a fine-tuning task. The term "fine-tune" is used exclusively to describe the process of further tuning a model that has already been trained with our data.

A summary of our experiments is presented in Figure 1 where we display different configurations of inclusion of language directions. A straightforward approach to address the problem would be to build one MT model in each language direction (**OneDirect**, Figure 1a). We use this setting as the baseline. One common alternative to benefit from both languages is to concatenate the datasets (**OneConcat**, Figure 1b). In our case, we append the training sentence pairs from JA-MS into that of JA-ID. Then we execute the training assuming JA→ID direction (which is preferable to JA→MS because our pretrained model, MBART-50, has not been trained on Malaysian sentences).

This study explores mostly how the translation quality can improve when the MT models of similar languages such as Indonesian and Malaysian are trained together instead as a separate translation direction. Therefore, we build a model where the train consist on a multilingual training using both JA-ID and JA-MS data (**MultDirect**, Figure 1c).

Additionally, we are also interested in exploring the impact of introducing additional similar language pairs. Particularly, we explore increasing the language direction in two cases: **MultPart**, which involves adding EN→ID and EN→MS directions, including therefore an additional language in the source side (Figure 1d); and **MultFull**, which imply adding exclusively sentences on fully-related languages of the target sides, i.e. Indonesian and Malaysian (Figure 1e).

Note that, except for **OneConcat**, in all the experiments we only change the number of language directions in the training process. The training data of each language pair remains always the same.

4 Experimental Settings

To conduct the experiments, we build models based on the pretrained MBART-50 (Tang et al., 2020) model built in Fairseq (Ott et al., 2019). It consists of a transformer (Vaswani et al., 2017) model with 12 layers both in the encoder and the decoder. All the sentences, regardless of the language, are tokenized using the same sentencepiece (Kudo and Richardson, 2018) model. The vocabulary of encoder and decoder is shared.

We use the data from "CCMatrix" (Fan et al., 2021; Schwenk et al., 2021) and "TED2020" (Reimers and Gurevych, 2020) for training and evaluation, respectively. Then, we use "FLORES-101" (Goyal et al., 2022) dataset for evaluation. The size of these datasets can be found in Table 1. The number of sentences of JA-MS is much smaller than JA-ID. In addition, we also calculated the number of shared vocabulary between the Indonesian and Malaysian parts of our training dataset. As can be seen in Table 2, both languages shared a substantial amount of vocabulary in our dataset in particular.

Dataset	JA-ID	JA-MS
train	7.7M	1.7M
dev	156K	11K
test	1K	1K

Table 1: Number of sentences for each dataset.

Language	#Vocab	%Common
ID	485928	21.2%
MS	166682	61.8%
ID ∩ MS	103017	-

Table 2: Number of shared vocabulary between ID and MS in the CCMatrix dataset (Schwenk et al., 2021).

In those experiments were more language direction are added (i.e. **MultPart** and **MultFull**), we also use the datasets from "CCMatrix". The sizes of these are displayed in Table 3.

We use L1-L2 notation to refer to a dataset of pairs of sentences and L1→L2 to specify the translation direction. In the case of multiple translation directions from the same source language, we use L1→{L2,L3} notation. Finally, L1↔L2 implies that the translation directions are both L1→L2 and L2→L1.

5 Experimental Results

The performances of the models are evaluated using both BLEU (Papineni et al., 2002) metric, which is based on the overlap of *n*-grams, and chrF2 (Popovic, 2015) which is a character-based

Dataset	size
EN-ID	15.7M
EN-MS	5.4M
MS-ID	7.8M

Table 3: Number of sentences of the additional datasets.

metric. We present the results in Table 4. Each row shows a different model, in which the dataset shown in the column "Language Directions" is used for the training.

In the first subtable, we show the performance of **OneDirect**, i.e. bilingual MT trained on a single direction, JA→ID (row 1) or JA→MS (row 2). We use these models as baselines.

