Beyond the Mode: Sequence-Level Distillation of Multilingual Translation Models for Low-Resource Language Pairs

Aarón Galiano-Jiménez,[+] **Juan Antonio Pérez-Ortiz,**[*+]
Felipe Sánchez-Martínez,[+] **Víctor M. Sánchez-Cartagena**[+]

[+]Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain
[*]Valencian Graduate School and Research Network of Artificial Intelligence, ValgrAI

{aaron.galiano,japerez,fsanchez,vm.sanchez}@ua.es

Abstract

This paper delves into sequence-level knowledge distillation (KD) of multilingual pre-trained translation models. We posit that, beyond the approximated mode obtained via beam search, the whole output distribution of the teacher contains valuable insights for students. We explore the potential of n-best lists from beam search to guide student's learning and then investigate alternative decoding methods to address observed issues like low variability and under-representation of infrequent tokens. Our research in data-limited scenarios reveals that although sampling methods can slightly compromise the translation quality of the teacher output compared to beam search based methods, they enrich the generated corpora with increased variability and lexical richness, ultimately enhancing student model performance and reducing the gender bias amplification commonly associated with KD.

1 Introduction

Neural machine translation (NMT) is an essential tool for communication and understanding between people who speak different languages. While there are NMT models that offer high performance for translation of high-resource languages, low-resource languages, with limited available training data, still pose a significant challenge for NMT (Goyal et al., 2020).

Multilingual NMT models can help address this issue by leveraging information from high-resource languages (Tran et al., 2021). While large language models (LLMs) also benefit from high-resource languages and deliver high-quality translations for them, studies have shown that they underperform compared to traditional encoder-decoder models when translating low-resource languages (Zhu et al., 2024).

In recent years, several multilingual pre-trained models —such as NLLB-200 (NLLB Team et al., 2022), DeltaLM (Ma et al., 2021), and MADLAD-400 (Kudugunta et al., 2023)— have been released, often outperforming bilingual models trained from scratch for low-resource languages. However, despite their performance, the high hardware requirements of multilingual NMT models make them impractical for general use. Knowledge distillation (KD) (Hinton et al., 2015) addresses this challenge by transferring knowledge from a large *teacher* model to a smaller, more efficient *student* model.

KD techniques can be classified into word-level (Hinton et al., 2015) and sequence-level (Kim and Rush, 2016). Word-level KD mimics the teacher's probability distribution for each token, while sequence-level KD trains the student using a synthetic corpus generated by the teacher. In both cases, the same corpus used to train the teacher is used for the distillation. Although both approaches have been widely studied in bilingual contexts (Wang et al., 2021; Zhang et al., 2018), they rely on the availability of the training parallel data, an issue for pre-trained multilingual models developed by private companies, where such data is often inaccessible. It is also possible that the desired translation direction benefits from transfer learning, but there is no parallel corpus available for the target language pair.

Without access to the parallel corpus used to train the teacher, sequence-level KD can be applied by translating a monolingual corpus to create a synthetic parallel corpus for training the student model (Lai et al., 2021; Yu et al., 2021). Sequence-level KD has typically been conducted using beam search (Graves, 2012; Kim and Rush, 2016), but this is detrimental to lexical richness. Beam search prioritizes maximizing the probability of the generated sequence, leading to repetitions, low variability (Kulikov et al., 2019) and under-representation of infrequent tokens (Müller and Sennrich, 2021). Consequently, a corpus generated via beam search is less lexically rich com-

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 6676–6691
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

pared to a native corpus (Holtzman et al., 2020). The over-representation of the most likely tokens by beam search limits the student model's exposure to a smaller range of plausible translations, as beam search translations are biased towards the most common patterns. This reduces the model's robustness, flexibility and ability to generalize. As demonstrated by Ahn et al. (2022), training with biased data amplifies the social and gender biases present in the teacher model. Beam search's emphasis on high probability results may increase the frequency of stereotypical or biased translations, explaining why distilled models show more bias than the teacher (Vamvas and Sennrich, 2021).

In this paper, we explore whether beam search is the best way to extract knowledge from a large multilingual NMT model for sequence-level KD, and whether sampling methods (Fan et al., 2018a), that generate more human-like text (Holtzman et al., 2020), are a feasible or complementary alternative. We posit that the teacher's full output distribution holds valuable insights for the student, thus supplementing the limited information captured by the approximated mode of the target language distribution obtained via beam search (Eikema and Aziz, 2020). In order to extract a broader range of knowledge from the teacher, we propose to generate several translations from the same source sentence, using the n-best list of beam search[1] or diverse beam search (Vijayakumar et al., 2018), or multiple iterations with sampling methods. We hypothesize that these multiple translations will reduce the over-representation of the most likely tokens, helping the student model to generalize better and reduce model gender bias.

This leads us to formulate the following research questions in sequence-level KD:

- RQ1: Is it effective to generate multiple translations from the same source sentence?
- RQ2: Is it possible to reduce gender bias amplification with multiple translations?
- RQ3: Can other decoding methods overcome the limitations of beam search?
- RQ4: What is the influence of factors such as the size of the corpus translated by the teacher and the sampling hyperparameters that control deviation from the mode?

To the best of our knowledge, this is the first paper to compare the properties of decoding methods focused on extracting knowledge from a pre-trained multilingual model using only monolingual corpus.

This paper is structured as follows: next section describes related work. Sec. 3 presents the experimental settings. Sec. 4 shows the experiments carried out[2] to analyse the key variables in the distillation process and the results of each experiment, followed by the concluding remarks and future work in Sec. 5.

2 Related work

The role of decoding methods. Neural models generate output tokens by producing a probability distribution over the target vocabulary at each decoding step. There are two ways for selecting these output tokens: deterministic methods, which prioritize high-probability tokens but offer low variability (Kulikov et al., 2019), and stochastic methods, which sample from the distribution but can lead to incoherent text (Basu et al., 2021). For *directed generation* tasks, such as NMT, beam search is commonly used because the output is closely tied to the input, and variability is less critical. In contrast, *open-ended* tasks, like story generation, require more diverse and human-like output (Holtzman et al., 2020). While several studies analyse decoding methods (Su et al., 2022; DeLucia et al., 2021; Wiher et al., 2022) and evaluate the quality of the resulting corpus (Pillutla et al., 2021), their focus on LLM and open-ended tasks, which limits their applicability to NMT.

