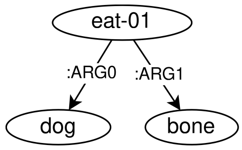
Should Cross-Lingual AMR Parsing go Meta? An Empirical Assessment of Meta-Learning and Joint Learning AMR Parsing

Anonymous ACL submission

Abstract

Cross-lingual AMR parsing is the task of predicting AMR graphs in a target language when training data is available only in a source language. Due to the small size of AMR training data and evaluation data, cross-lingual AMR parsing has only been explored in a small set of languages such as English, Spanish, German, Chinese, and Italian. Taking inspiration from Langedijk et al. (2022), who apply meta-learning to tackle cross-lingual syntactic parsing, we investigate the use of meta-learning for cross-lingual AMR parsing. We evaluate our models in k-shot scenarios (including 0-shot) and assess their effectiveness in Croatian, Farsi, Korean, Chinese, and French. Notably, Korean and Croatian test sets are developed as part of our work, based on the existing *The Little Prince* English AMR corpus, and made publicly available. We empirically study our method by comparing it to classical joint learning. Our findings suggest that while the meta-learning model performs slightly better in 0-shot evaluation for certain languages, the performance gain is minimal or absent when k is higher than 0.



(a) AMR graph

```
(eat-01
 :ARG0 (dog)
 :ARG1 (bone))
```

(b) linearized AMR

Figure 1: “The dog eats a bone.”

1 Introduction

Abstract Meaning Representation (Banarescu et al., 2013, AMR) represents the meaning of texts as rooted and directed acyclic graphs. AMR graphs capture the underlying semantics of input texts while abstracting away from their syntactic realizations. Nodes in AMR graphs are not explicitly mapped to their input token. Hence, it is an unanchored formalism. AMRs are widely used to enhance the capabilities of NLP systems such as question answering (Deng et al., 2022; Kapanipathi et al., 2021), text summarization (Liao et al., 2018; Liu et al., 2015), or human-robot interaction (Bonial et al., 2019, 2023).

AMR was originally designed for English texts only. However, Damonte and Cohen (2018) demonstrated that AMR could be used for other languages such as Spanish, Italian, Chinese, and German. Since then, many approaches have adopted AMR parsing for multilingual AMR parsing (Procopio et al., 2021; Blloshmi et al., 2020; Xu et al., 2021; Cai et al., 2021; Sheth et al., 2021). However, one of the main challenges for this task is the lack of data. Currently, training data are only available in English (Knight et al., 2017, 2020) and evaluation data in 6 languages: English, German, Spanish, Italian, Chinese (Damonte and Cohen, 2018; Li et al., 2021),[1] and French (Kang et al., 2023). To overcome the lack of training data in target languages, previous approaches create silver training data in the target languages. This is done through machine translation (Damonte and Cohen, 2018; Blloshmi et al., 2020) under the assumption that a text conveying the same meaning should have a shared AMR graph across languages. Similarly, parallel corpora with English AMR parsers are also employed to create silver data (Xu et al., 2021; Blloshmi et al., 2020). Another approach uses English data for training and then evaluates the model in the target language in a zero-shot manner (Procopio et al., 2021). Since evaluation data is available in five languages, most of these proposals focus on this small set of languages.

In this study, our goal is to apply AMR parsing for more diverse languages that have been less explored in previous work and tackle the lack of training data with k-shot learning. Taking inspiration from Langedijk et al. (2022), who applied

[1] In Chinese AMR 2.0 (Li et al., 2021), AMR concepts are annotated in Chinese.

meta-learning for k-shot cross-lingual syntactic parsing, we apply meta-learning for cross-lingual AMR parsing. To examine the efficiency of the method, we compare the meta-learning approach to a classical joint learning method.