5.1 Results of Multilingual Settings

In the first set of experiments, we evaluate the models trained in multiple language directions without including additional datasets.

The subtable *"Similar language data multilingual training"* row 4 includes the results of the **MultiDirect**, which is trained in a multilingual setting in both JA→ID and JA→MS direction together. If we compare the results of these models to those of **OneDirect** we observe an increase in performance of 0.2 and 0.4 BLEU points increase.

Note that by following this configuration, the target side is not mixed and there is a clear distinction between Indonesian and Malaysian during the training. Despite that, due to the shared vocabulary, the combination is mutually beneficial. This configuration is also more efficient than simply concatenating both datasets as in **OneConcat**, which underperformed the baselines.

5.2 Results of Augmentation with More Language Directions

In the second set of experiments, we introduced additional language pairs in the training.

In the subtable *"Partially-related language data multilingual training"* we show the results of **MultPart** model. These models include EN→ID and EN→MS directions. Therefore, the set of source languages is extended with a language that is very different from Indonesian or Malaysian (or Japanese). The performance of this configuration increased when compared to those of bilingual models.

We also include the results of **MultFull** model in the subtable *"Fully-related language data multilingual training"*. This consist of three parts as ID-MS data can be integrated in three different ways: (i) ID→MS direction (row 6); (ii) ID→MS direction (row 7); and (iii) both direction ID↔MS (row 8). Although the sentences in these configurations were the same, the biggest impact is observed when both related languages are present in the target. The Inclusion of MS↔IN direction achieves a performance similar to that of **MultPart**. Note that there is some difference in the number of sentences, according to Table 3, the EN→{ID,MS} extension has $15.7M + 5.4M = 21.1M$ sentences and ID↔MS has $2 * 7.8M = 15.6M$.

Interestingly, according to both BLEU and chrF2 metrics, by including only ID→MS or MS→ID directions (rows 6 and 7), the performance is lower than the **OneDirect** baseline. Therefore, simply adding more language directions is not a guarantee of improvement. This effect may be a consequence of the model aiming to find an optimal equilibrium of performance between more language pairs, and therefore it may underperform on those that we are interested in.

5.3 Additional Stage of Fine-tuning

As seen in the previous section, including several languages may harm the quality of the translation of the directions we are evaluating because the training needs to be optimized for more languages.

We suspect that the models could be further optimized for the task on hand. For this reason, we also fine-tune for an additional stage in JA→{ID,MS} directions. The results are shown in the *"+fine-tune"* rows of Table 4.

In these rows, we observe that the performance can increase further. Moreover, some models that underperformed the **OneDirect** baselines, such as **MultFull** where only MS→ID or ID→MS were included, surpassed the baselines after executing an additional fine-tune.

A question that still is left to answer is whether this second stage of fine-tuning should be performed on JA→ID and JA→MS individually or JA→{ID,MS} together. We present in Table 5 the fine-grained results. In this case, the difference in performance is not big. If we compare the + *fine-tune JA→ID* or + *fine-tune JA→MS* rows with their corresponding + *fine-tune JA →{ID, MS}* row in the same subtables, the differences are in the $[-0.2, +0.1]$ BLEU range and $[-0.41, +0.08]$ chrF2 range.

6 Discussion

6.1 Error Analysis of Translation Examples

In Table 6 we show some examples of the translations generated by our models. In particular, we show the outputs of: (i) **OneDirect**, i.e. the baseline models; (ii) **OneConcat**, where training data is concatenated; (iii) **MultPart**, with addition of