Concerning NMT, despite some proposed methods to increase translation variability (Kool et al., 2019; Leblond et al., 2021; Hewitt et al., 2022), beam search remains the most widely used, especially for sequence-level KD (Kim and Rush, 2016). The reason for the extensive use of beam search is due to its balance between computational efficiency and output quality. Though approaches like Minimum Bayes' Risk (MBR) decoding (Kumar and Byrne, 2004) yield better translation quality, they are too slow for practical use, even with recent optimisations (Freitag et al., 2023; Vamvas and Sennrich, 2024). While the relationship between corpus quality and variability has been explored in open-ended tasks (Zhang et al., 2021), it is understudied for KD in NMT.

[1] Kim and Rush (2016) suggested using n-best lists from beam search but concluded that the 1-best "worked well".

[2] The code is available at https://github.com/transducens/sampling-distillation

KD techniques. Regarding KD of multilingual translation models, some studies employ multiple teacher models, multilingual (Do and Lee, 2023) or bilingual (Tan et al., 2019), to distil knowledge into a single multilingual student. In contrast, our approach distills knowledge from a single teacher into a bilingual student, differing from Gumma et al. (2023), who compress a single multilingual teacher into a multilingual student. Similarly, De Gibert et al. (2023) distil a high-resource pair from NLLB and then fine-tune the student model on the low-resource languages.

Concerning bilingual student models, some methods use high-resource languages related to low-resource ones for distillation, training the student with both languages as sources and English as the target (Song et al., 2023). In contrast, our study is not limited to English as the target language. Galiano-Jiménez et al. (2023) fine-tune the teacher for specific language pairs and train the student model with a mix of parallel and forward and back-translated synthetic data.

Recently, MBR has been used to generate sequences with an LLM for fine-tuning an encoder-decoder translation model (Finkelstein and Freitag, 2024). The study concluded that a student model fine-tuned with MBR-generated outputs outperforms one fine-tuned using beam search translations. Wang et al. (2024) extended this approach by incorporating multiple references, similar to our proposed method. They concur that extracting multiple sentences from the teacher model better captures its probability distribution, leading to improved student models. However, our work explores a broader range of scenarios and concludes that the choice of decoding method should depend on both the teacher model's translation quality and the size of the available corpus.

3 Experimental settings

This section details the experiments carried out, covering decoding methods, language pairs, the model and corpora used, and the evaluation criteria.

3.1 Decoding methods

This study examines a selected set of decoding methods, including beam search and diverse beam search as deterministic approaches, and nucleus sampling and top-k as stochastic approaches.[3]

Beam search: At each decoding step beam search keeps the n highest probability paths (Graves, 2012). This has the advantage of identifying high probability sequences that start with less likely initial tokens and would have been ignored by greedy decoding, which always chooses the most probable token.

Diverse beam search: It is a variant of beam search that tries to produce more diverse results. Instead of maintaining a single list of the most likely paths, it divides the n paths into G groups and applies a penalty to prevent them from being similar to each other (Vijayakumar et al., 2018). As recommended by the authors, we used n=G, i.e. as many groups as n, with only one sequence per group, and λ=0.5.

Top-k: The k most likely next tokens are filtered and the probability mass is redistributed among them (Fan et al., 2018b). A small k means less variability and progressively more similarity to greedy decoding. For our experiments, we kept the original proposal of k=10 (Fan et al., 2018b), which has proven to work well for generating synthetic corpora for back-translation (Zhang et al., 2020).

Top-p (nucleus sampling): It chooses from the smallest possible set of tokens whose cumulative probability exceeds the probability p (Holtzman et al., 2020). The probability mass is then redistributed among this set of tokens. This way, the size of the set of tokens can dynamically increase and decrease according to the next token's probability distribution. Following Eikema and Aziz (2022), we set p to 0.7 in our experiments.

3.2 Models, language pairs and data

Models. We used NLLB-200 1.3B and NLLB 3.3B (NLLB Team et al., 2022) as teacher models to assess the generalization of our approach to different model sizes.

Our students are encoder-decoder Transformer models in the *base* configuration, as defined by Vaswani et al. (2017, Tbl. 3). With 65M parameters, our student models are notably compact, representing just 5% of the size of the NLLB 1.3B model. For more details on the architecture and training, see Appendix B.

[3] We exclude ancestral sampling from our analysis because the NLLB model was trained using label smoothing, which elevates the likelihood of rare events, leading to translations of significantly lower quality.

Langs	BS	DBS	top-p	top-k
eng-swh	33.1	32.2	25.1	21.5
eng-ibo	16.1	13.7	12.4	10.9
eng-bam	6.8	6.1	4.9	4.6
swh-eng	42.9	42.4	34.8	28.8
ibo-eng	30.3	30.3	24.3	20.9
bam-eng	17.8	16.2	14.4	12.6
bam-swh	11.6	8.3	8.9	8.3

Table 1: BLEU scores of NLLB 1.3B on the FLORES+ devtest dataset when decoding with beam search (BS), diverse beam search (DBS), top-p (average of 3 runs) and top-k (average of 3 runs). The NLLB-3.3B scores can be found in Appendix C.

Language pairs. For this study, we selected languages based on the quality of the translations performed by the teacher (Table 1) and the size of the available corpora. Our objective is to have multiple scenarios that allow us to analyse the impact of different variables at both generation and training time. The languages we have chosen are English (eng), Swahili (swh), Igbo (ibo) and Bambara (bam). Translation directions to be distilled are as follows:

- **eng-swh, eng-ibo, eng-bam**. Scenario in which we have almost unlimited monolingual source corpora and different qualities of translation.
- **swh-eng, ibo-eng, bam-eng**. Scenario where we have small amount of monolingual data to translate, but enough to try out various sizes in some cases. As the teacher has learned a lot of English and beam search limits the vocabulary we can extract, we theorise that sampling methods allow us to extract more knowledge.
- **bam-swh**. Small amount of monolingual data and low quality translation. The teacher's knowledge is based on transfer learning, generalised from the other translation directions and monolingual knowledge of the source and target languages.