Our contributions to cross-lingual AMR parsing are as follows:

- This work presents the **first empirical study on meta-learning applications on cross-lingual AMR parsing**.
- We train and evaluate our model in languages less explored for AMR parsing: Korean, Croatian, French, and Farsi.
- We **publish new evaluation data in Korean and Croatian**, based on *The Little Prince*.
- We release a multilingual AMR parser that can be evaluated in many languages in k-shot. We also release the code to train and evaluate the model.

2 Meta Crosslingual AMR

Seq2seq AMR Parsing In sequence-to-sequence AMR parsing (Bevilacqua et al., 2021), AMR parsing is viewed as generating a sequence of tokens representing AMR nodes and edges. AMR graphs should be first linearized in a single-line format (see Figure 1) to feed it to a sequence-to-sequence model. We linearize AMR graphs following van Noord and Bos (2017), which includes light pre-processing such as removing variables and wiki link.[2] We refer the readers to van Noord and Bos (2017) for a comprehensive understanding of the linearization process. To generate AMR graphs from multi-lingual inputs, we employ mBart (Tang et al., 2020) model, a pre-trained multilingual sequence-to-sequence model, as done by Procopio et al. (2021).

MAML for Cross-lingual AMR Parsing We use MAML (Finn et al., 2017) for cross-lingual AMR parsing. MAML learns good initial parameters θ that can be tuned to unseen tasks with only a few optimization steps and a few training data examples. MAML trains a model to be good at adapting to new tasks only with a few examples by *simulating the k-shot training and evaluation* during the training. We apply MAML to train our multilingual AMR parser so that it adapts quickly to new tasks, which are in our case, new languages. The training procedure is described below.

Step 1: At each iteration step, the initial model (Θ) is copied once per language i. For each i, $2 \times K$ examples are randomly sampled from D_i^{train} and divided into the support and the query set (K each). Using the support set, the model is temporarily updated with stochastic gradient descent with learning rate α (Eq. 1). Iterate through the support set for P adaptation steps to obtain Φ_i:

$$\Phi_i \leftarrow \Theta - \alpha \bigtriangledown_{\Theta} \mathcal{L}(\Theta_i). \tag{1}$$

Next, the loss is computed to evaluate the temporary model Φ_i on the query set. The loss $\mathcal{L}_i(\Phi_i)$ is saved for the next step. The entire step is called an 'inner loop' and the inner loop is repeated over the entire task batch, that is, for the number of all training languages I.

Step 2: $\mathcal{L}_i(\Phi_i)$ is summed up over training languages to update the initial model Θ by stochastic gradient descent with a learning rate β. This entire step is called an 'outer loop':[3]

$$\Theta \leftarrow \Theta - \beta \sum_i \bigtriangledown_{\Phi_i} \mathcal{L}_i(\Phi_i). \tag{2}$$

Step 3: Repeat Step 1 and Step 2 until the total number of training steps.

3 Experimental Setup

Silver Training/Validation Data We aim to train a multilingual AMR parser that adapts quickly to new languages, specifically French, Chinese, Korean, Farsi, and Croatian, with k examples. Our method is similar to that of Langedijk et al. (2022) in applying meta-learning for a k-shot cross-lingual parsing task, but our training data is only available in English, whereas they have multilingual training data. To create multilingual training data, we apply machine translation as in previous approaches (Damonte and Cohen, 2018; Xu et al., 2021; Blloshmi et al., 2020). We adopt DeepL[4] and translate English AMR training data (Knight et al., 2020, LDC2020T02) into 13 languages: German, Italian, Romanian, Finnish, Russian, Turkish, Japanese, Czech, Dutch, Polish, Swedish, Estonian, and Indonesian. The 13 languages were chosen for compatibility with our training model, mBart

[2] We employ the implementation code available at https://github.com/RikVN/AMR for graph preprocessing and post-processing.