No.	Language Directions	JA → ID		JA → MS	
		BLEU	chrF2	BLEU	chrF2
OneDirect - *Single direction training with task data - Fig.1a*					
1.	JA → ID	18.2	49.88	-	-
2.	JA → MS	-	-	15.6	48.86
OneConcat - *Similar language data concatenation - Fig.1b*					
3.	JA → (ID + MS)	17.6	49.15	9.5	41.48
MultDirect - *Similar language data multilingual training - Fig.1c*					
4.	JA →{ID, MS}	18.4	50.14	16.0	49.18
	+ fine-tune JA → ID	18.3	49.40	-	-
	+ fine-tune JA → MS	-	-	16.8	50.13
MultPart - *Partially-related language data multilingual training - Fig.1d*					
5.	JA →{ID, MS}, EN → {ID,MS}	19.3	50.38	16.2	49.20
	+ fine-tune JA →{ID, MS}	20.0	51.69	17.0	50.31
MultFull - *Fully-related language data multilingual training - Fig.1e*					
6.	JA →{ID, MS}, ID → MS	17.0	49.19	15.7	49.16
	+ fine-tune JA →{ID, MS}	18.9	50.49	16.9	49.76
7.	JA →{ID, MS}, MS → ID	17.7	49.67	14.7	48.43
	+ fine-tune JA →{ID, MS}	18.5	49.91	16.6	48.96
8.	JA →{ID, MS}, ID ↔ MS	19.3	50.61	16.6	49.59
	+ fine-tune JA →{ID, MS}	19.9	51.22	16.8	49.94

Table 4: Experiment results.

No.	Language Directions	JA → ID		JA → MS	
		BLEU	chrF2	BLEU	chrF2
OneDirect - *Single direction training with task data - Fig.1a*					
1.	JA → ID	18.2	49.88	-	-
2.	JA → MS	-	-	15.6	48.86
MultPart - *Partially-related language data multilingual training - Fig.1d*					
5.	JA →{ID, MS}, EN → {ID,MS}	19.3	50.38	16.2	49.20
	+ fine-tune JA →{ID, MS}	20.0	51.69	17.0	50.31
	+ fine-tune JA → ID	20.1	51.54	-	-
	+ fine-tune JA → MS	-	-	16.8	50.35
MultFull - *Fully-related language data multilingual training - Fig.1e*					
8.	JA →{ID, MS}, ID ↔ MS	19.3	50.61	16.6	49.59
	+ fine-tune JA →{ID, MS}	19.9	51.22	16.8	49.94
	+ fine-tune JA → ID	20.0	51.30	-	-
	+ fine-tune JA → MS	-	-	16.6	49.53

Table 5: Results of different fine-tuning combination.