Data. English and Swahili have the most extensive corpora, from which we selected a subset of 1 million sentences. For Igbo, we used a corpus comprising 451,789 sentences, while for Bambara we employed a corpus containing 108,187 sentences. All corpora used are freely available. Specific details on the corpora can be found in Appendix A.

As development and test sets we use the FLORES+[4] (NLLB Team et al., 2022) *dev* and *devtest*, respectively.

3.3 Evaluation metrics

We evaluate two main elements: the synthetic corpora and the models trained on them.

Corpora. We focus on their vocabulary, the fidelity of translations, and the sentence variability. We assess lexical richness by measuring vocabulary diversity using Zipf's Law and counting unique words and sentences to compare corpora. Variability is measured by analysing the diversity of translations generated from the same source sentence using self-BLEU (Zhu et al., 2018), with lower values indicating greater variability.

For translation quality, we rely on BLEU (Papineni et al., 2002) and chrF (Popović, 2015) to evaluate teacher output on FLORES+ dataset, given the lack of neural learned metrics for these languages. Additionally, we report COMET (Rei et al., 2020) for the supported languages.

All corpora were generated by translating the respective monolingual corpus with the teacher, using the transformers library (Wolf et al., 2020) and the desired decoding method.

Models. We use two main metrics to evaluate the students' performance: first, their translation quality, evaluated in the same way as the teacher's, with the test set translated using beam search (n=5); and second, gender biases. Due to the lack of annotated datasets for these languages, we measure gender bias using contrastive conditioning (Vamvas and Sennrich, 2021), as outlined in Sec. 4.4.

4 Experiments and results

Our approach is to apply sequence-level KD using different decoding methods and generating multiple translations. The aim is to experiment with different sizes of monolingual corpora and to analyse the features of the generated corpora in order to test how these features affect the quality of the student models.

4.1 Effect on distillation

Sampling methods typically yield lower performance for machine translation compared to beam

[4] https://github.com/openlanguagedata/flores

search and diverse beam search as shown in Table 1. However, in this section, we assess whether, despite this drop in performance, sampling methods offer superior data for training student models.

We translated 100k sentences using beam search (n=10), diverse beam search (n=10=G, λ=0.5), top-p (p=0.7) and top-k (k=10), generating 10 translations per sentence. While top-p and top-k rely on sampling, beam search and diverse beam search select the 10 highest-probability candidates. This process yielded a training corpus of 1 million parallel sentences, with each source sentence translated 10 times into the target language. We then divided each corpus into four blocks based on the number of target sentences per source, with the complete corpus containing 10 samples, and the other blocks containing 5, 3, and 1 sample(s). The version with a single sentence from beam search corresponds to standard sequence-level KD. Subsequently, we trained student models on these generated training samples.

Results. Fig. 1 shows the performance of student models, with scores reflecting the average BLEU or chrF on three training runs.[5] See Appendix C.3 for the results of both metrics for all models. Results from NLLB 3.3B, which exhibit a similar pattern, are provided in Appendix C.2.

The results of paired approximate randomization (Riezler and Maxwell, 2005) statistical significance tests are shown in Appendix C.3. We compared all student models with the model trained with one translation from beam search as the first baseline and the model trained with 10 translations from beam search as the second baseline. Except for the eng-ibo models, all translation directions showed statistically significant differences compared to standard sequence-level KD. When compared to the second baseline, models trained with diverse beam search translations for ibo-eng, bam-eng, and eng-swh did not show differences.

As expected, student models trained on beam search (or diverse beam search) outputs generally performed best than sampling methods when only one translation per sentence was generated. However, as the number of translations per sentence increased, models trained on sampled data outperformed those trained with beam search and diverse beam search. The gap between beam search and sampling methods is especially notable for bam-swh. When distilling NLLB 1.3B, students trained with 5 or 10 beam search translations performed worse than those trained with only 3 or 1 translations. In this case, traditional KD with beam search failed due to the poor quality of translations and lack of data, but sampling methods enabled student models to approach the teacher's performance. In general, the difference between beam search and top-p student models is greater when the target is English. This is in line with our hypothesis that, as the teacher has been trained with so much English and we are working with such small corpora, the sampling methods allow us to extract more information than beam search.

To ensure that the improvements seen with sampling methods were not simply due to a particularly good translation among the multiple outputs, we conducted an additional experiment. For the eng-swh corpus generated by top-p and 10 translations, we selected the best translation for each sentence based on COMET without reference. We used only these selected translations to train a student model. The resulting performance was similar to that of a model trained with just one translation, confirming that the improvements observed with sampling methods were driven by the diversity of multiple translations.

Finally, we distilled eng-swh and eng-bam using MBR to evaluate its effectiveness despite its slower performance. Following Finkelstein and Freitag (2024), we used epsilon sampling (Hewitt et al., 2022) to generate 256 candidates with $\epsilon = 0.02$, selecting the top 10 translations based on fastChrF (Vamvas and Sennrich, 2024) as the utility function. For eng-swh, we observed that MBR produces less variability compared to top-p, resulting in worse student models. In contrast, for Bambara, an extremely low-resource language, the teacher's probability distribution is highly distributed, so the teacher benefits from translation with MBR. In this scenario, the results obtained with MBR were slightly better than those obtained with top-p. However, as top-p offers similar performance while being significantly faster, we consider it the preferable option.

Generated data analysis. To explain these results, we measured the variability between the translations using self-BLEU (Zhu et al., 2018). As expected, deterministic methods produce sentences with low variability (even diverse beam search), whereas sampling methods, particularly

[5] Note that, for the sampling methods, translations were generated again by the teacher in each training run.