[3] We apply First-Order MAML to avoid computation overhead (second-order derivative requires heavy computation)

[4] https://www.deepl.com

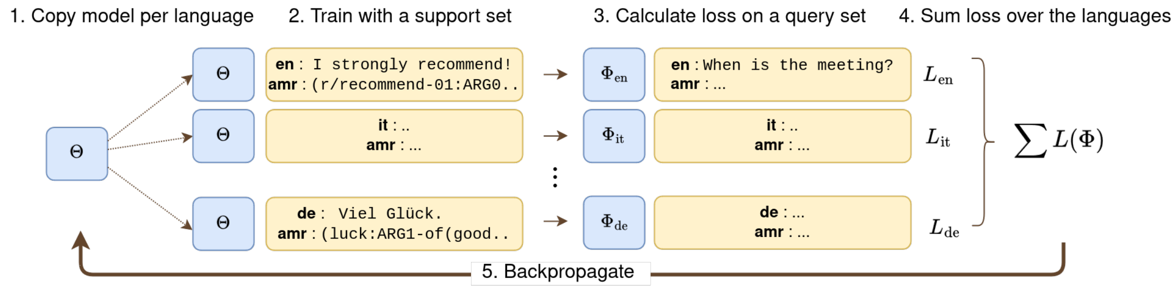


Figure 2: One training step for MAML cross-lingual AMR parsing.

(Tang et al., 2020), and for language diversity. They cover 5 language families: Indo-European (Germanic, Romance, Slavic), Uralic, Turkic, Japonic, and Austronesian. For each training language, there are 55,635 pairs of sentences and their corresponding AMR graph. To assess the translation quality, we evaluated the training data with the reference-free evaluation metric COMET (Rei et al., 2020). The COMET score of 13 languages is 83.8±0.8. We use a total of 14 languages including English for our training data. We use Spanish as the validation language and use the Spanish evaluation set from AMR 2.0 (Damonte and Cohen, 2020). For k-shot evaluation during the validation and test step, k random examples from the English dev set are translated to each evaluation language.

Gold Test Data We evaluate our model in French, Chinese, Korean, Farsi, and Croatian. For French, Chinese, and Farsi, we employ *The Little Prince* AMR corpus annotated in each language, respectively from Kang et al. (2023), `https://amr.isi.edu/` and Takhshid et al. (2022).[5] For Croatian and Korean, we create our test sets by manually aligning *The Little Prince* corpus in each language to corresponding English AMR graphs. After manual alignment, we excluded pairs exhibiting semantic discrepancies between the aligned sentence and its English counterpart, such as pairs where additional or omitted information was observed in the aligned sentences.[6] This leaves us with, respectively, 1,527 and 1,543 pairs for Korean and Croatian. A few examples of the final dataset are given in Appendix A. We make the test set publicly available.[7]

Meta-Training and Evaluation We adopt `mBart-large-50` model (Tang et al., 2020) from the transformers library (Wolf et al., 2020) to train our multilingual AMR parser. To implement model-agnostic meta-learning, we employ the `learn2learn` library (Arnold et al., 2020). Parameters used for the training are provided in Appendix B. Our goal is to evaluate the model's performance in new languages that were not seen during the training, specifically, French, Chinese, Korean, Farsi, and Croatian. To this end, for both validation and testing, we employ k-shot learning, where the model is fine-tuned with k examples for the test language before evaluation. We report evaluation scores with varying k size using SMATCH (Cai and Knight, 2013), an evaluation metric for AMR graphs.

Baseline with Joint Learning We train a baseline model with a joint learning method for comparison with our approach. We use the same mBart model and the training data as described in Section 3. To assess the effectiveness of our method compared to joint learning, we carry out the two experiments in settings as similar as possible (e.g. training data, hyper-parameters, learning scheduler, k-shot evaluation). Hyperparameter details are given in Appendix B.

