Neural Machine Translation between Low-Resource Languages with Synthetic Pivoting

Khalid N. Elmadani and Jan Buys
Department of Computer Science
University of Cape Town
ahmkha009@myuct.ac.za, jbuys@cs.uct.ac.za

Abstract

Training neural models for translating between low-resource languages is challenging due to the scarcity of direct parallel data between such languages. Pivot-based neural machine translation (NMT) systems overcome data scarcity by including a high-resource *pivot* language in the process of translating between low-resource languages. We propose *synthetic pivoting*, a novel approach to pivot-based translation in which the pivot sentences are generated synthetically from both the source and target languages. Synthetic pivot sentences are generated through sequence-level knowledge distillation, with the aim of changing the structure of pivot sentences to be closer to that of the source or target languages, thereby reducing pivot translation complexity. We incorporate synthetic pivoting into two paradigms for pivoting: cascading and direct translation using synthetic source and target sentences. We find that the performance of pivot-based systems highly depends on the quality of the NMT model used for sentence regeneration. Furthermore, training back-translation models on these sentences can make the models more robust to input-side noise. The results show that synthetic data generation improves pivot-based systems translating between low-resource Southern African languages by up to 5.6 BLEU points after fine-tuning.

Keywords: neural machine translation, pivot-based translation, low-resource translation

1. Introduction

Neural Machine Translation (NMT) is the state-of-the-art approach for automatic translation, producing high-quality output text when translating between high-resource languages (Wu et al., 2016). However, translation between low-resource languages is more challenging due to the limited availability of high-quality parallel corpora between such languages (Burlot and Yvon, 2018). Moreover, it is easier to find parallel data between a high-resource language and a low-resource language than between two low-resource languages.

Pivot-based NMT approaches for translating between low-resource languages leverage the availability of parallel data between a high-resource language (the pivot language) and the source and target low-resource languages, respectively. In this scenario both low-resource languages have a reasonable amount of parallel data with the pivot language, compared to a much smaller amount of parallel data directly between the low-resource pair.

There are two main ways to make use of the pivot language (see figure 1). The first is to translate from the source language to the pivot language, and then from the pivot language to the target language, using two separate NMT models (cascading). The second approach generates synthetic direct translation sentences by translating the pivot language sentences to either or both the source and target languages, and then trains an NMT model for directly translating from the source to the target using the synthetic translated data (Park et al., 2017; Currey and Heafield, 2019). In both approaches, some model parameters are used unnecessarily to learn the word order of the pivot language, which is not actually necessary for the model to learn correctly as long the final system can translate from the source into the target language.

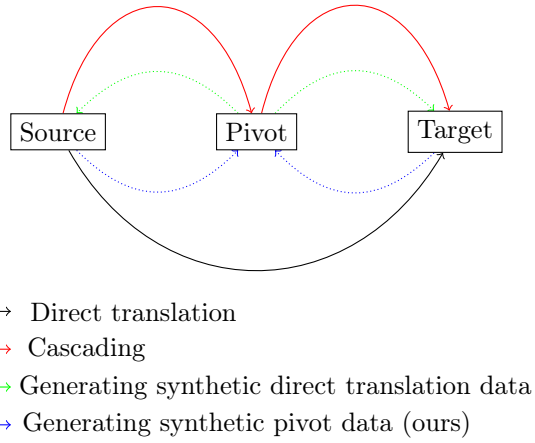

Figure 1: Pivoting translation paradigms. The arrows indicate the translation process in the case of direct translation and cascading. The dotted arrows indicate how the training data is generated.

In this paper, we propose *Synthetic Pivoting*, a novel approach to pivot-based translation in which the pivot sentences are generated synthetically from both the source and target languages. We investigate different ways of regenerating pivot sen-

LREC-COLING 2024, pages 12144–12158
20-25 May, 2024.

tences with the aim of pushing them towards the structure (word order) of source and target languages. Our approach is intended for translation between related low-resource languages that have similar linguistic structure, in which intermediate reordering towards the word order of the pivot language is particularly wasteful. We show how changing the structure of pivot sentences affects the performance of different pivot-based NMT systems.

Our approach consists of two phases. In the first phase, synthetic pivot sentences are generated through sequence-level knowledge distillation from NMT models (§4). These NMT models include Auto-regressive (AT) and Non-autoregressive (NAT) models. The sentences generated through sequence-level knowledge distillation are more structurally similar to the input-side language than the original sentences (Zhou et al., 2021). In the second phase we apply the two standard paradigms for pivoting: cascading (§5) and generating synthetic direct translation data §6, while replacing the original pivot sentences with the regenerated sentences from the first phase. We refer to these modified pivot-based approaches as *Synthetic Cascading* and *Synthetic Direct Translation*, respectively.

We evaluate the proposed approaches using two pairs of closely related low-resource Southern African languages: Xhosa and Zulu ($xho \leftrightarrow zul$) from the Nguni Language family, and Sepedi and Tswana ($nso \leftrightarrow tsn$) from the Sotho-Tswana language family. We use English (eng) as pivot language in both cases. Results show gains of 0.3 BLEU points on $xho \leftrightarrow zul$ and 5.6 BLEU points on $nso \leftrightarrow tsn$ over the bilingual baseline for each language pair. Additional gains are obtained when using a multilingual NMT model for cascading or generating synthetic direct translation data.

2. Background

Sequence-to-sequence models (Bahdanau et al., 2016; Gehring et al., 2017; Vaswani et al., 2017) have proven to be effective in the MT task. They utilize the expressiveness of neural networks and the sequential property of language in training autoregressive Neural Machine Translation (NMT) models, which has became the standard approach for training MT models. In autoregressive translation (AT), each token in the target sentence is generated conditioned on the previous target tokens and the source sentence.

2.1. Non-Autoregressive Machine Translation

An alternative approach is non-autoregressive translation (NAT), where the whole target sentence is generated simultaneously (Gu et al., 2018; Lee et al., 2018). The main motivation for NAT was to improve inference speed over AT which generates the output one token at a time. In NAT the output tokens are conditionally independent of each other given the source sentence, which can lead to inconsistencies in the output sequence. Some recent approaches proposed modified architectures and the training procedure that increase the dependency between output tokens (Stern et al., 2019; Ghazvininejad et al., 2019; Gu et al., 2019). However, due to the complexity of the training data, NAT models still lag behind AT in terms of translation quality.

Gu et al. (2018) proposed using sequence-level knowledge distillation (Kim and Rush, 2016) to train NAT models using synthetic target sentences generated from an AT model. These sentences tend to be simpler than the actual target sentences in terms of word order and lexical choice. Zhou et al. (2021) found that when aligning real and synthetic target sentences to source sentences, synthetic data has less reordering compared to real data, i.e., their structure is shifted to some degree towards the source language. Although the translation quality of synthetic target sentences might be lower, the desired effect of reducing structural changes enables training better NAT models. Another side effect of knowledge distillation is mode reduction: distilled data tend to contain fewer lexical choices per source word, which lowers the difficulty of learning (Ding et al., 2020).

2.2. Pivot-based NMT

Pivot-based NMT approaches enable translating between two low-resource languages via a high-resource language (Johnson et al., 2017). When no direct source-target data is used they are an instance of zero-shot NMT.