#1	Source	群島は半島の北120 kmに位置します。最大の島はキング ジョージ島で、そこにビジャ・ラス・エストレージャスの集落があります。
	Ref. (EN)	The archipelago lies 120 km north of the Peninsula. The largest is King George Island with the settlement of Villa Las Estrellas.
	OneConcat	島 terletak 120 km di utara Semenanjung, dan pulau terbesar adalah Pulau King George, di mana Anda dapat menemukan permukian di Bija Las Estrellas.
Japanese to Indonesian	Ref. (ID)	Kepulauan ini terletak 120 km dari utara Semenanjung. Pulau terbesar adalah Pulau King George yang memiliki pemukiman Villa Las Estrellas.
	OneDirect	Pulau ini terletak di 120 km sebelah utara Semenanjung, dan pulau terbesarnya adalah Pulau King George, di mana terdapat permukiman Raja Ruth-Estrajas.
	MultDirect	Kepulauan ini terletak di 120 km sebelah utara semenanjung, dan pulau terbesar adalah Pulau King George, di mana terdapat permukiman Raja Las Straights.
	MultPart	Gugusan ini terletak 120 km sebelah utara semenanjung, pulau terbesar adalah Pulau King George, di mana terdapat permukiman warga Bija Ras Estonia.
Japanese to Malaysian	Ref. (MS)	Kepulauan itu terbentang 120 km utara di Semenanjung. Pulau yang terbesar adalah King George Island dengan penempatan Villa Las Estrellas.
	OneDirect	Pulau ini terletak 120 km di utara Semenanjung, dan pulau terbesarnya adalah Pulau King George, di mana Anda dapat menemukan permukiman Angkor.
	MultDirect	Kepulauan ini terletak 120 km di utara semenanjung, dan pulau terbesar adalah King George Island, di mana terdapat perkampungan Vijay Rasheed.
	MultPart	Terletak 120 km di utara semenanjung, pulau terbesar adalah Pulau King George, di mana terdapat sebuah pengumpulan Bija Ras Estonia.
#2	Source	最後に、昆虫、げっ歯類、トカゲ、鳥といったはるかに多数の小さい獲物を餌とする小型の猫（野良猫を含む）が数多く生息しています。
	Ref. (EN)	Finally, there are many small cats (including loose pet cats) that eat the far more numerous small prey like insects, rodents, lizards, and birds.
	OneConcat	Akhirnya, banyak kucing-kucing kecil (termasuk kucing liar) yang memberi makan banyak mangsa kecil seperti serangga, gigi betina, kutub, dan burung.
Japanese to Indonesian	Ref. (ID)	Yang terakhir, ada banyak kucing kecil (termasuk kucing peliharaan yang lepas) yang memakan jauh lebih banyak mangsa kecil seperti serangga, binatang pengerat, kadal, dan burung.
	OneDirect	Akhirnya, ada banyak serangga, jamur, katak, dan burung yang jauh lebih sedikit mangsa (termasuk kucing liar) yang memberi makan kucing-kucing kecil.
	MultDirect	Akhirnya, ada banyak kucing kecil (termasuk kucing liar) yang memberi makan banyak mangsa kecil seperti serangga, belalang, kadal, dan burung.
	MultPart	Akhirnya, ada banyak kucing kecil (termasuk kucing liar) yang memakan banyak mangsa yang lebih kecil, seperti serangga, belalang, kadal, dan burung.
Japanese to Malaysian	Ref. (MS)	Akhir sekali, terdapat banyak kucing bersaiz kecil (termasuk kucing liar) yang memakan lebih banyak mangsa kecil seperti serangga, tikus, cicak dan burung.
	OneDirect	Akhirnya, banyak kucing-kucing kecil (termasuk kucing liar) yang memakan banyak mangsa yang lebih kecil seperti serangga, gigi betina, topeng, dan burung.
	MultDirect	Akhir sekali, terdapat banyak kucing kecil (termasuk kucing liar) yang memberi makan kepada banyak mangsa kecil seperti serangga, belalang, cicak dan burung.
	MultPart	Akhir sekali, terdapat banyak kucing kecil (termasuk kucing liar) yang memakan banyak mangsa yang lebih kecil, seperti serangga, beruang kutub, cicak dan burung.
#3	Source	科学者たちは、暗黒物質は、通常の物質と同じように他の暗黒物質にも影響を与えていると結論づけました。
	Ref. (EN)	The scientists were able to conclude that the dark matter affect other dark matter in the same way regular matter does.
	OneConcat	: Para ilmuwan menyimpulkan bahwa materi gelap juga mempengaruhi materi gelap lainnya seperti biasa.
Japanese to Indonesian	Ref. (ID)	Para ilmuwan dapat menyimpulkan bahwa materi gelap mempengaruhi materi gelap lainnya dengan cara yang sama seperti materi biasa.
	OneDirect	Para ilmuwan menyimpulkan bahwa materi gelap juga mempengaruhi materi gelap lainnya seperti materi normal.
	MultDirect	Para ilmuwan menyimpulkan bahwa materi gelap mempengaruhi materi gelap lainnya seperti materi biasa.
	MultPart	Para ilmuwan menyimpulkan bahwa materi gelap juga mempengaruhi materi gelap lainnya, seperti materi biasa.
Japanese to Malaysian	Ref. (MS)	Para saintis dapat menyimpulkan bahawa jirim gelap menjejaskan jirim gelap yang lain dalam cara yang sama dengan jirim biasa.
	OneDirect	Para ilmuwan menyimpulkan bahwa materi gelap juga mempengaruhi materi gelap lainnya seperti biasa.
	MultDirect	Para saintis menyimpulkan bahawa bahan gelap mempunyai kesan yang sama dengan bahan gelap yang lain seperti bahan biasa.
	MultPart	Para saintis menyimpulkan bahawa bahan gelap juga mempunyai kesan ke atas bahan gelap yang lain, sama seperti bahan biasa