4 Results and Discussion

We assessed our model across five languages in k-shot learning. Table 1 displays the evaluation results for different shot settings (k) where $k = 0, 32, 128$. In the 0-shot evaluation, MAML demonstrates higher performance for most evaluation languages, except for Croatian. Nevertheless, the performance gap is minimal, making it difficult to

[5] The original Farsi dataset consists of AMR concepts in Farsi. Since we employ AMR graphs with English concepts, we use only the input texts of the corpus and graphs from the English AMR corpus.

[6] A native Korean speaker manually aligned and filtered the data. For Croatian, we automatically translated Croatian text into English with Google Translate (`https://translate.google.com/`) and checked the semantic discrepancy with its English counterpart.

[7] The URL will be provided upon publication.

draw firm conclusions regarding the method's advantage. In the k-shot evaluation, the performance gap between the two models diminishes, with either the average score showing no significant difference (128-shot) or the baseline model outperforming the MAML model (32-shot). These observations suggest that while MAML may offer benefits in 0-shot evaluation for certain languages, its advantage is not consistent across all languages. In k-shot learning scenarios, the benefit is minimal or null. On the other hand, the joint-learning method shows competitive results regardless of its methodological simplicity. We hypothesize that substantial overlap between inputs and outputs in the training data across languages has contributed to these results. Our training data comprises translations of AMR 3.0 into multiple languages, resulting in overlapped AMR graphs and shared patterns in input texts. In this context, the joint-learning model may learn the similarities between training data directly, allowing the model to learn the task more efficiently.

Surprisingly, both MAML and baseline models exhibit a performance decrease when fine-tuned in 32-shot, compared to not being fine-tuned at all. We hypothesize that the mBart pre-trained model has already enough knowledge of our target languages and fine-tuning the model with only a few examples in each language may impair the model's capacity. This could also be attributed to the domain difference between the fine-tuning dataset and the test dataset. The fine-tuning dataset includes content from general fields such as online forums, journals, and web blogs, whereas the test dataset consists of *The Little Prince*, a novel written in the 1940s. Consequently, the domain shift between the two datasets may have contributed to the model's inability to generalize effectively to the test domain.

We provide additional analysis of our models in Appendix C (effect of the number of considered languages and of the translation quality).

	fr	zh	ko	fa	hr	avg
base_0-shot	56.4	45.6	42.1	46.3	**51.4**	48.4
MAML_0-shot	**56.5**	**46.1**	**42.2**	**46.7**	50.8	**48.5**
base_32-shot	**56.3**	**45.4**	**42.0**	**46.1**	**51.3**	**48.3**
MAML_32-shot	55.5	45.1	41.1	45.9	48.9	47.3
base_128-shot	**56.5**	45.9	42.0	46.6	**51.5**	48.5
MAML_128-shot	56.0	**46.2**	**42.2**	**46.8**	51.3	48.5

Table 1: SMATCH scores of the baseline and the MAML model (k-shot evaluation).

5 Related Work

Meta-learning, also known as *learning to learn*, is a learning paradigm that allows a model to quickly learn a new task with only a few examples. This is made possible by the prior knowledge that the model has acquired through a series of different tasks. In cross-lingual applications, each task corresponds to a different language. The closest approach to ours is Langedijk et al. (2022), who adopt MAML for cross-lingual dependency parsing. They train a dependency parser on a set of languages using MAML and then evaluate the model on unseen languages to investigate the model's ability to adapt quickly. In contrast, we focus on a *semantic* parsing task with an unanchored formalism. In addition, they have multilingual training data at hand, whereas we generate our silver multilingual data by machine translation from English data. Another difference is that they use a graph-based biaffine model for parsing, whereas we use a seq2seq model with a linearized graph. Sherborne and Lapata (2023) applied meta-learning to cross-lingual SQL parsing. While useful at representing (and executing) database queries expressed in natural language, SQL is not a general-purpose semantic formalism like AMR. To the best of our knowledge, our work is the first to apply MAML for cross-lingual AMR parsing.

