Generating from AMRs into High and Low-Resource Languages using Phylogenetic Knowledge and Hierarchical QLoRA Training (HQL)

William Soto Martinez, Yannick Parmentier, Claire Gardent

Generating from AMRs into High and Low-Resource Languages using Phylogenetic Knowledge and Hierarchical QLoRA Training (HQL)

William Soto Martinez
Université de Lorraine / LORIA
william-eduardo.soto-martinez@loria.fr

Yannick Parmentier
Université de Lorraine / LORIA
yannick.parmentier@loria.fr

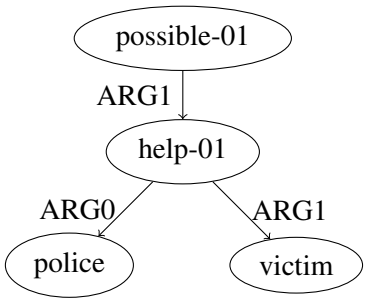
Claire Gardent
CNRS/LORIA and Université de Lorraine
claire.gardent@loria.fr

Abstract

Previous work on multilingual generation from Abstract Meaning Representations has mostly focused on High- and Medium-Resource languages relying on large amounts of training data. In this work, we consider both High- and Low-Resource languages capping training data size at the lower bound set by our Low-Resource languages i.e., 31K training instances. We propose two straightforward techniques to enhance generation results on Low-Resource while preserving performance on High- and Medium-Resource languages. First, we iteratively refine a multilingual model to a set of monolingual models using Low-Rank Adaptation - this enables cross-lingual transfer while reducing over-fitting for High-Resource languages as the monolingual models are trained last. Second, we base our training curriculum on a tree structure which permits investigating how the languages used at each iteration impact generation performance on High and Low-Resource languages. We show an improvement over both mono and multilingual approaches. Comparing different ways of grouping languages at each iteration step we find two beneficial configurations: grouping related languages which promotes transfer, or grouping distant languages which facilitates regularisation.

1 Introduction

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a representation language used to encode the meaning of sentences. Figure 1 shows an example AMR graph and some of its possible verbalisations in 4 different languages. AMR-to-Text generation is the task of verbalizing the meaning encoded by an AMR graph. While there has been constant progress on this task for the English language (Hoyle et al., 2021; Ribeiro et al., 2021b,c; Bevilacqua et al., 2021) and some other High-Resource (HR) and Medium-Resource (MR) languages (Fan and Gardent, 2020; Ribeiro et al., 2021a; Xu et al., 2021; Martínez Lorenzo et al., 2022; Sobrevilla Cabezudo and Pardo, 2022), not much attention has been given to this task on Low-Resource (LR) languages.



Eng: The police could help the victim.

Deu: Die Polizei konnte dem Opfer helfen.

Spa: La policía podría ayudar a la víctima.

Ita: La polizia potrebbe aiutare la vittima.

Figure 1: An example AMR graph and its meaning in English, German, Spanish and Italian.

Previous work on machine translation (MT) exposes a complex trade-off between High- and Low-Resource languages. While Koehn and Knowles (2017) show that neural MT models have a steep learning curve leading to poor performance in Low-Resource scenarios, Lin et al. (2020); Aharoni et al. (2019) demonstrate that multilingual training mitigates this effect. Conversely, Conneau et al. (2020) observe that the noise resulting from multilingual training negatively affects HR languages while NLLB Team et al. (2022) show that curriculum learning (Bengio et al., 2009) can help reduce over-fitting on LR languages. Phylogenetic knowledge has sometimes been used to handle this tradeoff both in multilingual NLU tasks such as dependency parsing, part of speech tagging, and natural language inference (Faisal and Anastasopoulos, 2022) and in NLG tasks such as

Proceedings of the 17th International Natural Language Generation Conference, pages 70–81
September 23–27, 2024.

Knowledge Graph-to-Text generation (Soto Martinez et al., 2023). Recent work (Meng and Monz, 2024) has also shown that training on closely related languages facilitates transfer while training on distant languages has a regularization effect. Finally, Parameter-Efficient Fine-Tuning approaches have proven useful in learning new tasks and languages for text generation of LR languages (Vu et al., 2022) while keeping memory requirements low during training.

In this work, we focus on AMR-to-Text generation and propose two simple yet efficient techniques to improve transfer from High- to Low-Resource languages while preserving performance on HR languages. First, we iteratively refine a multilingual model to a set of monolingual models using Low-Rank Adaptation (LoRA) (Hu et al., 2021). We hypothesise that this promotes cross-lingual transfer, limits the impact of data sparsity for LR languages and reduces over-fitting of HR languages as the monolingual models are trained last. Second, we base our training curriculum on a tree structure whose nodes indicate which languages are included in the training data at each step of the iteration. Using phylogenetic knowledge, we group together High- and Low-Resource languages which are either closely related or distant. In this way, we can investigate how using different phylogenetic-based training strategies impact performance.

We apply our approach to 6 LR and 6 HR languages from two families (Germanic and Romance) and compare it to a multilingual model, monolingual models and a generate-and-translate pipeline. Overall, we observe improvement over both the multilingual and the monolingual approaches. In line with Soto Martinez et al. (2023)'s results, we find that the quality of the generate-and-translate approach varies with the quality of machine translation for the target languages. Finally, we observe similar performance for the two ways of grouping languages, which seems to confirm the intuition that training on related languages promotes transfer while training on distant languages facilitates regularisation.

2 Related Work

AMR-to-Text Generation beyond English. Using Europarl texts and silver AMRs derived from the English part of that corpus, Fan and Gardent (2020) train a multilingual AMR-to-Text generation model for 21 EU languages. They pre-train the graph encoder and the language models on millions of graph and monolingual sentences. The AMR-to-Text generation model is trained on 400K to 8.2M (graph, text) pairs depending on the target language. Focusing on the four languages of the AMR3.0 test set (German, Italian, Spanish, Chinese, LDC2020T07)[1], Ribeiro et al. (2021a) show that combining a large 1.9M dataset of (silver AMR, human-written text) pairs with a small dataset of 36.5K (gold AMR, machine-translated text) pairs yield better results than using each dataset separately when fine-tuning mT5_{base}. Xu et al. (2021) extend Ribeiro et al. (2021a)'s work using multi-task learning. Their model is first pre-trained on six tasks (AMR-to-English, English-to-AMR, English-to-X, X-to-English, AMR-to-X, and X-to-AMR) with millions of (silver AMR, human-written text) pairs. It is then fine-tuned on 2 tasks (AMR-to-X and English-to-X) on 36.5K (gold AMR, gold English, machine-translated X text). Evaluating on German, Spanish and Italian, they show that their approach outperforms previous work. Martínez Lorenzo et al. (2022) fine-tune a model using 55.6K (gold AMRs, machine-translated text) pairs. Their model is based on SPRING (Bevilacqua et al., 2021), a bidirectional AMR-to-text and text-to-AMR model pretrained on 200K (silver AMR, human-written English text) and fine-tuned on the AMR3.0 data for English.

Different from these approaches, we consider both high- and Low-Resource languages, restrict our approach to a Low-Resource scenario and propose a novel training strategy to derive monolingual models from a multilingual one.

Curriculum learning. Bengio et al. (2009) showed that curriculum learning can lead to improved performance over a random training order and Xu et al. (2020) propose a dynamic curriculum learning approach that relies on training loss and model competence to increase the difficulty of the training samples shown to the model. To train their massively multilingual machine translation model, the NLLB Team et al. (2022) use a curriculum learning approach in which LR languages are introduced later into the training pool. They show that this helps reduce over-fitting for these languages. Similarly, Kuwanto et al. (2023) propose a curriculum learning approach where the model is first pretrained on monolingual data for

[1] https://catalog.ldc.upenn.edu/LDC2020T07

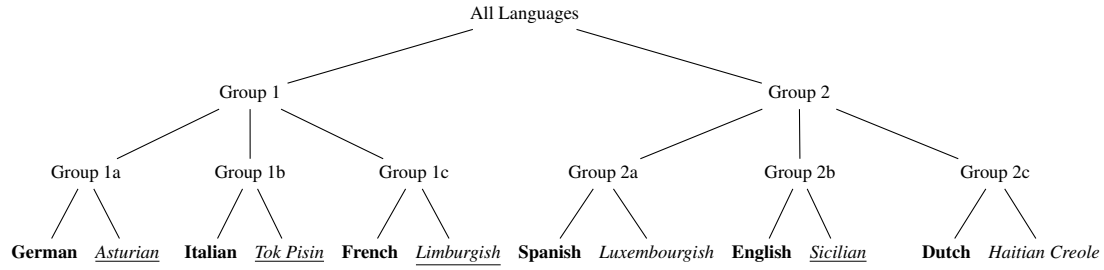
English and a target LR language as well as synthetic code-switching data in a second step.

We expand on these approaches by proposing a tree-structured curriculum where the nodes indicate the set of languages used at each step of the curriculum.

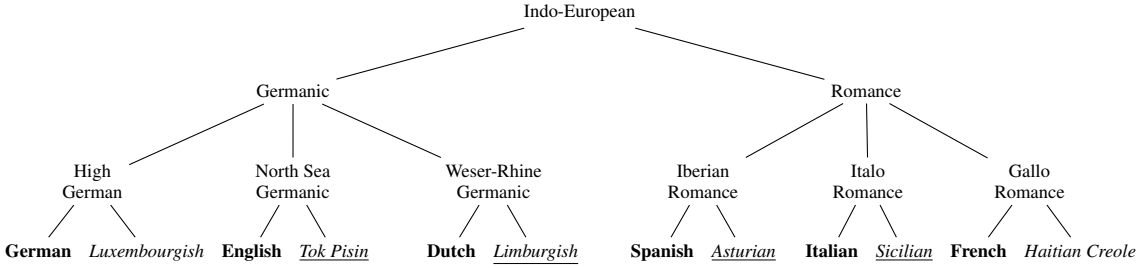
Exploiting Phylogenetic Knowledge. As illustrated in Figure 2b, a language phylogenetic tree highlights the proximity or distance between languages. Previous works have shown that phylogenetic knowledge can be leveraged to improve the performance of multilingual models, particularly for LR languages. Neubig and Hu (2018) show that training machine translation models on a pair of closely related high- and Low-Resource languages improves performance on LR languages. Faisal and Anastasopoulos (2022) stacked bottleneck adapters (Houlsby et al., 2019) for different levels of a phylogenetic tree to tackle diverse NLU tasks (dependency parsing, part of speech tagging, and natural language inference) on a variety of languages. Soto Martinez et al. (2023) used a soft prompt-inspired technique (Lester et al., 2021) to provide a model with information about the phylogenetic tree on RDF-to-Text generation of Celtic languages. For AMR-to-Text, Fan and Gardent (2020) noted that training on a pair of closely related languages of the same language family yields better results than training on a pair of languages from the same family that are more distant. Finally, Meng and Monz (2024) studied transfer learning in machine translation models and noted that closely related languages have a strong transfer effect and that augmenting the number of related languages further enhances performance. Interestingly, they also observed that introducing a balanced amount of distant language instances during training can provide unexpected regularizing effects.

Following up on these approaches, we use phylogenetic knowledge to guide curriculum learning and we study the effect of grouping closely related languages as well as grouping distant languages.

Low-Rank Adaptation. Hu et al. (2021) introduced Low-Rank Adaptation (LoRA), a Parameter-Efficient Fine-Tuning (PEFT) alternative to standard bottleneck adapters and prompt tuning approaches. Evaluating on multiple NLG datasets for summarization and Data-to-Text Generation, they showed their approach outperformed Full Fine Tuning (FFT) and matched or outperformed other PEFT techniques on GPT-2 models (Radford et al., 2019). Following Faisal and Anastasopoulos (2022), we propose to train a LoRA adapter for each iterative step of our curriculum learning training, stacking them as we go.



(a) Distant Languages Hierarchy (DLH)



(b) Phylogenetic Tree Hierarchy (PTL)

Figure 2: Training hierarchies tested. The top one (DLH) maximizes the language difference within nodes of each level. The bottom one (PTL) minimizes the language difference within nodes of each level. High-Resource languages are in **bold**, Low-Resource languages are in *italics* and languages unseen by the pretrained base model are underlined.

3 A brief overview of LoRA and QLoRA

LoRA is a Parameter-Efficient Fine-Tuning approach where, during training, the weights of the original base model (W_0) are frozen and two low-rank, trainable, decomposition matrices (A and B) are added to selected layers of the model, reworking the output hidden state of the layers (h) to the addition of the original weights and the product of the low-rank matrices (AB) as shown in Equation 1.

$$h = W_0 x + ABx \quad (1)$$

AB happens to be a good approximation of a full fine-tuning weight update while requiring fewer parameters to be trained. Notably, after having trained A and B on some task or language, we can compute their final product (AB) and merge this product into the original weights (W_0) via simple matrix addition thereby creating a new model specialised for the target task or language. Thus the same model can be iteratively fine-tuned on multiple tasks or languages. In our approach, we start from a pre-trained multilingual model and iteratively derive 12 monolingual models from this initial model in 4 steps, starting by fine-tuning this model tuned on 12 languages (Step 0) and iteratively fine-tuning models for 6, 2 and 1 languages (Steps 1, 2 and 3).

By merging the weights of the original model with the parameters learned in the LoRA matrices, the final models have no inference overhead, which distinguishes LoRA from other PEFT approaches. Furthermore, since LoRA matrices are smaller than the base model, LoRAs for multiple tasks or languages can be trained and switched faster and without requiring as much storage space as other approaches.

Another advantage of LoRA adaptation is that it lowers the memory requirements for fine-tuning very large models compared with full fine-tuning. To further reduce memory requirements during training, Dettmers et al. (2024) proposed QLoRA, where unquantized LoRA modules are applied to a quantized model. While training quantized weights is unstable (Wortsman et al., 2023), only training the few unquantized weights of the LoRA module makes this approach stable.

4 Task

We aim to verbalise AMR graphs into both high- and Low-Resource languages. To factor out the impact of training data size, we keep this size constant across languages restricting the number of distinct training instances per language to 31K, the Lower bound set by the language with fewer resources. In this way, differences between languages can be traced back to differences between models and training strategies rather than to the size of the available data for each language.