Table 6: Translation examples.

partially-related language data; and (iv) **MultFull** using *ID↔MS* configuration. For (iii) and (iv), we show the output of fine-tuned version (i.e. + *fine-tune JA →{ID, MS}*, which are those that shown the highest performance).

In the first sentence of the table, the translation of “群島” (i.e. “archipelago” in English) is referred as “kepulauan” in the reference. However, the baselines incorrectly generate “pulau” which means “island” and the **OneConcat** method simply copied a Japanese character from the source. The models with additional language directions produced the same word as the reference (i.e. “kepulauan”), or “gugusan”, which is also correct.

In this sentence, we can also find an example of the limitation of these systems, which is the translation of proper nouns from katakana (which are the Japanese characters used to transliterate foreign terms into Japanese). For example, the source sentence includes the term “ビジャ・ラス・エストレージャス” which is the transliteration from Spanish of the name of the settlement called “Villa Las Estrellas”. In the translations, only **OneConcat** system was partially correct. The other models proposed different incorrect romanizations of the name.

The word 最後に (i.e. “finally”) is translated as “akhirnya” by baseline models. Although it is correct, in Malaysian “akhir sekali”, is more common. This term is only present in those models with additional data included. Note also that on this models, the term is clearly differentiated in Indonesian and Malaysian. The translation of “cats” (as in “many cats”) can be either “banyak kucing” or “kucing-kucing”, however some baselines (i.e. JA→MS and JA→(ID+MS)) generate a wrong combination of “banyak kucing-kucing”. This is corrected in the models with extra data. A similar outcome happens with the translation of the list “昆虫、げっ歯類、トカゲ、鳥 ” (“serangga, binatang pengerat, kadal, dan burung”, which in English is “insects, rodents, lizards, and birds”). The baselines fail to translate correctly some of these whereas **MultPart** and **MultFull** models translate them accurately.

In the last sentence, the translations of “scientists” (科学者たち), “material” (物質), or “that” are translated into Indonesian as “ilmuwan”, “materi” and “bahwa”, respectively. In Malaysian, these terms are more commonly translated as “saintis”, “jirim” and “bahawa”. We can see that in the reference, and also we find more occurrences

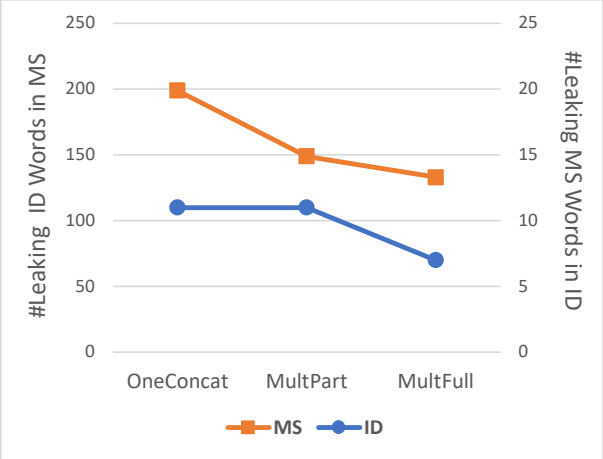

Figure 2: Number of vocabulary that leaks into the other language. Our proposed training strategy successfully decrease the number, which yields better fluency in the generated translation.

of them in the JA-MS training data. Despite that, the model trained on JA→MS produced the Indonesian terms instead. This shows that the pretrained MBART-50 model (built only with Indonesian data) has influenced the output. The models with additional data were able to learn this difference, and we see that the Malaysian outputs include the Malaysian terms.