Cascading involves training two separate models, one for translating from the source language to the pivot and the other for translating from the pivot to the target language. The final translation system is a cascade of the two models. A drawback of this approach is error propagation (Johnson et al., 2017), i.e., translation errors made by the source-to-pivot model are passed to the pivot-to-target model.

The second pivoting paradigm involves generating a source-target synthetic dataset for training the source-to-target translation model, instead of having two decoding stages (source-to-pivot and pivot-to-target). The synthetic translation dataset is generated by translating pivot sentences to source/target using pre-trained pivot-to-source/target models. There are several ways of generating synthetic source-target parallel data:

1. Translating the pivot side of the source-pivot

Language Pairs	WMT22_african
eng-xho	8.6M
eng-zul	3.8M
xho-zul	1M
eng-tsn	5.9M
eng-nso	3M
nso-tsn	235K

Table 1: Training data size: Number of parallel sentences for all language pairs.

data to the target language and translating the pivot side of the pivot-target data to the source language (Park et al., 2017).

2. Translating pivot monolingual sentences into both source and target languages (Currey and Heafield, 2019).

Yang et al. (2022) used sequence-level knowledge distillation to generate a distilled source-target dataset using three teacher models, which is then used to train a multilingual NMT model. Chen et al. (2017) used word-level KD in guiding a source-to-target model (student) through a pivot-to-target model (teacher). One reason for the improvements from using synthetic data is that some synthetic sentences are actually of higher quality than their original counterparts (Briakou and Carpuat, 2022).

In this work we propose replacing the original pivot sentences with synthetic ones with either of the pivoting paradigms. This enables us to validate the hypothesis that regenerating pivot sentences to have a structure closer to that of the source and target languages will improve pivot-based translation. We regenerate pivot sentences using knowledge distillation with autoregressive and non-autoregressive translation models as potential ways of restructuring the pivot sentences.

2.3. Related Work

In recent years, there have been several attempts to improve NMT for low-resource languages through multilingual training (Firat et al., 2016; Lakew et al., 2018; Neubig and Hu, 2018). Kumar et al. (2021); Neubig and Hu (2018) proposed frameworks for adapting the existing NMT systems to new low-resourced languages. Zhang et al. (2021) used transfer learning to adapt NMT models from translating between high-resource languages to low-resources. Moreover, Kumar et al. (2021); Lakew et al. (2018) fine-tuned NMT models on self-generated data, which led to greater performance gains in low-resource and zero-resource directions.

eng	How did the light wave travel through air?
zul	Igagasi lokukhanya lalihamba kaniani emoyeni?
xho	Iliza lokukhanya lalihamba niani emoyeni?

Figure 2: Example word alignment between English, Xhosa, and Zulu.

3. Experimental Setup

Data As our approach is aimed at translating between closely related low-resource languages, we evaluate on two pairs of such languages. We select Xhosa (xho) and Zulu (zul) from the Nguni Language family and Sepedi (nso) and Tswana (tsn) from the Sotho-Tswana language family. We evaluate translation in both directions, and select English (eng) as pivot language for both language pairs due to parallel data availability. An example of word alignment between English, Xhosa and Zulu is given in Figure 2.

For training and validation, we use a subset of WMT22_african.[1] Table 1 shows available number of sentences for each language pair. We reserve the first 3000 sentences from each language pair for validation and the rest for training. We use the Flores dev set to evaluate the performance of different regeneration methods for pivoting.[2] It contains 997 parallel sentences for each language pair. Additionally, we report the results of the best translation systems as evaluated on the Flores devtest set, which contains 1012 parallel sentences for each language pair.

Vocabulary The low-resource languages we are working with are agglutinative. The Nguni languages are written conjunctively, meaning that words may consist of multiple morphemes without separation. As a preprocessing step for these languages (Xhosa and Zulu) we use a combination of BPE and supervised morphological segmentation with CRFs (Moeng et al., 2021) (see appendix A). Sepedi and Tswana are disjunctive (morphemes are space-separated), so we use BPE only.

For the translation between Xhosa and Zulu, we train a multilingual vocabulary on the $eng - xho$, $eng - zul$, and $xho - zul$ datasets, with a vocab size of $30K$. We use the same vocabulary for training all models in the pivot-based approaches. We follow the same approach with translation between Sepedi and Tswana.

Model Architecture All models were trained with the Fairseq toolkit (Ott et al., 2019). We

[1] https://huggingface.co/datasets/allenai/wmt22_african

[2] https://huggingface.co/datasets/facebook/flores

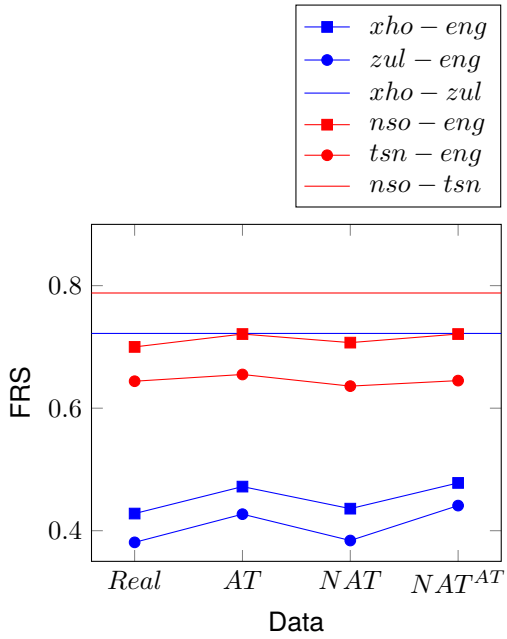

Figure 3: Fuzzy Reordering Score (FRS) of different synthetic datasets compared with real data among different language pairs. A larger score indicates more monotonic alignments.

used the `transformer-base` architecture (Vaswani et al., 2017) for training all AT models and the `Levenshtein Transformer` (Gu et al., 2019) for training NAT models.

We used the following hyper-parameters for all models: Adam optimizer with learning rate = $5e-4$, $\beta_1 = 0.9$ and $\beta_2 = 0.98$, inverse square root learning rate scheduler with warmup updates of $4,000$ for AT models and $10,000$ for NAT models, label smoothed cross entropy criterion for AT models and `nat_loss` for NAT models with label smoothing of 0.1 and dropout of 0.3. All models were trained on either an `Nvidia A100` full card ($40GB$) or a division of half a card ($20GB$) for 45 epochs with a batch size of 12288 tokens for AT models and 8192 tokens for NAT models. We select the best checkpoint based on validation loss.

Multilingual Model Additionally, we perform some experiments using a multilingual NMT model to generate synthetic data or to perform pivoting. We use the multilingual model of Elmadani et al. (2022) that was trained to translate to and from 8 Southern African languages, including all 4 languages we consider in this paper. However it was trained to translate directly between only one of the four language pairs, so we also consider fine-tuning the model with direct translation data between these language pairs as an additional baseline.

4. Synthetic Pivot Sentence Generation

4.1. Notation

We start by defining the general notation used for models and datasets. Given two languages, A and B, let $A_B - B_A$ represent the parallel dataset between the two languages, where A_B and B_A denote the language A and language B sides of the bitext, respectively. Let $A \xrightarrow{AT} B$ and $A \xrightarrow{NAT} B$ be the A to B autoregressive and non-autoregressive translation models, respectively, trained on the $A_B - B_A$ dataset. B_A^{AT} denotes the distilled B_A sentences generated from model $A \xrightarrow{AT} B$. A NAT model trained on dataset $A_B - B_A^{AT}$ with synthetic target-side data is represented as $A \xrightarrow{NAT} B^{AT}$.

4.2. Approach

Let S, T, and P be the source, target, and pivot languages, respectively. Our goal is to translate between the two languages $S \leftrightarrow T$, where S and T are closely related languages.

We aim to regenerate P sentences to have a similar structure to S and T. Our approach is divided into two steps: First, synthetic pivot generation, where we use AT and NAT models to regenerate P sentences from S and T. The second step is to replace the actual pivot sentences with synthetic ones and use them in the pivoting approaches from §2.2.

Given datasets $S_P - P_S$ and $T_P - P_T$, we regenerate P_S and P_T sentences using the three types of translation models. We pass S_P and T_P sentences to all models to generate three synthetic versions of P_S and P_T:

1. P_S^{AT} and P_T^{AT} are generated from $S \rightarrow P$ and $T \rightarrow P$ autoregressive models trained on real data.
2. P_S^{NAT} and P_T^{NAT} are generated from $S \rightarrow P$ and $T \rightarrow P$ non-autoregressive models trained on real data.
3. $P_S^{NAT^{AT}}$ and $P_T^{NAT^{AT}}$ are generated from $S \rightarrow P$ and $T \rightarrow P$ non-autoregressive models trained on synthetic data generated from autoregressive models.

Synthetic pivot sentence generation using the multilingual NMT model follows the same approach, except that since we don't have a multilingual NAT model, some NAT model combinations are not applicable here.

4.3. Reordering evaluation

Figure 3 shows the Fuzzy Reordering Score (FRS) (Talbot et al., 2011) of the real data and the

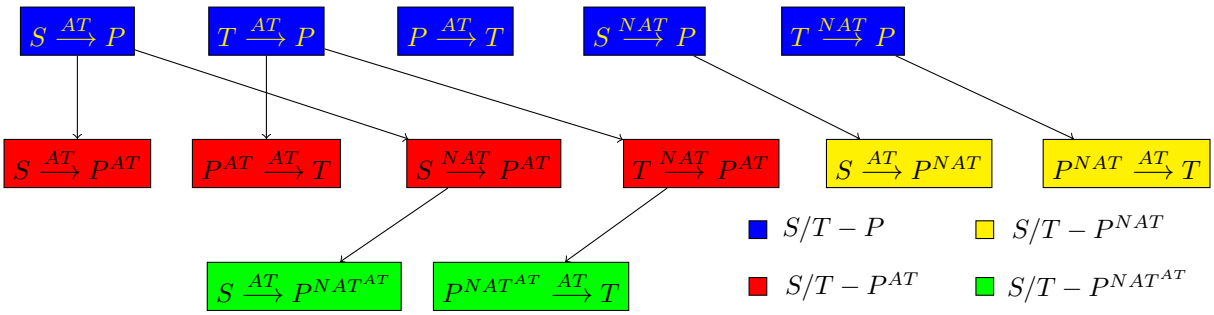

Figure 4: The procedure we took for training the $S \rightarrow P$ and $P \rightarrow T$ models. The arrows indicate generating P synthetic sentences from a model and using them to train another model. The colour scheme indicates the data we used for training the model.

three synthetic datasets among different language pairs. It also shows the FRS of direct translation datasets, $xho - zul$ and $nso - tsn$. We use FRS to measure the change in word order during synthetic data generation: A larger score indicates more monotonic alignments between the two languages. We used the `fast_align` library (Dyer et al., 2013) for generating the alignments used to compute the FRS (see appendix B for details). The figure shows that similar languages have similar FRS with English among different datasets. The FRS between similar languages is always higher than the FRS between any of these languages and English. Furthermore, the auto-regressively generated data has less reordering than real data and non-auto-regressively generated data. Zhou et al. (2021) argued that NAT models need training data with more monotonic alignments to perform better because it is hard for these models to learn complicated alignments. However, our results show that sentences generated from NAT models have *fewer* monotonic alignments compared to the ones generated by AT models.

5. Synthetic Cascading

5.1. Approach

In the cascading approach, source sentences are translated to the pivot language through a source-to-pivot ($S \rightarrow P$) model, which is then in turn translated to the target language through a pivot to target ($P \rightarrow T$) model. The $S \rightarrow P$ and $P \rightarrow T$ models are trained separately. To train either, the real dataset and three synthetic datasets from §4 are available; for each dataset we can train either AT or NAT model. That give 8 sets of $S \rightarrow P$ models and 8 sets of $P \rightarrow T$ models, leading to 64 possible combinations. However, for $S \rightarrow P$, we did not train NAT models on synthetic datasets generated from another NAT models (P_S^{NAT} and $P_S^{NAT^{AT}}$). For $P \rightarrow T$, we did not train any NAT models. We ended up with 6 $S \rightarrow P$ and 4 $P \rightarrow T$ models (24 combinations). Figure 4 shows which dataset was used in training each of these models. The arrows between models indicate that the top model was used to generate the pivot side of the training data for the model at the lower level.

We perform cascaded translation by using the $S \rightarrow P$ model to translate the S side of the test set (S_T) to P. This output is translated to T using the $P \rightarrow T$ model. The overall translation quality of a system depends on two factors: the performance of individual $S \rightarrow P$ and $P \rightarrow T$ models, and the degree of matching between the $S \rightarrow P$ and $P \rightarrow T$ models. We are particularly interested in the robustness of $P \rightarrow T$ models to the input they are given. Therefore we say that a *robust* $P \rightarrow T$ model can maintain its performance regardless of how the $S \rightarrow P$ model was trained.

5.2. Results

Table 2 reports the results of the cascading approach for translation between Xhosa and Zulu and between Tswana and Sepedi. Both the results of the individual source to pivot and pivot to target models, and for the cascaded source to target translation are given. Due to space limitations we don't report the results of all model combinations: we report the results of the combination corresponding to each source-to-pivot model, and the average performance for each of the pivot-to-target models. The latter reflects the relative robustness of the different pivot-to-target models to pivot sentences generated by the various input models. In other words, it better reflects the ability of a back-translation model to adapt to noise on the input side - no matter what the source of the input pivot sentence is, it can translate to the target language robustly. Moreover, the average score informs us which Pivot-to-Source/Target model to choose in order to translate the Pivot sentences into the Source and Target languages in the Synthetic Direct Translation approach. See tables 6, 7, 8 and 9 in the appendix for the full cascading results.

	xho → zul		tsn → nso		zul → xho		nso → tsn	
	pivot	cascade	pivot	cascade	pivot	cascade	pivot	cascade
$\mathbf{S} \xrightarrow{\mathbf{AT}} \mathbf{P}$	**26.1**	**9.2**	**11.8**	**5.6**	**28.4**	8.7	**14.4**	**8.4**
$\mathbf{S} \xrightarrow{\mathbf{AT}} \mathbf{P^{AT}}$	24.0	8.7	8.9	5.1	26.6	8.3	7.3	4.7
$\mathbf{S} \xrightarrow{\mathbf{NAT}} \mathbf{P}$	18.7	7.7	6.4	3.7	18.5	5.7	6.9	4.9
$\mathbf{S} \xrightarrow{\mathbf{AT}} \mathbf{P^{NAT}}$	20.3	7.6	5.5	3.5	20.7	6.9	5.0	4.0
$\mathbf{S} \xrightarrow{\mathbf{NAT}} \mathbf{P^{AT}}$	23.4	8.8	2.6	2.4	25.3	8.2	3.7	3.7
$\mathbf{S} \xrightarrow{\mathbf{AT}} \mathbf{P^{NAT^{AT}}}$	22.8	8.7	0.6	0.5	25.9	**8.8**	1.2	1.1
$\mathbf{P} \xrightarrow{\mathbf{AT}} \mathbf{T}$	15.2	7.8	**8.5**	2.6	12.2	7.2	10.1	3.3
$\mathbf{P^{AT}} \xrightarrow{\mathbf{AT}} \mathbf{T}$	**15.8**	8.3	7.6	3.3	**12.7**	7.8	**13.2**	4.0
$\mathbf{P^{NAT}} \xrightarrow{\mathbf{AT}} \mathbf{T}$	13.4	8.3	6.4	3.2	11.1	7.3	10.3	4.3
$\mathbf{P^{NAT^{AT}}} \xrightarrow{\mathbf{AT}} \mathbf{T}$	14.2	8.4	6.3	3.1	12.2	7.6	6.7	4.0

Table 2: Pivot translation results with bilingual models (dev set BLEU scores). Results are reported for both the individual pivot models (Source to Pivot and Pivot to Target) and the full cascade (Source to Target), with English as pivot language. For the cascaded results we report the *best* result for each source to pivot model (out of all the pivot to target models), as well as the *average* result when using each of the pivot to target models (over all the source to pivot models).

We find that the combination of the best source-pivot and pivot-target models often but no always lead to the best cascaded translation performance. The source-to-pivot model trained on the original data almost aways lead to the best cascaded performance as well, but for the choice of pivot-to-target models the picture is more mixed: The back-translation model $\mathbf{P^{AT}} \xrightarrow{\mathbf{AT}} \mathbf{T}$ usually results in the best performance, even though in some cases the model trained on original data has higher pivot-to-target performance, and the $\mathbf{P^{NAT}} \xrightarrow{\mathbf{AT}} \mathbf{T}$ models are more robust for some language pairs. Where the structure of the pivot sentences are changed more drastically by the source-pivot model, the matching in model type between the $S \rightarrow P$ and $P \rightarrow T$ models becomes a bigger factor in explaining the overall performance.

Synthetic Training Data and Model Robustness A potential explanation of the results is the degree of robustness in each $P \rightarrow T$ model. Regenerating English sentences from a translation model can add noise (Edunov et al., 2018). Back-translation models are more tolerant to source-side noise due to the noise added to the training data. If the quality of the generated data is good enough, the back-translation model can benefit from both robustness and an increase in individual performance.

Multilingual model results As an additional experiment we use the multilingual autoregressive model to generate synthetic pivot sentences. We only consider translation between Sepedi and Tswana, as the bilingual model results indicate that this relatively lower-resourced language pair benefits more from synthetic pivoting than translation between Xhosa and Zulu. The results are given in Table 3, only considering translation between Sepedi and Tswana. We experiment both with using the multilingual model directly for the cascaded generation ($S \xrightarrow{AT} P(m)$ and $P \xrightarrow{AT} T(m)$), and using synthetic pivot data generated by the multilingual model to train bilingual models for the cascade. The results show that using the multilingual model to perform both $S \rightarrow P$ and $P \rightarrow T$ outperforms all synthetic cascading systems in overall system performance. We hypothesize that synthetic pivoting does not help improve multilingual cascading systems due to the lack of transferability of multilingual

	tsn → nso		nso → tsn	
	pivot	cascade	pivot	cascade
$\mathbf{S} \xrightarrow{\mathbf{AT}} \mathbf{P(m)}$	**20.3**	**14.4**	**26.9**	**13.9**
$\mathbf{S} \xrightarrow{\mathbf{AT}} \mathbf{P^{AT}}$	15.2	13.3	19.4	12.5
$\mathbf{S} \xrightarrow{\mathbf{NAT}} \mathbf{P^{AT}}$	9.1	8.8	10.4	7.6
$\mathbf{S} \xrightarrow{\mathbf{AT}} \mathbf{P^{NAT^{AT}}}$	9.5	8.9	11	7.9
$\mathbf{P} \xrightarrow{\mathbf{AT}} \mathbf{T(m)}$	**23.1**	11.2	**18.2**	10.3
$\mathbf{P^{AT}} \xrightarrow{\mathbf{AT}} \mathbf{T}$	18	10.5	13.9	10.1
$\mathbf{P^{NAT^{AT}}} \xrightarrow{\mathbf{AT}} \mathbf{T}$	11.6	8.7	11.5	9.4

Table 3: Pivot translation results with synthetic data from an (autoregressive) multilingual model (dev set BLEU scores). Results are reported in the same format as Table 2.

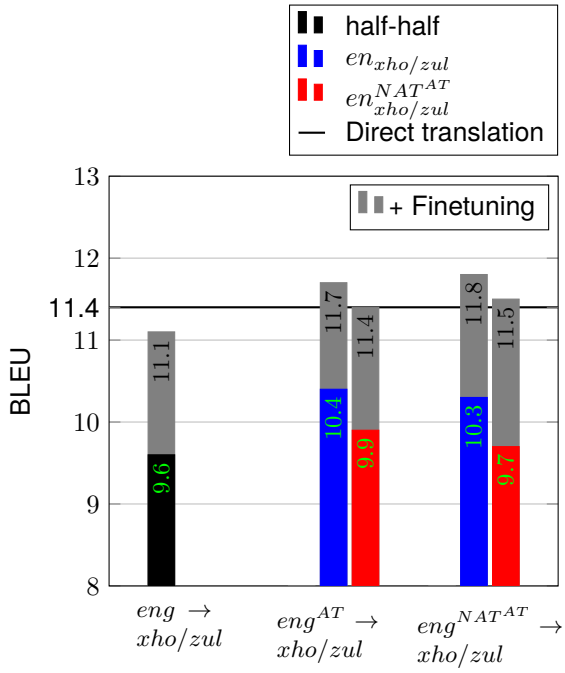

Figure 5: BLEU scores of different $xho \rightarrow zul$ models for synthetic direct translation.

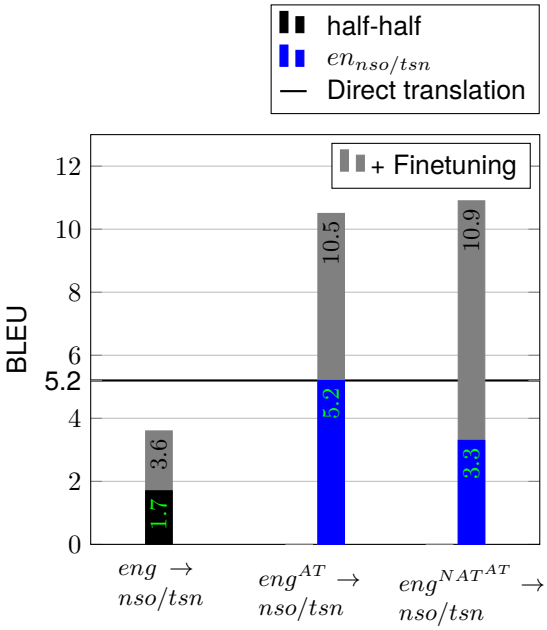

Figure 6: BLEU scores of different $tsn \rightarrow nso$ models for synthetic direct translation.

information. Generating monolingual sentences from a multilingual model is not enough to obtain multilinguality improvements.

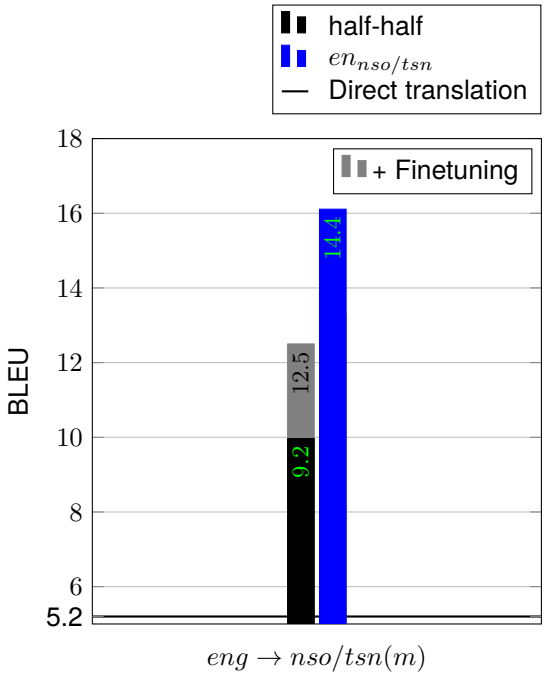

Figure 7: BLEU scores of $tsn \rightarrow nso$ models, using a multilingual model for synthetic direct translation.

6. Synthetic Direct Translation

6.1. Approach

We extend the pivot-based approach proposed by Park et al. (2017), who proposed generating synthetic $S-T$ sentences from P. Park et al. (2017) used $P \xrightarrow{AT} S$ to translate the pivot side of $P_T - T_P$ to the source language and $P \xrightarrow{AT} T$ to translate the pivot side of $P_S - S_P$ to the target language. This generates an $S-T$ dataset with the size of $S_P - P_S$ and $P_T - T_P$ combined. In this dataset, half of the sentences on each side are synthetic, while the other half is real; both the source and target sides contain a combination of real and synthetic sentences. We call this approach *half-half*.

We propose translating all pivot sentences ($P_S + P_T$) to *both* source and target languages using different types of $P \rightarrow S/T$ models. Additionally, the pivot sentences to be translated may themselves be synthetic (following §4). We refer to this as *synthetic-only* pivoting. We select the synthetic data generation models based on the cascading results: We choose the best AT-based $P \rightarrow T$ model and the best NAT-based $P \rightarrow T$ model based on robustness. The direct translation models trained on synthetic data only can then be fine-tuned on available real $S-T$ sentences.

6.2. Results

Xhosa to Zulu Figure 5 shows the performance of the synthetic-only pivoting approach compared

to the half-half approach on $xho \rightarrow zul$ translation. The first bar represents the half-half baseline of Park et al. (2017). We use two sources of eng sentences, real sentences and $eng^{NAT^{AT}}$ sentences generated from the $xho, zul \xrightarrow{NAT} eng^{AT}$ models. Then, we use the selected $eng \rightarrow xho, zul$ models to translate these sentences from both sources to xho and zul. In the figure, the bars' colours represent the source of the eng sentences used to generate xho and zul synthetic data, while the x-axis shows the model used for translating eng pivot sentences. For example, for the result in the first blue all real eng sentences ($en_{xho} + en_{zul}$) were translated to both xho and zul, using the $eng^{AT} \rightarrow xho$ and $eng^{AT} \rightarrow zul$ models, respectively. As a final step we fine-tune all models on real $xho - zul$ data (see figure 8 in the appendix for $zul \rightarrow xho$ translation).

The results show that the performance of all the models that use synthetic pivot sentences are relatively similar; neither the choice of eng data sources nor of the synthetic data generation models have a substantial impact. However, these models are slightly better than the baseline approach. The gains from fine-tuning are also similar across models. However, using real eng sentences for translation is generally better regardless of the type of the translation model. Moreover, none of the fine-tuned models outperformed the direct translation baseline with more than 0.4 BLEU points. We argue that the translation between xho and zul does not reflect truely low-resource translation scenarios, as the available bitext between the two languages includes more than $1M$ pair of sentences. Although it might not seem to be a lot of data, the fact that the two languages are structurally similar supports the hypothesis that less data would be needed to train a translation model from scratch.

Tswana to Sepedi The translation from Xhosa to Zulu reflects the importance of the synthetic data in the model's robustness but not in the overall translation quality of pivot-based systems. Figure 6 shows the performance of the synthetic data generation approach for $tsn \rightarrow nso$ translation, where much less direct bitext is available. All models obtain very large gains after fine-tuning. The fine-tuned models also surpass direct translation by a large margin (in contrast to Xhosa to Zulu translation). Therefore, lower-resource languages can benefit more from pivot-based approaches than higher-resource languages. We also see that the baseline half-half approach performs poorly compared to our approaches that use fully synthetic $nso - tsn$ parallel data (see figure 9 in the appendix for $nso \rightarrow tsn$ translation).

Multilingual Synthetic Direct Translation We also experimented with synthetic direct translation using the multilingual model. Again only translation between Tswana and Sepedi is considered. Figure 7 shows the performance of the synthetic-only pivoting approach compared to the half-half approach on translating from Tswana to Sepedi (see figure 10 in the appendix for $nso \rightarrow tsn$ translation). We find that synthetic-only pivoting outperforms half-half in Tswana to Sepedi translation. However, in this case fine-tuning the synthetic-only pivoting model on real data *harms* translation quality, suggesting that the synthetic data is more informative than the real data.

7. Final Results

Table 4 shows the performance of baselines and the best models trained using synthetic pivot data. All approaches are evaluated on Flores `devtest` set. The best performance in all translation directions (excluding the multilingual models) is obtained by translating all pivot sentences to source and target languages using back-translation models, followed by training the $S \rightarrow T$ model on the generated data and fine-tuning on real data. The results also confirm that the translation between Sepedi and Tswana benefits from including synthetic data more than the translation between Xhosa and Zulu.

We compare approaches using the multilingual model Elmadani et al. (2022) separately, as training on a large number of language pairs lead to complementary gains. The multilingual direct translation model was fine-tuned on direct translations between all 4 language pairs. For $xho \leftrightarrow zul$ our best synthetic pivoting approach outperforms the multilingual models. However on $nso \leftrightarrow tsn$ all the multilingual approaches outperform our best synthetic pivoting approach.