For our experiments, we select a combination of 6 Low- and 6 High-Resource languages (as classified by the NLLB Team et al. (2022)). We select these languages so that they can be grouped in a balanced phylogenetic tree (see Figure 2b). Table 1 includes further information about the selected languages noting in particular, how much training data per language was seen by our underlying pretrained $\text{mT5}_{\text{large}}$ base model.

Language	Code	H/L	% PT Data
German	DEU	High	3.05%
Luxembourgish	LTZ	Low	0.68%
English	ENG	High	5.67%
Tok Pisin	TPI	Low	**0.00%**
Dutch	NLD	High	1.98%
Limburgish	LIM	Low	**0.00%**
Spanish	SPA	High	3.09%
Asturian	AST	Low	**0.00%**
Italian	ITA	High	2.43%
Sicilian	SCN	Low	**0.00%**
French	FRA	High	2.89%
Haitian Creole	HAT	Low	0.33%

Table 1: Target languages, their ISO 639-3 code, whether they are high- or Low-Resource (H/L) languages, and how much of the base model pretraining data (PT Data) they cover.

5 Hierarchical QLoRA (HQL)

To mitigate the effects of data scarcity (over-fitting) and multilingual training (noise), we propose a variation of curriculum learning that leverages both phylogenetic knowledge and the modularity and memory efficiency of LoRAs to iteratively refine a base multilingual model into a set of monolingual models.

Base Model. Our base model is $\text{mT5}_{\text{large}}$ (Xue et al., 2021)[2], a multilingual encoder-decoder model which we extend with LoRA modules to support modular Parameter-Efficient Fine-Tuning and 4-bit quantization to reduce memory footpring during training.

Refining Models. We learn 12 monolingual models by iteratively fine-tuning a model trained in

[2] https://huggingface.co/google/mt5-large

12 languages in four steps as follows. In the first step (Level 0), the base model ($mT5_{large}$) is fine-tuned on 12 languages using LoRA fine-tuning. The resulting model – which is created by merging $mT5_{large}$'s weights with the A and B matrices as explained above – is then fine-tuned on two sets of 6 languages yielding two 6-language models, each trained with a separate LoRA module (Level 1). We repeat this process twice: first, fine-tuning the two 6-language models into 6 bilingual models (Level 2) and second, fine-tuning each of the bilingual models into 12 monolingual models (Level 3). Algorithm 1 in Appendix A specifies our training strategy in more detail.

Choosing Language Groups. Which set of languages should be used at each step of the iteration? Our training strategy follows a four-level deep tree where each node in the tree determines the set of languages used for fine-tuning the parent model. Based on previous work, we compare the effect of two training hierarchies as shown in Figure 2.

Meng and Monz (2024) showed that balanced amounts of data from distant languages during training can act as a regularizing factor. Accordingly, our first strategy consists in increasing the average distance between languages for each node in our training hierarchy. This produces the Distant Languages Hierarchy depicted in Figure 2a.

Conversely, multiple previous studies have pointed to the benefits of training multilingual models on closely related languages (cf. Section 2). Based on this, our second training hierarchy follows the phylogenetic tree shown in Figure 2b where at each level of the hierarchy, the corresponding LoRA module is trained on smaller, less diverse and more closely related groups of languages. Under this Phylogenetic Tree Hierarchical QLoRA (PTHQL) approach, the expectation is to increase the transfer learning and reduce the noise of other languages as training progresses.

6 Experimental Setup

6.1 Data

As parallel (AMR, text) data only exists for a restricted set of languages, we use both machine translation and AMR-parsing to create multilingual training and test data.

Training Data. The AMR 3.0 dataset (Knight, Kevin et al., 2020)[3] includes 55.6K (gold AMR, human-written text) pairs where the texts are in English. We create training data for our target languages using machine translation and language identification scores as follows. First, we translate the English texts to our target languages using a 4-bit quantized NLLB-3.3B model (NLLB Team et al., 2022)[4]. Second, we filter the machine-translated texts using the GlotLID (Kargaran et al., 2023)[5] language identification model and removing all instances with a score less than 0.5. Third, we keep the top 31K instances for each language so that the quantity of training data is the same for all languages. This yields a dataset of 31K (gold AMR, machine-translated texts) for each of our target languages except English where texts are human-written.

In addition, we create a small parallel dataset for all our target languages where the AMR are silver and the texts are human-written. We derive this dataset from the FLORES-200 dataset of parallel texts (NLLB Team et al., 2022) and obtain silver AMR graphs by parsing the English texts of this dataset using AMR3-structbart-L (Drozdov et al., 2022)[6]. Since FLORES-200 does not include training data, we used the validation data for training. We then split the test data in half to create two small validation and test sets.

Test Data. We evaluate on (gold AMR, human-written text) for English, German, Spanish and Italian using LDC2020T07 (Damonte and Cohen, 2018; Damonte, Marco and Cohen, Shay, 2020)[7], which is a subset of AMR3.0 with gold AMR graphs and human translated and corrected texts. For the remaining 8 languages, we used our subset of the FLORES-200 test set of 506 (silver AMR, human-written text) pairs. While we could instead have used (gold AMR, machine-translated texts) derived from AMR3.0, we prefer to use silver AMR graphs paired with human-verified sentences. The rationale behind this decision is that the noise introduced by an AMR parser when producing the silver AMR graphs will be uniform across all tested languages, whereas the noise that machine-translated silver sentences have would vary across languages given the uneven performance of machine translation models. Table 2 summarizes the size and type

[3] https://catalog.ldc.upenn.edu/LDC2020T02

[4] https://huggingface.co/facebook/nllb-200-3.3B

[5] https://github.com/cisnlp/GlotLID

[6] https://github.com/IBM/transition-amr-parser/

[7] https://catalog.ldc.upenn.edu/LDC2020T07

of our data.

Dataset	Quality		Instances per Language		
	AMR	Text	Train	Test	Valid
FLORES-200	Silver	Gold	997	506	506
AMR 3.0	Gold	Silver	30 000	1 000	1 000
AMR3.0	Gold	Gold	N/A	1 371	N/A

Table 2: Our final datasets after preprocessing.

6.2 Training

Implementation Details. All our experiments are done using $\text{mT5}_{\text{large}}$ as the underlying base model via the Transformers [8] library. We use the PEFT [9] library to handle the LoRA implementation. The model is quantized to 4-bit precision for memory efficiency. Following (Dettmers et al., 2024), we apply LoRA to all linear layers of the model as this was shown to improve performance. Both Rank and Alpha are set to 256 using Rank-Stabilized scaling, these high values are selected given the model's need to learn both an entirely new task (AMR-to-Text vs Spam Correction) as well as generate into scarcely seen and previously unseen languages. As pointed out by Hu et al. (2021) new languages and tasks might require much higher ranks. The base model contains around 1.2B parameters and introducing the LoRA adds almost 300M new trainable parameters.

Training Scheme. We use a batch size of 8 and a maximum length per training instance of 256 tokens, which is similar to the values chosen by Ribeiro et al. (2021a) while keeping the total batch size as a power of 2 which benefits the training speed. This limit implies the truncation of around 8% of tokens on the input sequence but does not affect the output sequences.

To factor out the impact of training data size, we train each model on the same amount of data. For each language, we have 30 997 distinct instances and we train for one epoch on each level of the training hierarchy. Thus L0 models are trained on 371 964 (= 30 997 $\times$ 12) unique instances, L1 models on 185 982 instances, L2 on 61 994 instances and L3 on 30 997 instances. Hence by the end of the training, each monolingual model has seen 650 937 instances in total, with unique instances being seen 4 times across models, which is equivalent to 4 epochs on the full dataset.

[8] https://huggingface.co/docs/transformers
[9] https://huggingface.co/docs/peft

It is worth noting that, given the modularity of LoRAs and the way we can reuse the intermediate levels in the training of the new ones, the total number of instances used for training all 12 monolingual models is 1 487 856. In comparison, without our approach, directly fine-tuning 12 monolingual models that have seen 650 937 instances would require training on 7 811 244 instances (= 650 937 $\times$ 12). As explained in section 5, we consider two training hierarchies, the Distant Languages Hierarchy and a Phylogenetic Tree Hierarchy. A summary of all training hyperparameters can be found in Table 5 in Appendix B.

6.3 Models

We compare our approach with previous work and with three strong baselines.

6.3.1 Previous Work

F&G (Fan and Gardent, 2020) is an Encoder-Decoder multilingual model that supports 21 High- and Medium-Resource languages. The encoder includes structural embeddings and the model was fine-tuned on (silver AMR, human-written text) pairs with data sizes ranging from 400K to 8.2M pairs depending on the target language.

Ribeiro (Ribeiro et al., 2021a) is a mT5_{base} model that supports 4 HR languages and was fine-tuned on millions of (silver AMR, human-written text) and tens of thousands of (human AMR, machine-translated text) pairs for each target language.

Xu (Xu et al., 2021) consists of 3 Transformer models trained separately on 3 HR languages using multi-task pretraining on 6 tasks (AMR-to-English, English-to-AMR, English-to-X, X-to-English, AMR-to-X, and X-to-AMR) with millions of (silver AMR, human-written text) pairs. The models are then fine-tuned on 2 tasks (AMR-to-X and English-to-X) on 36.5K (gold AMR, gold English/machine-translated X text).

Martinez (Martínez Lorenzo et al., 2022) the $\text{mBART}_{\text{large}}$ model trained separately on 4 HR languages. We use the version trained on plain AMR inputs which was trained for up to 30 epochs on 55K (gold AMR, machine-translated text) pairs.

6.3.2 Baselines

Monolingual QLoRA (MonoQL). 12 monolingual models obtained by fine-tuning $\text{mT5}_{\text{large}}$ on each language separately using LoRA. We expect this model to perform worse than ours, particularly

on LR languages, due to the limited training data which can lead to either a lack of generalization or to over fitting. Each final model of our HQL approach has seen 650 937 instances during training (subsection 6.2). To allow for a fair comparison, we train each *MonoQL* model with that many instances.

Multilingual QLoRA (MultiQL). Fine-tuned $mT5_{large}$ using LoRA on data from all 12 languages. We expect this model to perform worse than ours due to the noise from the language mix. Since our HQL models are trained on 1 487 856 instances (cf. subsection 6.2), we let this multilingual model train up to that many instances.

Generate and Translate (Gen&Trans). We generate from AMR-to-English using the English *MonoQL*. Then we translate that output into the target languages with the same model used to generate our silver data (4-bit quantized NLLB-3.3B). We expect this model to mirror the uneven quality of machine translation models, performing well in HR but less well in LR languages.

6.4 Metrics

Following NLLB Team et al. (2022), we use BLEU, a simple surface-based metric that does not rely on training data, which is an advantage when dealing with multiple languages, particularly low-resource ones. We compute the scores with SacreBLEU (Post, 2018)[10] and the default settings (including *13a* tokenizer) for comparability with previous works. We also report Chrf++ and BLEURT [11] scores in Appendix C, however we discuss mostly BLEU given its widespread use in the past, being the only metric available on all previous works that use the same test as we do. We compute statistical testing via paired bootstrap resampling (Koehn, 2004) for BLEU and ChrF++ and Wilcoxon signed-rank test (Wilcoxon, 1945) for BLEURT-20 and report them on Appendix D.

7 Results

We report results obtained when generating from both Silver and Gold AMR comparing our approach with previous works and baselines and examining results on both High- and Low-Resource languages.

HQL outperforms or is on par with mono and multilingual baselines (Silver and Gold AMRs). On silver AMRs, HQL models are consistently better than both the mono and the multilingual baselines, except for Tok Pisin (Figure 3, Table 3, Figure 4). Statistical tests (Appendix D) confirm that the difference is statistically significant in most cases. On gold AMRs, the results are more mixed.

[10] https://github.com/mjpost/sacrebleu
[11] https://github.com/google-research/bleurt

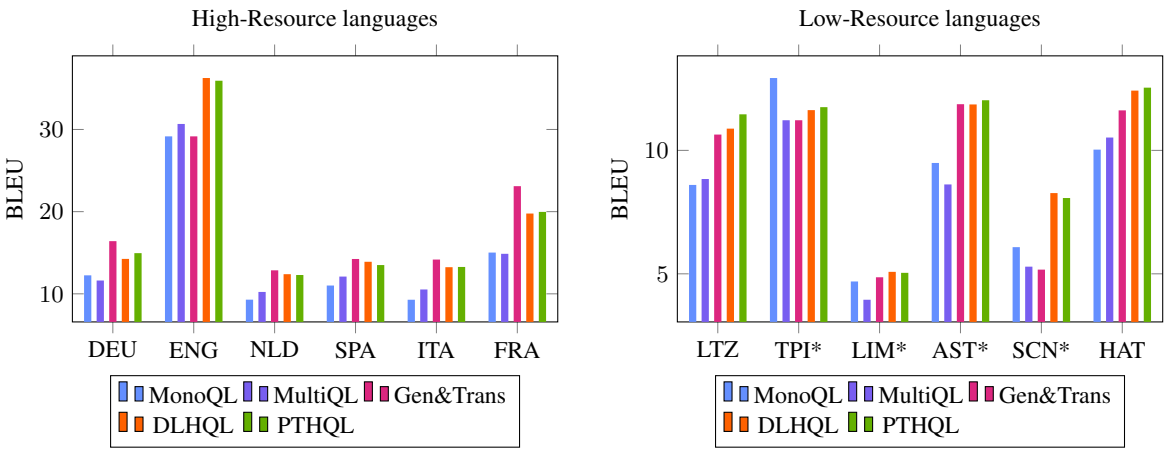


Figure 3: BLEU score on our sub set of FLORES-200 test data. *Languages unseen by the $mT5_{large}$ base model.

Model	DEU	LTZ	ENG	TPI	NLD	LIM	SPA	AST	ITA	SCN	FRA	HAT
MonoQL	12.2	8.6	29.2	**12.9**	9.3	4.7	11.0	9.5	9.3	6.1	15.0	10.0
MultiQL	11.6	8.8	30.7	11.2	10.2	4.0	12.1	8.6	10.5	5.9	14.9	10.5
Gen&Trans*	**16.4**	10.6	29.2	11.2	**12.9**	4.9	**14.2**	11.9	**14.2**	5.2	**23.1**	11.6
DLHQL	14.2	10.9	**36.3**	11.6	12.4	**5.1**	13.9	11.9	13.2	**8.3**	19.8	12.4
PTHQL	15.0	**11.5**	35.9	11.8	12.3	5.0	13.5	**12.0**	13.3	8.1	20.0	**12.5**

Table 3: BLEU score on our sub set of FLORES-200 test data. *English Gen&Trans is simply the result of MonoQL.

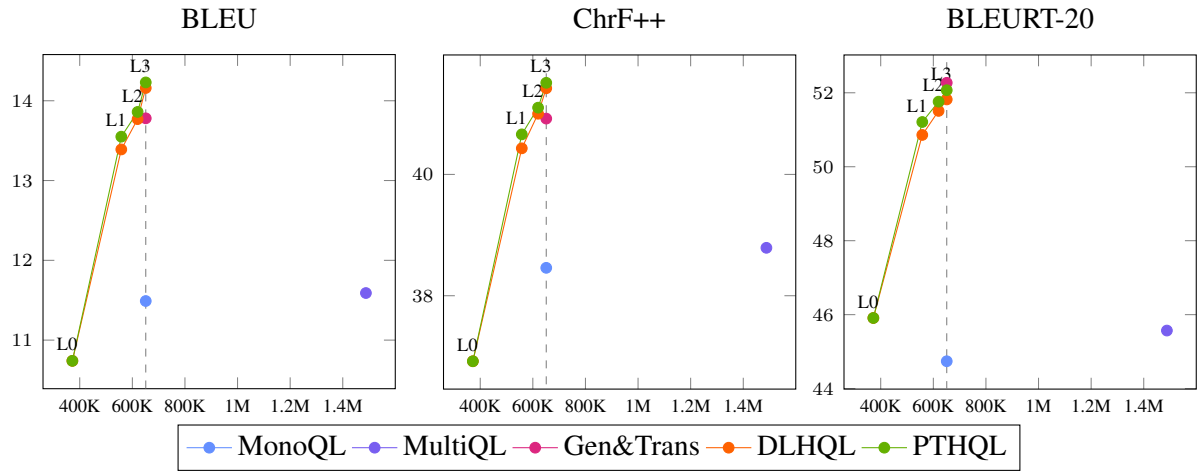


Figure 4: Average score (Y axis) across all 12 languages vs. total instances seen during training (X axis) for 3 metrics on our subs set of FLORES-200 test data. HQL models include results on all the intermediary levels of the hierarchy.

Our models outperform on Italian and German but not on English and Spanish - this is likely due to both languages being among the most represented in the pretraining data of the base model (Table 1).

HQL outperforms the Gen&Trans Baseline on all LR languages. While the Gen&Trans baseline outperforms our models on most HR languages (except English), our approach outperforms the Gen&Trans models on all LR languages (Figure 3). This shows the benefits of HQL for LR languages where MT yield low quality texts while our stacked LoRA approach seems to enhance transfer. Similar results are seen on other metrics (Appendix C) where HQL comes ahead in most LR languages.

We also see that two languages previously unseen by the base model (Tok Pisisn and Asturian) show a transfer effect as they perform on par with LR languages present in the base model's training data. For Limburgish and Sicilian, we conjecture that the low scores result from the low-quality of the machine translation as evidenced by the poor performance of the Gen&Trans baseline on these languages.

HQL optimizes faster than the three baseline models and on average, outperforms them all. Figure 4 plots the average BLEU, Chrf++, and BLEURT-20 score for all 12 languages against the number of instances seen during training. We see that already at level L2, our HQL models outperform all three baselines (monolingual, multilingual, Gen&Trans) on two of the metrics despite seeing fewer total training instances. The graph also shows that each new level of the hierarchy improves performance.

HQL performs on par with previous work (Gold AMRs). Table 4 compares our results with previous works on Gold AMRs. In HR Romance languages, our HQL approach outperforms all previous works, in English, the score is close to the best-performing model and in German, our model underperforms both Xu's and Lorenzo's approach - possibly due to differences in training data size and the impact of multi-task learning.

Model	DEU	ENG	SPA	ITA
F&G	15.3	24.9	21.7	19.8
Ribeiro	20.6	—	30.7	26.4
Xu	**25.7**	—	31.4	28.4
Martinez	23.2	44.8	34.6	29.0
MonoQL	18.2	**49.2**	38.6	22.7
MultiQL	19.8	42.9	34.1	27.2
Gen&Trans*	**28.0**	**49.2**	**39.6**	**33.8**
DLHQL	21.2	44.2	37.4	29.2
PTHQL	22.8	43.4	37.2	29.7

Table 4: BLEU score on LDC2020T07 test data. English Gen&Trans is simply the result of MonoQL.

HQL performs well compared to previous works despite being trained on fewer data. In previous work, *F&G, Ribeiro* and *Xu* trained on 400k to 8.9M synthetic training pairs per language while the *Martinez* model is trained for up to 30 epochs on close to 55K monolingual instances. In contrast, our models are trained on 4 epochs and less than 31K instances per language. Despite this, our models come close to and in some cases, outperform those previous approaches, while also enabling support for LR languages.