6.2 Vocabulary Leak

As seen in the previous subsection, we could find some words that are not present in the training data of their corresponding language in the generated translation. Some of these words, however, can be found in the training data of the counterpart language. This suggests that some terms were unintentionally transferred (i.e. leaked) from one language into another. This might imply that although both languages are similar, the model is mixing the language too much. This will make the model hypotheses becomes unnatural due to decreasing fluency (i.e. sounds like a native Malaysian is speaking Indonesian or vice versa)

For example, in the hypothesis of **OneConcat** model, there were 11 words in the Indonesian output that came from JA-MS data, and 199 words in the Malaysian output that came from JA-ID. Hence, the vocabulary leak is more likely to happen from Indonesian to Malaysian. This is explained by the fact that MBART-50 model is built on Indonesian, and also because the size of JA-ID data is larger (Table 1).

Despite that, **MultPart** and **MultFull** mod-

els, which were trained with more sentences, the Indonesian-to-Malaysian vocabulary leak decreased from 199 to 149 (inclusion of EN→{ID,MS}) and to 133 (inclusion of ID↔MS). In the opposite direction, only the **MultFull** caused to decrease from 11 to 7 occurrences. Note that the difference in scale between Indonesian and Malaysian (133 to 7) is due to the dataset size and domain differences.

Nevertheless, the decrease in leaking words number (Figure 2) shows that our proposed method is crucial to let the model better differentiate between the two languages. Although we have shown that the two languages are similar, the model still needs to differentiate between two languages. Therefore, the integrity of the vocabulary is maintained, which increases fluency in the generated translation.

7 Conclusion and Future Work

In this work, we explored different techniques to improve the training of MT models to translate from Japanese into Indonesian and Malaysian. As finding resources in these target languages is not always easy, we focused on how to benefit from their similarities in multilingual conditions.

The results showed how training in both directions jointly boosts the translation quality of each translation. However, an interesting outcome is that simply including additional language pairs alone does not necessarily lead to improvements. In some cases, an additional step of fine-tuning was required so the models achieve higher performances than those built in a single direction.

We believe that the outcomes of these experiments are also applicable to other languages. In the future, we want to explore this approach for other language families, such as Romance or Slavic, or even dialects or variations of the same language. Some of them may also be low-resourced and it may be difficult to find enough data to build competitive MT models. On top of that, the techniques investigated in this study are complementary to other strategies (e.g. data augmentation, zero-shot learning) that are commonly used to increase the performance of low-resourced MT models. Accordingly, we suppose that a combination of them with our proposed training strategy could further improve the translation performance.

References

AiTi Aw, Sharifah Mahani Aljunied, Lianhau Lee, and Haizhou Li. 2009. Piramid: Bahasa Indonesia and Bahasa Malaysia translation system enhanced through comparable corpora. In *TCAST*.

Johanes Effendi, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2018. Multi-paraphrase augmentation to leverage neural caption translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 181–188, Brussels.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd workshop on neural machine translation and generation*, pages 18–24, Melbourne, Australia,.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1379–1390, Lisbon, Portugal. Association for Computational Linguistics.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 44:179–222.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht,

Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, Minneapolis, USA.

Scott H Paauw. 2009. *The Malay contact varieties of Eastern Indonesia: A typological comparison*. State University of New York at Buffalo.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Maja Popovic. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Édouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Bangkok, Thailand.

Yuuki Sekizawa, Tomoyuki Kajiwara, and Mamoru Komachi. 2017. Improving Japanese-to-English neural machine translation by paraphrasing the target language. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 64–69, Taipei, Taiwan.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *The 5th International Conference on Learning Representations, ICLR*, Toulon, France.

Raymond Hendy Susanto, Septina Dian Larasati, and Francis Tyers. 2012. Rule-based machine translation between indonesian and malaysian. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 191–200, Mumbai, India.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, USA.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2019. Paraphrases as foreign languages in multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 113–122, Florence, Italy.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin,Texas,USA.