6 Conclusion

This study investigates the effectiveness of meta-learning compared to joint learning in cross-lingual AMR parsing. We assess our models across less-explored languages for AMR parsing, including French, Chinese, Korean, Farsi, and Croatian. To facilitate evaluation, we develop new test sets for Korean and Croatian and release the data to promote AMR parsing in diverse languages. Our findings reveal that meta-learning exhibits minor performance gain compared to joint learning in 0-shot evaluation. The small gain diminishes for k-shot learning (when $k > 0$). Consequently, our results suggest that the joint learning method serves as a robust baseline, while meta-learning appears to be a sub-optimal approach for cross-lingual AMR parsing. We believe that this research provides valuable insights into the comparative efficacy of meta-learning and joint learning in cross-lingual AMR parsing, offering important guidance for future developments in cross-lingual AMR parsers.

Limitations

Our model does not outperform a simple monolingual model which is trained with AMR data in the target language translated by a MT system. However, our approach can be explored for low-resource languages for which machine translation is not available. In addition, we did not apply grid search to find the best learning rates for the baseline models and used the same learning rate as done by Procopio et al. (2021), who also employed mBart for sequence-to-sequence cross-lingual AMR parsing. This could have affected the results in favor of meta-learning. Nonetheless, this does not affect our conclusion of the empirical study to reveal the weakness of the meta-learning approach for cross-lingual AMR parsing. This study does not include evaluation scores on the AMR 2.0 multilingual test set, which could help position our models relative to the state-of-the-art models. There are two motivations for the omission. Firstly, the Spanish test set in AMR 2.0 is already used as our validation set. Therefore, the AMR graphs (they are shared across the 4 languages) are already exposed during the validation step. Secondly, German and Italian, evaluation languages in AMR 2.0, are already included in our training data. Since our goal is to evaluate our model for unseen target tasks, evaluating our model on these languages is not coherent with the objective. Despite the limitations, we believe that our study empirically shows the constraints of meta-learning for cross-lingual AMR parsing and provides valuable insights into the meta-learning application in the task.

References

Sébastien M R Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. 2020. learn2learn: A library for Meta-Learning research.

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.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500, Online. Association for Computational Linguistics.

Claire Bonial, Julie Foresta, Nicholas C. Fung, Cory J. Hayes, Philip Osteen, Jacob Arkin, Benned Hedegaard, and Thomas Howard. 2023. Abstract Meaning Representation for grounded human-robot communication. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 34–44, Nancy, France. Association for Computational Linguistics.

Claire N. Bonial, Lucia Donatelli, Jessica Ervin, and Clare R. Voss. 2019. Abstract Meaning Representation for human-robot dialogue. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 236–246.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021. Multilingual AMR parsing with noisy knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2778–2789, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Marco Damonte and Shay Cohen. 2020. Abstract meaning representation 2.0 - four translations ldc2020t07.

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.

Zhenyun Deng, Yonghua Zhu, Yang Chen, Michael Witbrock, and Patricia Riddle. 2022. Interpretable amr-based question decomposition for multi-hop question answering.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks.

Jeongwoo Kang, Maximin Coavoux, Didier Schwab, and Cédric Lopez. 2023. Analyse sémantique AMR pour le français par transfert translingue. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts*, pages 55–62, Paris, France. ATALA.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, Alfio Gliozzo, Sairam Gurajada, Hima Karanam, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yunyao Li, Francois Luus, Ndivhuwo Makondo, Nandana Mihindukulasooriya, Tahira Naseem, Sumit Neelam, Lucian Popa, Revanth Gangi Reddy, Ryan Riegel, Gaetano Rossiello, Udit Sharma, G P Shrivatsa Bhargav, and Mo Yu. 2021. Leveraging Abstract Meaning Representation for knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3884–3894, Online. Association for Computational Linguistics.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2017. Abstract meaning representation (amr) annotation release 2.0 - linguistic data consortium.