Distant vs. Close Languages. We observe almost no significant difference when training on distant (DLHQL) vs. closely related (PTHQL) languages. While this could confirm Meng and Monz (2024)'s observation that both are useful in inducing transfer and regularisation respectively, this could also be due to the restricted size of our training tree since because of computation constraints, we limited ourselves to a small number of languages which induces a strong overlap of training data between the two hierarchies: 100% on L0 and L3, 50% on L1 and L2, for a total training overlap of 81%. To further evaluate the difference between this approaches, future studies could reduce the overlap by selecting a larger hierarchy or by starting with a reduced number of instances and increasing their number as the training progresses through the levels.

8 Conclusion

We proposed a novel approach for multilingual AMR-to-Text generation and showed that it significantly outperforms fully monolingual and fully multilingual approaches. We demonstrated that, on LR languages, it can outperform a Gen&Trans approach, despite most training data being machine-translated. We compared different techniques for selecting a training hierarchy and found that, while the Phylogenetic approach usually achieves better results than the distant languages approach, differences were not significant.

9 Acknowledgments

We thank the anonymous reviewers for their feedback. We gratefully acknowledge the support of the French National Research Agency (Gardent; award ANR-20-CHIA-0003, XNLG "Multilingual, Multi-Source Text Generation"). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

10 Ethical Considerations

While there have been significant advances in multiple NLP tasks over the last couple of years, these benefits tend to focus on High-Resource languages. By researching how to improve performance over a more diverse set of languages we hope to make the field more inclusive and democratize the technology. This seems to us particularly relevant in Graph-to-Text tasks, which help verbalize text into more languages. Despite all these advantages, we are still aware of the shortcomings of these technologies. Current models are capable of generating inaccurate text and misleading users in High-Resource languages, and they remain even more unreliable on Low-Resource tasks.

Supplementary Materials Availability Statement: All the required code and data can be obtained, although some of the data is not free. Our source code for training the models can be found at https://gitlab.inria.fr/wsotomar/HQL-Hierarchical-QLoRA. The NLLB-200-3.3B model used for Machine Translation is available at https://huggingface.co/facebook/nllb-200-3.3B. The AMR3-structbart-L semantic parser is available at https://github.com/IBM/transition-amr-parser/. The Flores-200 data is available at https://huggingface.co/datasets/facebook/flores. The AMR 3.0 dataset (LDC2020T02) is available at https://catalog.ldc.upenn.edu/LDC2020T02. AMR 3.0 - 4 Translations dataset (LDC2020T07) is available at https://catalog.ldc.upenn.edu/LDC2020T07.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Sym-

metric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Damonte, Marco and Cohen, Shay. 2020. Abstract meaning representation 2.0 - four translations.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim, and Ramón Astudillo. 2022. Inducing and using alignments for transition-based AMR parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1086–1098, Seattle, United States. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Angela Fan and Claire Gardent. 2020. Multilingual AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Alexander Miserlis Hoyle, Ana Marasović, and Noah A. Smith. 2021. Promoting graph awareness in linearized graph-to-text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 944–956, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Knight, Kevin, Badarau, Bianca, Baranescu, Laura, Bonial, Claire, Griffitt, Kira, Hermjakob, Ulf, Marcu, Daniel, O'Gorman, Tim, Palmer, Martha, Schneider, Nathan, and Bardocz, Madalina. 2020. Abstract meaning representation (amr) annotation release 3.0.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, Alex Jones, and Derry Wijaya. 2023. Low-resource machine translation training curriculum fit for low-resource languages. In *PRICAI 2023: Trends in Artificial Intelligence: 20th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2023, Jakarta, Indonesia, November 15–19, 2023, Proceedings, Part III*, page 453–458, Berlin, Heidelberg. Springer-Verlag.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Abelardo Carlos Martínez Lorenzo, Marco Maru, and Roberto Navigli. 2022. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1727–1741, Dublin, Ireland. Association for Computational Linguistics.

Yan Meng and Christof Monz. 2024. Disentangling the roles of target-side transfer and regularization in multilingual machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1828–1840, St. Julian's, Malta. Association for Computational Linguistics.

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.

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. *Preprint*, arXiv:2207.04672.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Leonardo F. R. Ribeiro, Jonas Pfeiffer, Yue Zhang, and Iryna Gurevych. 2021a. Smelting gold and silver for improved multilingual AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 742–750, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021b. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021c. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo and Thiago Pardo. 2022. Exploring a POS-based two-stage approach for improving low-resource AMR-to-text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 531–538, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2023. Phylogeny-inspired soft prompts for data-to-text generation in low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–198, Nusa Dua, Bali. Association for Computational Linguistics.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. biom bull 1 (6): 80–83.

Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. 2023. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36:10271–10298.

Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2020. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3977–3989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2021. XLPT-AMR: Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. 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 896–907, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual

pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.