8. Conclusion

This paper investigated strategies to improve pivot-based NMT systems using synthetic pivot data. Training with synthetic data reduces the complexity of learning the pivot language by changing the structure of pivot sentences to be closer to the source or target languages. In our experiments we used the real sentences along with pivot sentences and synthetic datasets generated from different types of NMT models. The results indicate that synthetic-only pivoting can benefit from synthetic pivot data more than cascading, and the largest gains are obtained in the lowest resource settings.

Model	xho → zul		zul → xho		nso → tsn		tsn → nso	
	BLEU	CHRF2	BLEU	CHRF2	BLEU	CHRF2	BLEU	CHRF2
Bilingual Direct Translation	12.1	49.5	11.7	50.4	5.3	29.2	4.3	27.1
Cascading	9.0	45.2	8.9	45.0	6.3	29.6	3.8	25.1
Synthetic Cascading	9.0	45.7	9.4	46.6	7.6	32.1	4.6	27.9
Synthetic Direct Translation								
half-half	10.1	46.8	10.1	47.6	0.8	14.3	1.6	19.3
+fine-tuning	11.7	49.7	11.4	49.5	2.0	20.6	3.0	24.4
Synthetic-only pivoting	10.5	48.9	10.6	50.0	5.6	29.2	4.2	26.4
+fine-tuning	**12.4**	**50.2**	**11.9**	**50.5**	**10.9**	**36.6**	**9.7**	**36.3**
Multilingual Direct Translation	11.7	50.1	11.7	50.7	15.2	42.9	15.4	44.4
Multilingual Cascading	10.6	48.7	10.6	50.0	13.1	40.5	13.4	40.5
Multilingual Synthetic Direct					14.6	41.9	12.6	41.3

Table 4: Final translation results of the pivoting approaches evaluated on the Flores `devtest` set. For each of our synthetic approaches the best model combinations were chosen based on their BLEU scores on the Flores `dev` set.

9. Limitations

In this paper, we only investigated the case of translating between similar languages. However, our results suggest that while language similarity plays a role, it might not be as crucial as hypothesized and therefore pivot-based models for translating between unrelated languages might also benefit from this approach. We do not consider using additional monolingual pivot data, which could potentially improve system performance further. Our results show that the performance of the pivot-based approaches highly depends on the quality of the NMT model used for synthetic pivot data generation. This will make it hard to apply our approach in extremely low-resource scenarios where only a small amount of parallel data between low-resource languages and pivot language is available.

10. Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number: 129850). Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: `hpc.uct.ac.za`.

11. Bibliographical References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Eleftheria Briakou and Marine Carpuat. 2022. Can synthetic translations improve bitext quality? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4753–4766, Dublin, Ireland. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong. Association for Computational Linguistics.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2020. Understanding and improving lexical choice in non-autoregressive translation. *CoRR*, abs/2012.14583.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Khalid N. Elmadani, Francois Meyer, and Jan Buys. 2022. University of Cape Town's WMT22 system: Multilingual machine translation for southern african languages.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11179–11189. Curran Associates, Inc.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Surafel Melaku Lakew, Quintino Francesco Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Improving zero-shot translation of low-resource languages. In *International Workshop on Spoken Language Translation*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement.

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2021. Canonical and surface morphological segmentation for nguni languages.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Jaehong Park, Jongyoon Song, and Sungroh Yoon. 2017. Building a neural machine translation system using only synthetic parallel data.

Mitchell Stern, William Chan, Jamie Ryan Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *International Conference on Machine Learning*.

David Talbot, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Masakazu Seno, and Franz Och. 2011. A lightweight evaluation framework for machine translation reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 12–21, Edinburgh, Scotland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. UM4: Unified multilingual multiple teacher-student model for zero-resource neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4454–4460. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Meng Zhang, Liangyou Li, and Qun Liu. 2021. Two parents, one child: Dual transfer for low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2726–2738, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2021. Understanding knowledge distillation in non-autoregressive machine translation.

A. Combining BPE and Morphological Segmentation

Although CRF models can detect morphemes more accurately than BPE, we cannot rely solely on morphological segmentation models for preprocessing and vocabulary creation; morphological segmentation alone does not deal with the out of vocabulary token problem. We propose running BPE after surface segmentation to ensure accurate morphemes and a finite vocabulary without unknown tokens (UNK). However, this approach requires additional processing ensure that morpheme and subword boundaries and handled consistently.

The BPE algorithm divides the words into different space-separated sub-words. Then, it adds the end-of-word token (</w>) at the end of the last subword of each word. Adding this token aims to differentiate between the sub-words of different words, making it possible to recover the original text from the encoded text. Now we want to combine two segmentation strategies; morphological surface segmentation and then BPE. The straightforward solution is to encode the text in two phases. The first phase would add </w> tokens to distinguish between different words, while the second phase would add </m> to distinguish between morphemes. The decoding would also be done in two stages: remove </m>, then </w>. However, treating the </w> token as a regular token would result in splitting it into sub-tokens in the second stage (< / w >), which would lead to increasing the sequence length with meaningless tokens.

Our proposed solution is to remove the </w> token after the surface segmentation phase (first phase) after recording which morpheme was at the end of each word. Then we run BPE on the modified text, which does not contain a token to differentiate between the end of morpheme and the end of word. In this case, BPE adds </w> at the end of both words and morphemes. Then, we use the recordes information to keep </w> only at the end of words and remove it from the end of morphemes. Since we use one unique </w> token to indicate the end of a word, it is possible to do the decoding in one stage by only removing this token.

Table 5 shows the steps for combining Surface Segmentation with BPE. First, we segment the Xhosa sentence using Surface Segmentation. Then, we use a binary array to indicate if the morpheme is at the end of a word. We then remove the </w> tokens and train the BPE tokenizer on all segmented Xhosa and Zulu sentences and all original English sentences. After training and running BPE on all sentences, we use the saved binary array to filter out </w> tokens from the end of morphemes that are not at the end of their words.

B. Fuzzy Reordering Score (FRS)

The fuzzy reordering score (FRS) measures the structural similarity of two languages. This score is computed using an algorithm proposed by Talbot et al. (2011). It takes as input the parallel data between two languages and the alignments between source and target sentences.

We used the `fast_align` library (Dyer et al., 2013) for generating the alignments from parallel data. We trained the alignments for the real parallel data normally and used the `grow-diag-final` symmetrization heuristic. For the datasets that include a synthetic pivot side, we used the same alignment model trained for the corresponding real version of the dataset. For example, we used the alignment model trained with $xho_{eng} - eng_{xho}$ to produce the alignments for $xho_{eng} - eng^{AT}_{xho}$, $xho_{eng} - eng^{NAT}_{xho}$, and $xho_{eng} - eng^{NAT^{AT}}_{xho}$. We used `force_align` with `grow-diag-final` symmetrization heuristic. The alignments are extracted after tokenizing the datasets.

	Xhosa	English
sentence	Qwalasela umbuzo ongezantsi	Consider the question below
segment	Qwalasel a</w> u m buzo</w> o ng e zantsi</w>	
save ends	0 1 0 0 1 0 0 0 1	
remove ends	Qwalasel a u m buzo o ng e zantsi	
Train BPE		
BPE	Qwa lasel</w> a</w> u</w> m</w> buzo</w> o</w> ng</w> e</w> zant si</w>	Cons id er</w> the</w> question</w> be low</w>
filter	Qwa lasel a</w> u m buzo</w> o ng e zant si</w>	

Table 5: Preprocessing example: combining surface segmentation with BPE

	xho → zul		AT-based		NAT-based	
			eng $\xrightarrow{AT}$ zul	eng^{AT} $\xrightarrow{AT}$ zul	eng^{NAT} $\xrightarrow{AT}$ zul	eng$^{NAT^{AT}}$ $\xrightarrow{AT}$ zul
			15.2	15.8	13.4	14.2
AT-based	xho $\xrightarrow{AT}$ eng	26.1	8.7	**9.2**	8.7	**9.2**
	xho $\xrightarrow{AT}$ eng^{AT}	24.0	8.1	**8.7**	8.3	**8.7**
NAT-based	xho $\xrightarrow{NAT}$ eng	18.7	7.0	7.3	**7.7**	7.4
	xho $\xrightarrow{AT}$ eng^{NAT}	20.3	7.0	7.4	**7.6**	**7.6**
	xho $\xrightarrow{NAT}$ eng^{AT}	23.4	8.1	8.5	**8.8**	**8.8**
	xho $\xrightarrow{AT}$ eng$^{NAT^{AT}}$	22.8	8.1	8.6	8.4	**8.7**
	avg		7.83	8.28	8.25	8.4

Table 6: BLEU scores of 24 $xho \rightarrow zul$ cascading translation systems. The first column of the table represents $xho \rightarrow eng$ models, while the first row of the table represents $eng \rightarrow zul$. The second column and the second row include the individual performance of $xho \rightarrow eng$ and $eng \rightarrow zul$ models, respectively. **bold** represents the best performance for each $xho \rightarrow eng$ model, while underline represents the best performance for each $eng \rightarrow zul$ model. The colour scheme is the same as in figure 4; models that translate between the same language pair and have the same colour were trained on the same dataset.

	zul → xho		AT-based		NAT-based	
			eng $\xrightarrow{AT}$ xho	eng^{AT} $\xrightarrow{AT}$ xho	eng^{NAT} $\xrightarrow{AT}$ xho	eng$^{NAT^{AT}}$ $\xrightarrow{AT}$ xho
			12.2	12.7	11.1	12.2
AT-based	zul $\xrightarrow{AT}$ eng	28.4	8.0	**8.7**	7.8	8.2
	zul $\xrightarrow{AT}$ eng^{AT}	26.6	7.8	**8.3**	7.7	8.2
NAT-based	zul $\xrightarrow{NAT}$ eng	18.5	5.6	**5.7**	**5.7**	5.6
	zul $\xrightarrow{AT}$ eng^{NAT}	20.7	6.5	**6.9**	**6.9**	6.8
	zul $\xrightarrow{NAT}$ eng^{AT}	25.3	7.5	**8.2**	7.6	8.0
	zul $\xrightarrow{AT}$ eng$^{NAT^{AT}}$	25.9	8.0	**8.8**	8.2	8.6
	avg		7.23	7.76	7.32	7.56

Table 7: BLEU scores of 24 $zul \rightarrow xho$ cascading translation systems. the first column of the table represents $zul \rightarrow eng$ models, while the first row of the table represents $eng \rightarrow xho$. The second column and the second row include the individual performance of $zul \rightarrow eng$ and $eng \rightarrow xho$ models, respectively. **bold** represents the best performance for each $zul \rightarrow eng$ model, while underline represents the best performance for each $eng \rightarrow xho$ model. The colour scheme is the same as in figure 4; models that translate between the same language pair and have the same colour were trained on the same dataset.

nso → tsn			AT-based		NAT-based	
			eng $\xrightarrow{AT}$ tsn	eng^{AT} $\xrightarrow{AT}$ tsn	eng^{NAT} $\xrightarrow{AT}$ tsn	eng$^{NAT^{AT}}$ $\xrightarrow{AT}$ tsn
			10.1	13.2	10.3	6.7
AT-based	nso $\xrightarrow{AT}$ eng	14.4	6.4	**8.4**	7.8	6.0
	nso $\xrightarrow{AT}$ eng^{AT}	7.3	3.8	4.4	**4.7**	4.6
NAT-based	nso $\xrightarrow{NAT}$ eng	6.9	3.7	4.3	**4.9**	4.6
	nso $\xrightarrow{AT}$ eng^{NAT}	5.0	3.0	3.3	3.9	**4.0**
	nso $\xrightarrow{NAT}$ eng^{AT}	3.7	2.5	2.9	3.4	**3.7**
	nso $\xrightarrow{AT}$ eng$^{NAT^{AT}}$	1.2	0.5	0.7	0.9	**1.1**
avg			3.32	4.00	4.27	4.00

Table 8: BLEU scores of 24 $nso \rightarrow tsn$ cascading translation systems. the first column of the table represents $nso \rightarrow eng$ models, while the first row of the table represents $eng \rightarrow tsn$. The second column and the second row include the individual performance of $nso \rightarrow eng$ and $eng \rightarrow tsn$ models, respectively. **bold** represents the best performance for each $nso \rightarrow eng$ model, while underline represents the best performance for each $eng \rightarrow tsn$ model. The colour scheme is the same as in figure 4; models that translate between the same language pair and have the same colour were trained on the same dataset.

tsn → nso			AT-based		NAT-based	
			eng $\xrightarrow{AT}$ nso	eng^{AT} $\xrightarrow{AT}$ nso	eng^{NAT} $\xrightarrow{AT}$ nso	eng$^{NAT^{AT}}$ $\xrightarrow{AT}$ nso
			8.5	7.6	6.4	6.3
AT-based	tsn $\xrightarrow{AT}$ eng	11.8	4.5	**5.6**	4.8	4.7
	tsn $\xrightarrow{AT}$ eng^{AT}	8.9	4.0	**5.1**	4.6	4.7
NAT-based	tsn $\xrightarrow{NAT}$ eng	6.4	2.7	3.5	**3.7**	3.4
	tsn $\xrightarrow{AT}$ eng^{NAT}	5.5	2.4	3.3	3.4	**3.5**
	tsn $\xrightarrow{NAT}$ eng^{AT}	2.6	1.5	2.0	**2.4**	2.2
	tsn $\xrightarrow{AT}$ eng$^{NAT^{AT}}$	0.6	0.2	0.3	**0.5**	0.3
avg			2.55	3.3	3.23	3.13

Table 9: BLEU scores of 24 $tsn \rightarrow nso$ cascading translation systems. The first column of the table represents $tsn \rightarrow eng$ models, while the first row of the table represents $eng \rightarrow nso$. The second column and the second row include the individual performance of $tsn \rightarrow eng$ and $eng \rightarrow nso$ models, respectively. **bold** represents the best performance for each $tsn \rightarrow eng$ model, while underline represents the best performance for each $eng \rightarrow nso$ model. The colour scheme is the same as in figure 4; models that translate between the same language pair and have the same colour were trained on the same dataset.

nso → tsn			AT-based		NAT-based
			eng $\xrightarrow{AT}$ tsn(m)	eng^{AT} $\xrightarrow{AT}$ tsn	eng$^{NAT^{AT}}$ $\xrightarrow{AT}$ tsn
			18.2	13.9	11.5
AT-based	nso $\xrightarrow{AT}$ eng(m)	26.9	**13.9**	12.9	11.0
	nso $\xrightarrow{AT}$ eng^{AT}	19.4	**12.5**	12.2	11.3
NAT-based	nso $\xrightarrow{NAT}$ eng^{AT}	10.4	7.5	**7.6**	**7.6**
	nso $\xrightarrow{AT}$ eng$^{NAT^{AT}}$	11.0	7.4	7.8	**7.9**
avg			10.3	10.1	9.4

Table 10: BLEU scores of 12 $nso \rightarrow tsn$ cascading translation systems, with pivot data generated by a multilingual model. The first column of the table represents $nso \rightarrow eng$ models, while the first row of the table represents $eng \rightarrow tsn$. The second column and the second row include the individual performance of $nso \rightarrow eng$ and $eng \rightarrow tsn$ models, respectively. (m) indicates that we are using the multilingual model for this translation. **bold** represents the best performance for each $nso \rightarrow eng$ model, while underline represents the best performance for each $eng \rightarrow tsn$ model. The colour scheme is the same as in figure 4: Models that translate between the same language pair and have the same colour were trained on the same dataset.

tsn → nso			AT-based		NAT-based
			eng $\xrightarrow{\text{AT}}$ nso(m)	eng$^{\text{AT}}$ $\xrightarrow{\text{AT}}$ nso	eng$^{\text{NAT}^{\text{AT}}}$ $\xrightarrow{\text{AT}}$ nso
			23.1	18.0	11.6
AT-based	tsn $\xrightarrow{\text{AT}}$ eng(m)	20.3	**14.4**	12.4	9.3
	tsn $\xrightarrow{\text{AT}}$ eng$^{\text{AT}}$	15.2	**13.3**	11.8	9.1
NAT-based	tsn $\xrightarrow{\text{NAT}}$ eng$^{\text{AT}}$	9.1	8.4	**8.8**	8.0
	tsn $\xrightarrow{\text{AT}}$ eng$^{\text{NAT}^{\text{AT}}}$	9.5	8.6	**8.9**	8.6
avg			11.2	10.5	8.7

Table 11: BLEU scores of 12 $tsn \rightarrow nso$ cascading translation systems, with pivot data generated by a multilingual model. The first column of the table represents $tsn \rightarrow eng$ models, while the first row of the table represents $eng \rightarrow nso$. The second column and the second row include the individual performance of $tsn \rightarrow eng$ and $eng \rightarrow nso$ models, respectively. (m) indicates that we are using the multilingual model for this translation. **bold** represents the best performance for each $tsn \rightarrow eng$ model, while underline represents the best performance for each $eng \rightarrow nso$ model. The colour code is the same as in figure 4; Models that translate between the same language pair and have the same colour were trained on the same dataset.

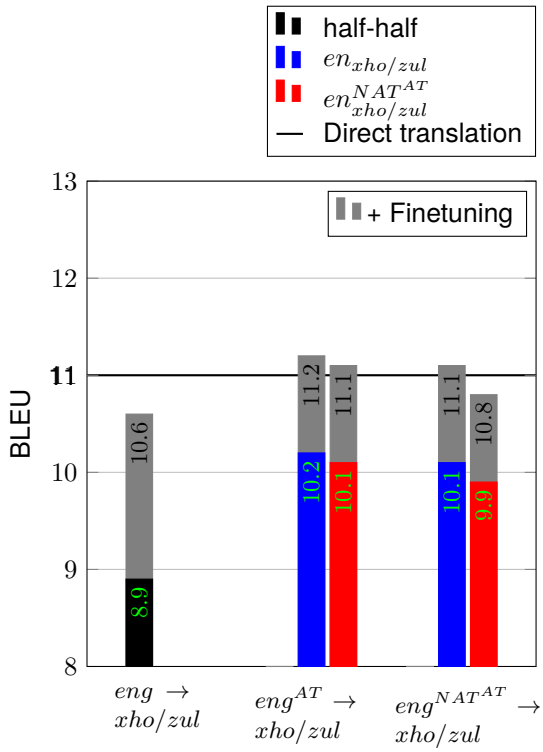

Figure 8: BLEU scores of different $zul \rightarrow xho$ models. The colour scheme indicates the source of the eng sentences that are later translated to xho and zul using the models on the x axis.

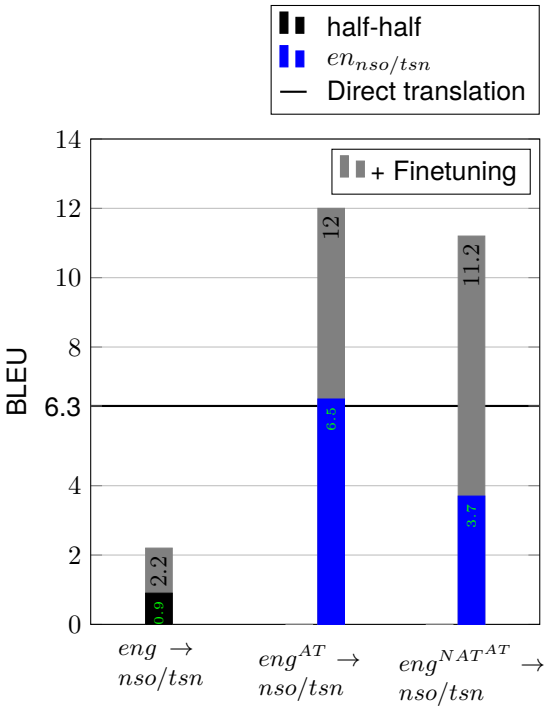

Figure 9: BLEU scores of different $nso \rightarrow tsn$ models. The colour schemee indicates the source of the eng sentences that are later translated to nso and tsn using the models on the x axis.

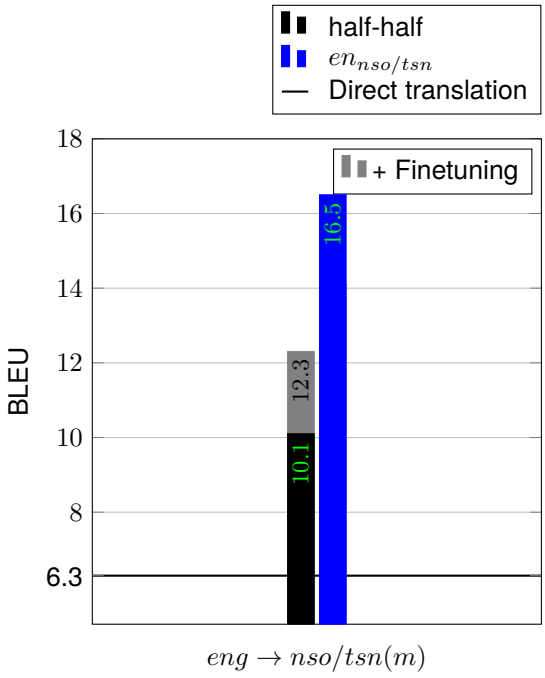

Figure 10: BLEU scores of different $nso \rightarrow tsn$ models, using a multilingual model for synthetic pivoting. The color code indicates the source of the eng sentences that are later translated to nso and tsn using the models on the x axis. (m) indicates that the nso and tsn sentences are generated from the multilingual model.