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

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8520, Dublin, Ireland. Association for Computational Linguistics.

Bin Li, Liming Xiao, Yihuan Liu, Yuan Wen, Li Song, Jayeol Chun, Minxuan Feng, Junsheng Zhou, Weiguang Qu, and Nianwen Xue. 2021. Chinese abstract meaning representation 2.0 ldc2021t13.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Luigi Procopio, Rocco Tripodi, and Roberto Navigli. 2021. SGL: Speaking the graph languages of semantic parsing via multilingual translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 325–337, Online. Association for Computational Linguistics.

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.

Tom Sherborne and Mirella Lapata. 2023. Meta-learning a cross-lingual manifold for semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:49–67.

Janaki Sheth, Young-Suk Lee, Ramón Fernandez Astudillo, Tahira Naseem, Radu Florian, Salim Roukos, and Todd Ward. 2021. Bootstrapping multilingual AMR with contextual word alignments. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 394–404, Online. Association for Computational Linguistics.

Reza Takhshid, Razieh Shojaei, Zahra Azin, and Mohammad Bahrani. 2022. Persian abstract meaning representation.

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

Rik van Noord and Johan Bos. 2017. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. *CoRR*, abs/1705.09980.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

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.

A Aligned Data Samples

en In the book it said : " Boa constrictors swallow their prey whole , without chewing it .
ko 그 책에는 이렇게 씌어 있었다. "보아 구렁이는 먹이를 씹지도 않고 통째로 집어삼킨다
hr U knjizi je pisalo: »Udavi gutaju svoj plijen cijel cjelcat, bez žvakanja.

en I pondered deeply , then , over the adventures of the jungle .
ko 나는 그래서 밀림 속에서의 모험에 대해 한참 생각해 봤다.
hr Zatim sam mnogo razmišljao o prašumskim pustolovinama,

en The little prince , who asked me so many questions , never seemed to hear the ones I asked him .
ko 어린 왕자는 내게 많은 것을 물어보면서도 내 질문에는 귀를 기울이는 것 같지 않았다.
hr Činilo se da mali princ, koji mi je postavljao brojna pitanja, nikada ne čuje moja.

en I was more isolated than a shipwrecked sailor on a raft in the middle of the ocean .
ko 대양 한가운데에 떠 있는 뗏목 위의 표류자보다 나는 더 고립되어 있었다.
hr Bio sam usamljeniji od brodolomca na splavi usred oceana.

B Training Hyperparameters

Meta Crosslingual AMR We train our model for 30,000 steps and evaluate the model every 500 steps with the Spanish validation set. Early stopping is applied, terminating training if the dev SMATCH score fails to improve for more than 7,500 steps. The number of fine-tuning cycles, called an adaptation step, is denoted as P. Unless specified otherwise, we set $P = 0$ and $k = 0$ (0-shot learning). MAML requires two learning rates, one for the inner loop (α) and one for the outer loop (β). We conducted a grid search to identify an optimal learning rate set and used $\alpha = 1 \times 10^{-5}$, $\beta = 3 \times 10^{-5}$ throughout the experiments. For β, we use a linear learning rate scheduler with 1,500 warm-up steps. Unless specified otherwise, we apply 1×10^{-5} to fine-tune a model before validation/testing. At each iteration step during the training, $2 \times K$ are sampled to form a query and a support set for each training language. As a result, the batch size N equals $2 \times K \times I$, where I denotes the number of training languages. By default, we assign $K = 8$ and $I = 14$, unless stated otherwise.

Baseline Model For the training set, we use a concatenation of the multilingual AMR training sets described in 3. At each iteration step, we randomly select N training examples from the concatenated training sets to calculate the loss and optimize the model accordingly. For the rest of the hyperparameters and test/evaluation method, we apply the same settings as described as above (e.g. learning rate scheduler, k-shot size) except for the learning rate since maml requires two learning rates α and β whereas joint-learning requires only one. We use a uniform learning rate for training 3×10^{-5} with a linear scheduler with 1500 warm-up steps.