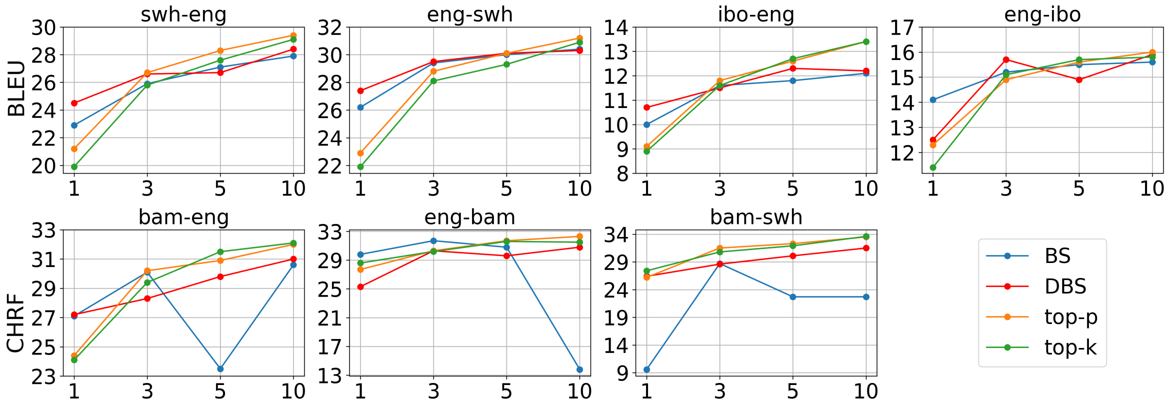

Figure 1: Average BLEU (first row) and chrF (second row) scores obtained by student models trained on samples generated with different decoding methods and varying number of translations per source sentence (x-axis).

top-k, yield more diverse ones (see Appendix C.1). However, top-k and top-p, despite generating more diverse translations, produce repeated translations, specially in the case of top-p, due to a narrow choice window when the model is confident. This, while potentially limiting diversity, can prevent hallucinations that could harm student training.

An analysis of the probabilities normalised by length (Figure 2) of the sentences produced shows a decrease in the probabilities of the deterministic methods. The sampling methods, on the other hand, produce sentences with a lower probability, but this probability remains stable. This, together with Eikema and Aziz (2020)'s observations on the inadequacy of the mode, may explain the decrease in quality observed in students trained with multiple beam search translations when working with languages for which the teacher is poorly fitted.

Note that it is not possible to increase the number of beam search translations without compromising their quality. In our experiments, we set n=k=10, resulting in only 10 beam search translations, each with progressively decreasing quality. In contrast, top-k sampling allows the generation of an arbitrary number of samples while maintaining their probabilities and showing a higher variability.

Regarding lexical diversity, the Zipf's distribution (Holtzman et al., 2020) analysis reported in Appendix C.1 shows that corpora amplified with sampling from smaller texts are more similar to native corpora than those produced by beam search with a single translation from larger texts. Intuitively, the generation of multiple translations by beam search might result in either very similar sentences, adding little value, or hallucinations when the model is forced into less probable paths. This depends on the model's knowledge of the language.

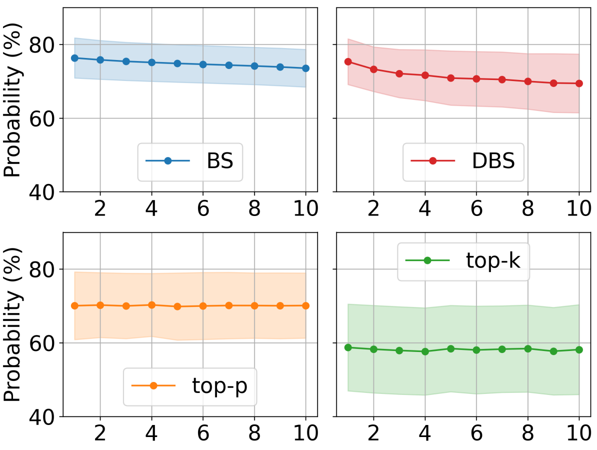

Figure 2: Probabilities of 10 swh-eng translations for each source sentence from FLORES+ devtest by NLLB 1.3B. The shaded areas around each line represent the standard deviation.

4.2 Impact of source corpus size

The size of the source corpus plays a crucial role, as a larger corpus allows for more knowledge to be extracted from the teacher. To analyse how this affects KD with sampling methods, we translated 100k, 500k and 1 million sentences, generating 10 translations per sentence. We compare each size with the result of translating the same corpus using beam search by generating a single translation.

Results. Figure 3 shows the performance of student models trained on samples generated from corpora of different sizes. The results for all translation directions can be found in Appendix C.3. As observed, the discrepancy between different decoding methods decreases as the corpus size increases. For corpora of 500k sentences, sampling methods still outperform beam search, while for corpora of 1 million sentences, sampling methods do not consistently yield superior results. Nonetheless, generating multiple translations remains advantageous. COMET results for the supported languages are shown in Figure 4.

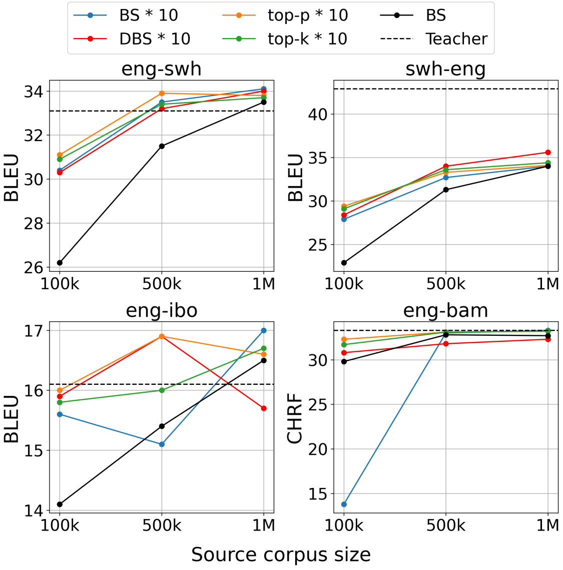

Figure 3: BLEU scores for different corpus sizes. BS corresponds to the standard sequence-level KD, and BS∗10 correspond to the use of beam search to generate 10 translations per source sentence.

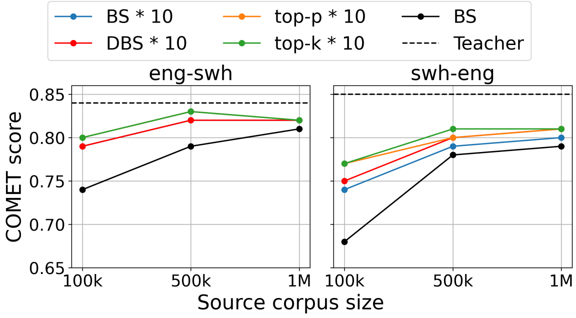

Figure 4: COMET scores for eng-swh and swh-eng. The results of top-p and top-k overlap in the eng-swh graph, as well as BS and DBS.

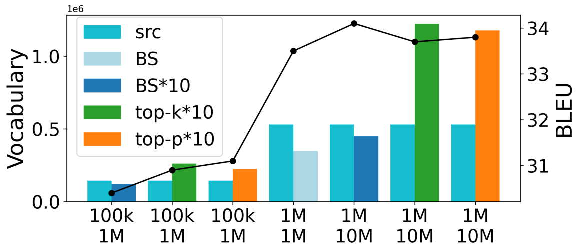

Figure 5: Relationship between the vocabulary (1×10^6) of the training corpus and the BLEU of the student models. The x-axis markers indicate sentences from the source corpus (first row) and sentences from the generated corpus (second row).

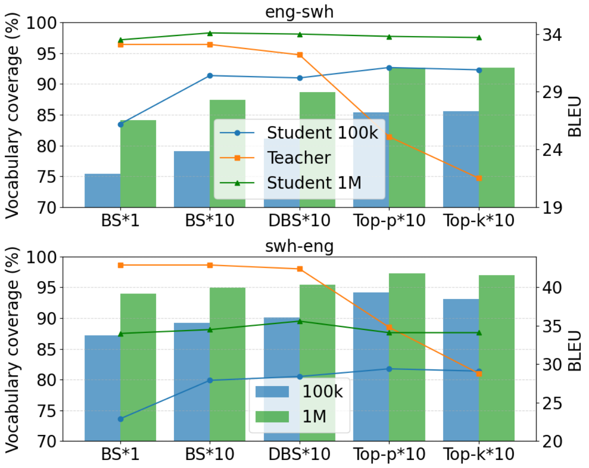

Figure 6: Effect of vocabulary coverage and teacher translation quality. The x-axis shows decoding methods ranked by variability. Columns show the percentage of test vocabulary (y-axis on the left) present in the training corpus. Lines show the BLEU (y-axis on the right) of the models trained with each corpus and the BLEU of the teacher with each decoding method.

Generated data analysis. To explore the importance of lexical richness in the translated corpus, we compared the number of unique words in both the source corpus and the generated corpus. Figure 5 illustrates the relationship between the size of the source vocabulary, the vocabulary produced by each decoding method, and the BLEU scores obtained by training student models on these corpora. The results show that sampling methods act as vocabulary amplifiers by generating multiple translations. However, it is important to know which part of this vocabulary is useful to the model. Figure 6 shows the percentage of the devtest target vocabulary present in the training corpus. It can be seen that until a certain coverage is reached (about 87% for eng-swh and 95% for swh-eng), increasing the coverage produces better student models, even if the teacher translations are worse. On the other hand, once this point is reached, it is more beneficial to prioritize translation quality.

In addition to quality, beam search can offer another benefit for KD. During training, models typically use teacher forcing, where the correct previous token is used as input, leading to a mismatch

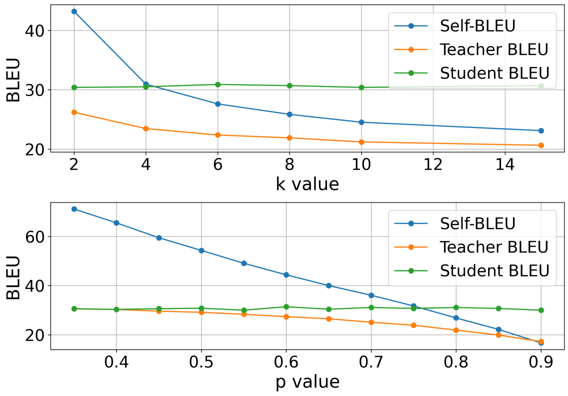

Figure 7: Relationship between the teacher translation quality and variability and the student models score for eng-swh. Initial corpus: 100k sentences

	eng-swh	eng-ibo	eng-bam
NLLB 1.3B	52.9	52.7	58.3
BS	49.2	49.4	50.8
BS * 10	51.0	50.2	50.3
DBS * 10	50.4	50.4	51.5
top-p * 10	51.7	50.5	52.3
top-k * 10	51.7	50.5	51.3

Table 2: Contrastive conditioning accuracy over WinoMT dataset. Higher scores are better and the blue scores mark the best student models.

between training and inference. This exposure bias (Ranzato et al., 2015) can be mitigated by using beam search outputs, which are closer to the tokens generated during inference. If the source corpus is sufficiently large, beam search can extract enough vocabulary, and its similarity to the inference process benefits the student model. This also explains the performance of the swh-eng model with a 1M source corpus using diverse beam search, which keeps the inference similarity of beam search while providing greater diversity.

4.3 Divergence from the mode vs. translation quality

The adjustment of the sampling parameters affects both output variability and translation quality. To gauge the sensitivity of our approach to the values of p and k, we conducted experiments on eng-swh, translating 100k English sentences. Fig. 7 illustrates the impact of p and k values on both the translation performance of the student and teacher models (measured with BLEU), and the similarity of the translations (measured with self-BLEU). As observed, higher values of p and k result in more diverse translations, albeit with poorer teacher performance, while maintaining similar performance for the student models. Finally, we repeated the experiment with 1 million sentences and found that the results were consistent with previous findings, confirming that the trade-off between quality and variability is independent of corpus size.

4.4 Analysis of gender bias

Sequence-level KD typically carries a bias amplification respect to the teacher model due to the over-representation of frequent tokens. To measure if this issue can be mitigated by generating multiple translations we used contrastive conditioning (Vamvas and Sennrich, 2021) to evaluate gender bias, using NLLB 1.3B as an evaluator model and the WinoMT dataset (Stanovsky et al., 2019). This method checks the probability of the translation generated by the evaluated model from the original source sentence using a disambiguated variation as the source of the evaluator model. For each decoding method, we evaluated the model trained with only one translation per source sentence and with 10 translations.

The results in Table 2 show how, for all methods, generating multiple translations for training reduces gender bias compared to training with only one translation.

5 Concluding remarks

This study investigates the effectiveness of generating multiple translations from the same source sentence in sequence-level KD with multilingual NMT encoder-decoder models and the effect of different decoding methods.

The results show that increasing the number of translations has a positive effect on the student model performance, especially when monolingual data is limited. Using this method, we achieve similar results to standard sequence-level KD with a much smaller monolingual corpus and improve the results with the same corpus size. Our method matches or slightly outperforms the teacher from English to low-resource languages but leaves a gap when translating into English. In multilingual models, it may not be possible to extract all the bilingual knowledge from the teacher model with only the synthetic parallel corpus of one language pair, since thanks to transfer learning, part of the translation ability comes from other translation directions. In NLLB, which is trained on different parallel corpora with English as the target, a small monolingual Swahili corpus translated into English by the teacher cannot capture all the English knowledge of the model (RQ1).

This approach also helps to reduce gender bias by increasing translation variability. This finding holds for all decoding methods, demonstrating the generalizability of the approach. Despite the overall good results, sampling methods achieve greater mitigation of bias by avoiding the over-representation of the most likely tokens inherent in beam search (RQ2).

Sampling methods allow for a more diverse corpus for learning when generating multiple translations, which is particularly beneficial for low-resource scenarios (ibo-eng, bam-eng, bam-swh). Nevertheless, with high-resource source languages, the quality of the translations and the mitigation of exposure bias obtained by beam search based methods, can compensate the low variability of these decoding methods, as occurs with eng-ibo, eng-bam and eng-swh. Especially, when the teacher model has a lot of knowledge about the source and target languages, it is able to produce multiple translations with a high probability. This explains why diverse beam search gives the best result for swh-eng when translating 1 million sentences (RQ3).

Regarding the deviation from the mode with sampling methods, the stability of the student models in the face of changing parameters p and k seems to indicate that the relationship between quality and variability remains balanced, allowing the student models to learn on the basis of one or the other (RQ4).

Future work. A promising avenue for future work is to test whether this approach remains effective with LLMs, where sampling methods are commonly used for text generation.

Regarding decoding methods, one possible approach is to combine different techniques to exploit the advantages of each, while another option is to focus on mitigating the weaknesses of certain methods. For example, stopping the generation of translations with beam search or diverse beam search when the generated translations present very low probabilities.

In terms of explainability and interpretability, we plan to explore the precise reasons for the decrease in student translation quality observed in certain language pairs when the number of beam search translations was increased.

Limitations

In spite of showing stronger correlations with human judgments than BLEU or chrF, we do not use neural-based machine translation evaluation metrics such as COMET for all languages, as the associated models do not cover the languages involved in our research. Furthermore, our research has specifically focused on a single model at different sizes (NLLB 1.3B and NLLB 3.3B). Our conclusions, therefore, might not be fully applicable to different models. The decision to limit our investigation to this specific model was primarily due to constraints in computational resources and time.

As regards, other approaches to KD, our approach has not been compared to word-level KD due to the requirement for parallel corpora, which was not completely feasible within some of our constraints. Regarding the decoding methods, our experiments testing the influence of different values of p and k for training student models were limited to the English-Swahili language pair. Consequently, the findings related to particular values of these parameters may not generalise across other language pairs. It is important to note, however, that the approaches we tested are designed to be applicable to any language pair within the multilingual model, offering broad relevance despite this limitation. Moreover, the selection of the values for p and k in our study was guided by established precedents within the literature.

We did not explore combining the synthetic sentences provided by different decoding methods, as our focus was on understanding individual contributions.

Lastly, due to the fact that NLLB was trained with label smoothing (NLLB Team et al., 2022), we were unable to explore ancestral sampling in our experiments. Label smoothing increases the probability of low-frequency events, which could result in lower-quality translations in scenarios that rely on ancestral sampling.

Ethics Statement

Knowledge distillation endeavors to produce smaller, more resource-efficient NMT systems, thereby diminishing energy requirements compared to the original systems and consequently aiding in the reduction of CO_2 emissions. Moreover, it lowers the entry barrier for deploying NMT models, as the resulting models work on lower-power hardware. Our student models are remarkably compact, operating at a mere 5% of the teacher model size. However, delving into knowledge distillation necessitates a substantial number of training iterations, each accompanied by its own energy consumption. For the experiments detailed in this paper, we trained 482 Transformer models employing NVIDIA GeForce RTX 2080 Ti GPUs. Furthermore, all corpora and tools utilized in this study are available under open source licenses, ensuring the complete reproducibility of the presented results.

Acknowledgments

This paper is part of the R+D+i project PID2021-127999NB-I00 funded by the Spanish Ministry of Science and Innovation (MCIN), the Spanish Research Agency (AEI/10.13039/501100011033) and the European Regional Development Fund A way to make Europe. Some of the computational resources used were funded by the European Regional Development Fund through project IDIFEDER/2020/003.

References

David Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. Mirostat: A neural text decoding algorithm that directly controls perplexity.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, Toronto, Canada. Association for Computational Linguistics.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding methods for neural narrative generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185, Online. Association for Computational Linguistics.

Heejin Do and Gary Geunbae Lee. 2023. Target-oriented knowledge distillation with language-family-based grouping for multilingual nmt. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018a. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018b. Hierarchical neural story generation.

Mara Finkelstein and Markus Freitag. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *The Twelfth International Conference on Learning Representations*.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, and Juan Antonio Pérez-Ortiz. 2023. Exploiting large pre-trained models for low-resource neural machine translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 59–68, Tampere, Finland. European Association for Machine Translation.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online. Association for Computational Linguistics.

Alex Graves. 2012. Sequence transduction with recurrent neural networks.

Varun Gumma, Raj Dabre, and Pratyush Kumar. 2023. An empirical study of leveraging knowledge distillation for compressing multilingual neural machine translation models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 103–114, Tampere, Finland. European Association for Machine Translation.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proc.*

Wouter Kool, Herke van Hoof, and Max Welling. 2019. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement.

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. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Wen Lai, Jindřich Libovický, and Alexander Fraser. 2021. The LMU Munich system for the WMT 2021 large-scale multilingual machine translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 412–417, Online. Association for Computational Linguistics.

Rémi Leblond, Jean-Baptiste Alayrac, Laurent Sifre, Miruna Pislar, Jean-Baptiste Lespiau, Ioannis Antonoglou, Karen Simonyan, and Oriol Vinyals. 2021. Machine translation decoding beyond beam search.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. ΔLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. 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 259–272, Online. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Maja Popović. 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. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Yewei Song, Saad Ezzini, Jacques Klein, Tegawende Bissyande, Clément Lefebvre, and Anne Goujon. 2023. Letz translate: Low-resource machine translation for luxembourgish.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *Seventh International Conference on Learning Representations*.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI WMT21 news translation task submission. In *Proc. of the Sixth Conference on Machine Translation (WMT)*, pages 205–215.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation.

Jun Wang, Eleftheria Briakou, Hamid Dadkhahi, Rishabh Agarwal, Colin Cherry, and Trevor Cohn. 2024. Don't throw away data: Better sequence knowledge distillation.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhengzhe Yu, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021. HW-TSC's participation in the

WMT 2021 large-scale multilingual translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 456–463, Online. Association for Computational Linguistics.

Dakun Zhang, Josep Crego, and Jean Senellart. 2018. Analyzing knowledge distillation in neural machine translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 23–30, Brussels. International Conference on Spoken Language Translation.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Yuhao Zhang, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, et al. 2020. The niutrans machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 338–345.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Corpora

The largest corpora correspond to English and Swahili. The English corpus is a fragment of OSCAR-3301 dataset[6] and for Swahili we used Monolingual African Languages from ParaCrawls, a collection of corpora available for the joint task Large-Scale Machine Translation Evaluation for African Languages at WMT22 (Adelani et al., 2022). The Igbo corpus was obtained from the same collection.

To clean these three corpora, we used `monocleaner` (Sánchez-Cartagena et al., 2018). We used the available ready-to-use language packages for English and Swahili and trained a model for Igbo using the Igbo part of the wmt22_african dataset.[7] We removed all sentences with a `monocleaner` score lower than 0.5 and, for English and Swahili, we then randomly picked one million sentences. For Igbo, our final corpus comprises 451,789 sentences.

For Bambara we collected all available corpora in Hugging Face.[8] For the MADLAD-400 (Kudugunta et al., 2023) corpus we used only the clean part. After concatenating these corpora, we removed duplicated sentences and the result was 108,187 sentences.

B Student models

Each student model consist of a transformer (Vaswani et al., 2017) with 6 layers for both the encoder and the decoder, embedding dimension of 512, feed-forward inner-layer dimension of 2048, and 8 attention heads. All our models were trained using the `Fairseq` toolkit[9] and a different joint bilingual SentencePiece (Kudo and Richardson, 2018) model for each language pair, trained on the training samples generated from the teacher with a vocabulary of 10,000 tokens. For training we used a learning rate of 0.0007 with the Adam (Kingma and Ba, 2015) optimizer (β_1=0.9, β_2=0.98), 8,000 warm-up updates and 8,000 max tokens. We used parameters dropout of 0.1 and updated the model after 2 training steps. The cross-entropy loss with label smoothing was computed on the development set after every epoch and the best checkpoint was selected after 6 validation steps with no improvement.

C Additional results

This sections reports additional results to measure the effect of decoding methods in sequence-level KD. In addition, the performance of NLLB-3.3B with the different decoding methods is shown in Table 3.

[6] https://huggingface.co/datasets/oscar-corpus/OSCAR-2301

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

[8] https://huggingface.co/datasets/RobotsMaliAI/bayelemabaga, https://github.com/masakhane-io/lafand-mt, https://wortschatz.uni-leipzig.de/en/download/Bambara, https://github.com/facebookresearch/flores/tree/main/nllb_seed, https://huggingface.co/datasets/bigscience/xP3, https://huggingface.co/datasets/allenai/MADLAD-400

[9] https://github.com/facebookresearch/fairseq

Langs	BS	DBS	top-p	top-k
eng-swh	33.9	32.7	28.9	22.1
eng-ibo	16.2	15.9	14.0	10.8
eng-bam	7.0	6.0	5.9	4.7
swh-eng	44.8	44.2	39.7	30.9
ibo-eng	32.0	31.1	28.1	21.9
bam-eng	17.5	17.1	15.1	12.1
bam-swh	10.8	10.7	9.3	7.0

Table 3: BLEU scores of NLLB 3.3B on the FLORES+ devtest dataset when decoding with beam search (BS), diverse beam search (DBS), top-p (average of 3 runs) and top-k (average of 3 runs).

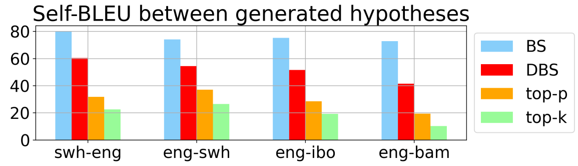

Figure 8: Similarity among the generated translations as evaluated by self-BLEU.

C.1 Lexical diversity

This section contains the experiments on lexical variability and diversity discussed in Section 4.1. Figure 8 shows the self-BLEU score obtained by each decoding method when generating 10 translations per source sentence.

Figures 9 compares the Zipf distribution of each generated corpus, translated from English to Swahili, together with the distribution of the native Swahili corpus of 1M sentences (mono_1M in the plot). The figure also includes the distribution of a corpus generated by translating the English corpus of 1M sentences with beam search, but with only one translation for each source sentence (BS_1M). It can be seen that sampling methods produced corpora closer to native language corpora than beam search, even in those cases in which a small corpus is amplified via sampling.

C.2 Experiments with 100k sentences

The results obtained using NLLB 3.3B are shown in Figure 10.

C.3 Experiments with 500k and 1 million sentences

Tables 4 and 5 show the BLEU and chrF scores of the trained student models together with the teacher scores. The results in Table 4 correspond to those in Figure 3, together with the language pairs for which the available corpus does not reach one million sentences.

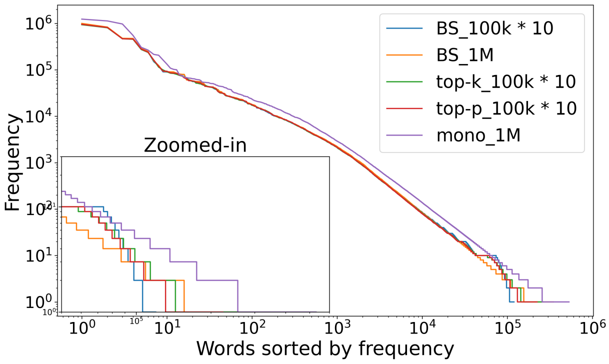

Figure 9: Zipf's distribution over Swahili corpora. Similar patterns were observed for the other languages.

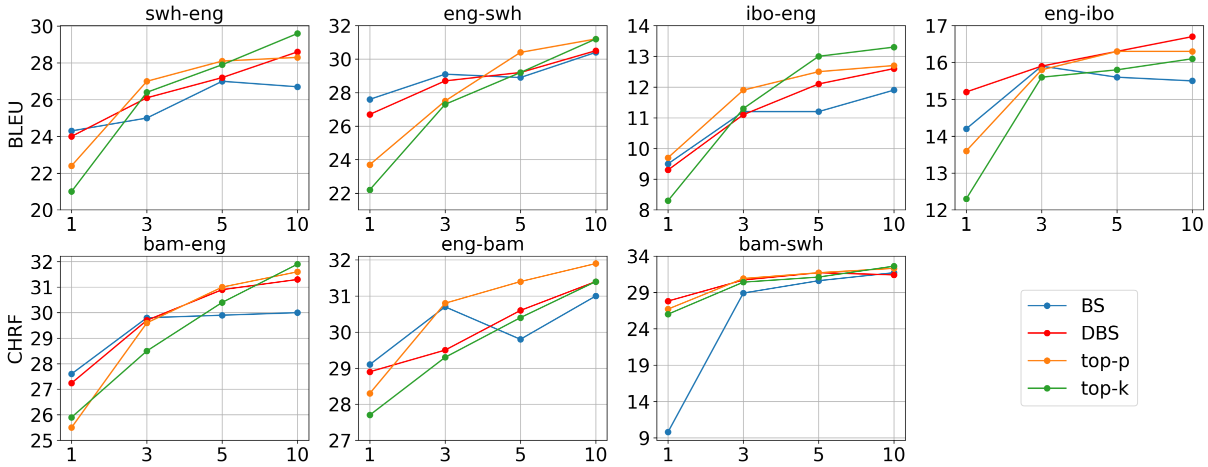

Figure 10: Average BLEU and chrF scores (y-axis) obtained by student models trained on samples generated by NLLB 3.3B with different decoding methods and varying number of translations per source sentence (x-axis).

	eng-swh	eng-ibo	eng-bam	swh-eng	ibo-eng	bam-eng	bam-swh
NLLB 1.3B	33.1	16.1	6.8	42.9	30.7	17.8	11.6
100k BS	26.2	14.1	4.7	22.9	10.0	5.8	2.1
100k BS * 10	30.4	**15.6**	0.3	27.9	12.1	8.7	1.2
100k DBS * 10	**30.2**	**15.9**	6.3	28.4	**12.4**	**8.3**	6.1
100k top-p * 10	31.1	**16.0**	6.6	29.4	13.4	9.9	7.7
100k top-k * 10	**30.9**	**15.8**	6.3	29.1	13.4	9.7	7.5
500k BS	31.5	15.4	6.3	31.3	14.2	–	–
500k BS * 10	33.5	15.1	6.4	32.7	15.5	–	–
500k DBS * 10	33.2	16.9	6.6	34.0	16.9	–	–
500k top-p * 10	33.9	16.9	6.9	33.3	16.2	–	–
500k top-k * 10	33.4	16.0	6.7	33.6	17.1	–	–
1M BS	33.5	16.5	6.5	34.0	–	–	–
1M BS * 10	34.1	17.0	6.8	34.5	–	–	–
1M DBS * 10	34.0	15.7	6.6	35.6	–	–	–
1M top-p * 10	33.8	16.6	7.1	34.1	–	–	–
1M top-k * 10	33.7	16.7	7.1	34.4	–	–	–

Table 4: BLEU scores on the FLORES+ devtest for several student models and the teacher. Underlined results are those that show no statistically significant difference compared to beam search with a single translation. Bolded results are those that show no statistically significant difference compared to beam search with 10 translations.

	eng-swh	eng-ibo	eng-bam	swh-eng	ibo-eng	bam-eng	bam-swh
NLLB 1.3B	61.8	43.2	33.3	64.8	54.0	40.2	38.3
100k BS	54.8	39.5	29.8	49.0	34.3	27.1	9.6
100k BS * 10	59.3	41.6	13.8	52.9	37.1	30.6	22.7
100k DBS * 10	59.0	41.9	30.8	53.6	37.6	31.0	31.4
100k top-p * 10	59.9	41.9	32.3	54.1	38.3	32.0	33.5
100k top-k * 10	59.4	41.8	31.7	54.1	38.2	32.1	33.6
500k BS	60.2	42.1	32.8	56.0	38.8	–	–
500k BS * 10	61.7	42.3	33.0	57.1	40.8	–	–
500k DBS * 10	61.7	43.1	31.8	57.8	42.2	–	–
500k top-p * 10	61.9	43.3	33.1	57.6	41.3	–	–
500k top-k * 10	61.7	42.5	33.1	58.1	41.7	–	–
1M BS	61.4	42.9	32.7	57.9	–	–	–
1M BS * 10	62.2	43.7	33.3	58.3	–	–	–
1M DBS * 10	62.0	42.6	32.3	59.6	–	–	–
1M top-p * 10	61.7	43.1	33.2	58.2	–	–	–
1M top-k * 10	61.8	43.2	33.2	58.5	–	–	–

Table 5: chrF scores on the FLORES+ devtest for several student models and the teacher model.