C Additional Analysis

We provide additional analysis of our approach focusing on how the training is affected by the number of training languages and translation sources. The results include 0-shot evaluation for both meta-learning and joint learning.

Q1: How does the number of languages affect the performance of the models?

To examine how the number of training languages impacts the model performance, we incrementally add more languages to the training data and we train three models respectively with 8, 12, and 14 languages. The first model is trained in German, English, Italian, Romanian, Russian, Turkish, Finnish, and Japanese. Then we add Czech, Dutch, Polish, and Swedish, and then finally we add Estonian and Indonesian. Note that for meta-learning, the batch size depends on the number of training tasks since we randomly sample K examples per language (batch size $= 2 \times K \times I$ where I denotes the number of training languages). To keep the batch size consistent across experiments while altering only the number of languages, when more than 8 languages are used for training, we randomly sample 8 languages per iteration step and select K training examples per language. Unless specified otherwise, each model is evaluated in a zero-shot manner for five languages: French, Chinese, Korean, Farsi, and Croatian.

Results Table 2 shows that both the MAML and baseline models have a positive correlation with the number of training languages. The baseline model has the largest gain when increasing the number of

	fr	zh	ko	fa	hr	avg
base_14langs	56.3	45.6	42.1	46.3	51.4	48.4
base_12langs	53.6	41.6	40.1	43.4	45.9	44.9
base_8langs	47.5	39.8	39.1	40.5	22.4	37.8
MAML_14langs	56.5	46.1	42.2	46.7	50.8	48.5
MAML_12langs	48.5	39.4	35.1	39.7	45.0	41.5
MAML_8langs	47.7	39.6	34.3	40.1	42.4	40.8

Table 2: SMATCH scores according to the number of training languages.

	fr	zh	ko	fa	hr	avg
base_DeepL	56.3	45.6	42.1	46.3	51.4	48.4
base_mBart	56.2	44.5	41.2	46.1	51.3	47.8
MAML_DeepL	56.5	46.1	42.2	46.7	50.8	48.5
MAML_mBart	55.6	45.1	40.8	46.1	48.9	47.3

Table 3: SMATCH scores according to the translation source.

languages from 8 to 12 language by 15.7%. MAML models, on the other hand, have the biggest gain when increasing the number of languages from 12 to 14 languages by 14.2%. Looking in detail per target language, however, in the MAML model, not all target languages benefit from adding more training languages. Comparing the two MAML models, trained respectively with 8 languages and 12 languages, the SMATCH score drops in Chinese and Farsi when adding four languages to the training data, whereas the baseline model shows a steady increase across target languages when adding more languages. In other words, the baseline model benefits uniformly from the inclusion of more training languages across all target languages, while the performance of the MAML model varies depending on the specific target language. In the MAML models, certain languages experience a decrease in performance despite the addition of more training languages. A caveat of this experiment is that the results may depend on the order in which the languages are added and their typological relationship to evaluation languages (we leave this investigation to future work).

Q2: How robust is the model with respect to translation quality?

To assess the impact of the translation source on our method, we employ an alternative translation model to translate our training data. Specifically, we use the mBart translation models, sourced from the Huggingface hub[8], to translate our training data into 13 languages. COMET score of the 13 translated texts is 80.7±1.4. Subsequently, we use this translation to train both the MAML and baseline models. Following this, we contrast the evaluation outcomes of these models with those trained using the DeepL translation.

Results For both the MAML and the baseline models, when using an open-source translation model mBart, the performance drops (see Table 3). In both cases, the Korean SMATCH score drops the most when using the mBart translation model. MAML model is more affected by this change. On the average score, the baseline model drops by 0.9%, whereas the MAML-model drops by 2.3%. This result shows that the meta-learning model is more sensitive to the input texts than the baseline model.

[8] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt