

Smelting Gold and Silver for Improved Multilingual AMR-to-Text Generation

Leonardo F. R. Ribeiro[†], Jonas Pfeiffer[†], Yue Zhang[‡] and Iryna Gurevych[†]

[†]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt

[‡]School of Engineering, Westlake University

ribeiro@aiphes.tu-darmstadt.de

Abstract

Recent work on multilingual AMR-to-text generation has exclusively focused on data augmentation strategies that utilize silver AMR. However, this assumes a high quality of generated AMRs, potentially limiting the transferability to the target task. In this paper, we investigate different techniques for automatically generating AMR annotations, where we aim to study which source of information yields better multilingual results. Our models trained on gold AMR with silver (machine translated) sentences outperform approaches which leverage generated silver AMR. We find that combining both complementary sources of information further improves multilingual AMR-to-text generation. Our models surpass the previous state of the art for German, Italian, Spanish, and Chinese by a large margin.¹

1 Introduction

AMR-to-text generation is the task of recovering a text with the same meaning as a given Abstract Meaning Representation (AMR) (Banarescu et al., 2013), and has recently received much research interest (Ribeiro et al., 2019; Wang et al., 2020; Mager et al., 2020; Harkous et al., 2020; Fu et al., 2021). AMR has applications to a range of NLP tasks, including summarization (Hardy and Vlachos, 2018) and spoken language understanding (Damonte et al., 2019), and has the potential power of acting as an *interlingua* that allows the generation of text in many different languages (Damonte and Cohen, 2018; Zhu et al., 2019).

While previous work has predominantly focused on monolingual English settings (Cai and Lam, 2020b; Bevilacqua et al., 2021), recent work has also studied multilinguality in meaning representations (Biloshmi et al., 2020; Sheth et al., 2021). Whereas Damonte and Cohen (2018) demonstrate

¹Our code and checkpoints are available at <https://github.com/UKPLab/m-AMR2Text>.

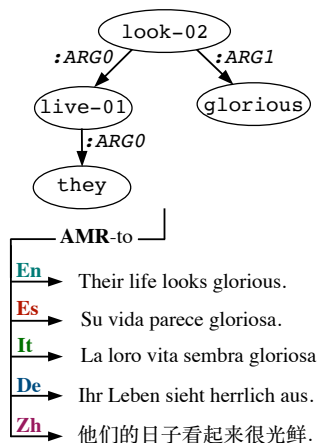


Figure 1: A generation example from English AMR to multiple different languages.

that parsers can be effectively trained to transform multilingual text into English AMR, Mille et al. (2018, 2019) and Fan and Gardent (2020) discuss the reverse task, turning meaning representations into multilingual text, as shown in Figure 1. However, gold-standard multilingual AMR training data is currently scarce, and previous work (Fan and Gardent, 2020) while discussing the feasibility of multilingual AMR-to-text generation, has investigated synthetically generated AMR as the *only source* of silver training data.

In this paper, we aim to close this gap by providing an extensive analysis of different augmentation techniques to cheaply acquire silver-standard multilingual AMR-to-text data: (1) Following Fan and Gardent (2020), we parse English sentences into silver AMRs from parallel multilingual corpora (SILVERAMR), resulting in a dataset consisting of grammatically correct sentences with noisy AMR structures. (2) We leverage machine translation (MT) and translate the English sentences from the gold AMR-to-text corpus to the respective target languages (SILVERSENT), resulting in a dataset with correct AMR structures but potentially unfaithful or non-grammatical sentences. (3) We experiment

with utilizing the AMR-to-text corpus with both gold English AMR and sentences in multi-source scenarios to enhance multilingual training.

Our contributions and the organization of this paper are the following: First, we formalize the multilingual AMR-to-text generation setting and present various cheap and efficient alternatives for collecting multilingual training data. Second, we show that our proposed training strategies greatly advance the state of the art finding that SILVERSENT considerably outperforms SILVERAMR. Third, we show that SILVERAMR has better relative performance in relatively larger sentences, whereas SILVERSENT performs better for relatively larger graphs. Overall, we find that a combination of both strategies further improves the performance, showing that they are complementary for this task.

2 Related Work

Approaches for AMR-to-text generation predominantly focus on English, and typically employ an encoder-decoder architecture, employing a linearized representation of the graph (Konstas et al., 2017; Ribeiro et al., 2020a). Recently, models based on the graph-to-text paradigm (Ribeiro et al., 2020b; Schmitt et al., 2021) improve over linearized approaches, explicitly encoding the AMR structure with a graph encoder (Song et al., 2018; Beck et al., 2018; Ribeiro et al., 2019; Guo et al., 2019; Cai and Lam, 2020b; Ribeiro et al., 2021).

Advances in multilingual AMR parsing have focused on a variety of different languages such as Brazilian Portuguese, Chinese, Czech and Spanish (Hajič et al., 2014; Xue et al., 2014; Migueles-Abraira et al., 2018; Sobrevilla Cabezudo and Pardo, 2019). In contrast, little work has focused on the reverse AMR-to-text setting (Fan and Gardent, 2020). We aim to close this gap by experimenting with different data augmentation methods for efficient multilingual AMR-to-text generation.

3 Multilingual AMR-to-Text Generation

In AMR-to-text generation, we transduce an AMR graph \mathcal{G} to a surface realization as a sequence of tokens $y = \langle y_1, \dots, y_{|y|} \rangle$. As input we use an English-centric AMR graph where the output y can be realized in different languages (see Figure 1).

3.1 Approach

We employ mT5 (Xue et al., 2021), a Transformer-based encoder-decoder architecture (Vaswani et al.,

2017), motivated by prior work (Ribeiro et al., 2020a, 2021) that leverages T5 (Raffel et al., 2019) for AMR-to-text generation.

We define $x = \text{LIN}(\mathcal{G})$, where LIN is a function that linearizes \mathcal{G} into a sequence of node and edge labels using depth-first traversal of the graph (Konstas et al., 2017). x is encoded, conditioned on which the decoder predicts y autoregressively.

Consequently, the encoder is required to learn language agnostic representations amenable to be used in a multilingual setup for the English AMR graph; the decoder attends over the encoded AMR and is required to generate text in different languages with varied word order and morphology.

To differentiate between languages, we prepend a prefix “translate AMR to <tgt_language>:” to the AMR graph representation.² We add the edge labels which are present in the AMR graphs of the LDC2017T10 training set to the encoder’s vocabulary in order to avoid considerable subtoken splitting – this allows us to encode the AMR with a compact sequence of tokens and also learn explicit representations for the AMR edge labels. Finally, this multilingual approach allows us to have more AMR data on the encoder side when increasing the number of considered languages. This could be particularly helpful when using languages with little training data.

3.2 Data

Since gold-standard *training* data for multilingual AMR-to-text generation does not exist, data augmentation methods are necessary. Given a set of gold AMR training data for English and parallel corpora between English and target languages, we thus aim to identify the best augmentations strategies to achieve multilingual generation.

As our monolingual AMR-to-text training dataset, we consider the LDC2017T10 dataset (GOLDAMR), containing English AMR graphs and sentences. We evaluate our different approaches on the multilingual LDC2020T07 test set by Damonte and Cohen (2018) consisting of gold annotations for Spanish (ES), Italian (IT), German (DE) and Chinese (ZH).³ For our multilingual parallel sentence corpus we consider data from different sources. For ES, IT and DE, we use: **Europarl-v7** (Koehn, 2005), an aligned corpus of European Union parlia-

²For example, for AMR-to-Spanish we use the prefix “translate AMR to Spanish:”.

³This dataset was constructed by professional translators based on the LDC2017T10 test set.

| | BLEU | | | | | BERTscore | | | | |
|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | ES | IT | DE | ZH | All | ES | IT | DE | ZH | All |
| MT (Fan and Gardent, 2020) | 21.6 | 19.6 | 15.7 | - | - | - | - | - | - | - |
| Multilingual model (Fan and Gardent, 2020) | 21.7 | 19.8 | 15.3 | - | - | - | - | - | - | - |
| MT | 27.6 | 24.2 | 19.4 | 23.3 | 23.6 | 87.1 | 85.7 | 83.5 | 79.9 | 84.0 |
| SILVERAMR | 23.3 | 21.2 | 16.9 | 20.1 | 20.4 | 84.5 | 83.7 | 82.0 | 76.3 | 81.6 |
| SILVERSENT | 28.3 | 24.3 | 18.9 | 22.2 | 23.4 | 87.3 | 85.7 | 83.5 | 79.6 | 84.0 |
| SILVERAMR + GOLDAMR | 28.2 | 24.9 | 19.4 | 22.9 | 23.9 | 87.6 | 85.9 | 83.9 | 79.5 | 84.2 |
| SILVERSENT + GOLDAMR | 28.5 | 24.6 | 19.2 | 22.3 | 23.7 | 87.3 | 85.8 | 83.6 | 79.6 | 84.0 |
| SILVERAMR + SILVERSENT | 30.7 | 26.4 | 20.6 | 24.2 | 25.5 | 87.8 | 86.3 | 84.1 | 80.5 | 84.7 |
| SILVERAMR + SILVERSENT + GOLDAMR | 30.4 | 26.1 | 20.5 | 23.4 | 25.1 | 88.0 | 86.3 | 84.1 | 80.1 | 84.6 |

Table 1: Results on the multilingual LDC2020T07 test set. When training on multiple seeds, the standard deviation is between 0.1 and 0.3 BLEU. The results of our models compared to the MT baseline are statistically significant.

mentary debates; **Tatoeba**,⁴ a large database of example sentences and translations; and **TED2020**,⁵ a dataset of translated subtitles of TED talks. For ZH, we use the **UM-Corpus** (Tian et al., 2014).

3.3 Creating Silver Training Data

We experiment with two augmentation techniques that generate silver-standard multilingual training data, described in what follows.

SILVERAMR. We follow Fan and Gardent (2020) and leverage the multilingual parallel corpora described in §3.2 and generate AMRs for the respective English sentences.⁶ While the multilingual sentences are of gold standard, the AMR graphs are of silver quality. Similar to Fan and Gardent (2020), for each target language we extract a parallel dataset of 1.9M sentences.

SILVERSENT. We fine-tune mT5 as a translation model for English to the respective target languages, using the same parallel sentences used in SILVERAMR. Then, we translate the English sentences of GOLDAMR into the respective target languages, resulting in a multilingual dataset that consists of gold AMRs and silver sentences. The multilingual training dataset contains 36,521 examples for each target language.

4 Experiments

We implement our models using mT5_{base} from HuggingFace (Wolf et al., 2020). We use the Adafactor optimizer (Shazeer and Stern, 2018) and employ a linearly decreasing learning rate schedule without warm-up. The hyperparameters we tune include the batch size, number of epochs and learning

rate.⁷ The models are evaluated in the multilingual LDC2020T07 test set, using BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), chrF++ (Popović, 2015) and BERTscore (Zhang et al., 2020) metrics. We compare with a MT baseline – we generate the test set with an AMR-to-English model trained with T5 (Ribeiro et al., 2021) and translate the generated English sentences to the target language using MT. For a fair comparison, our MT model is based on mT5 and trained with the same data as the other approaches.

Training Strategies. We propose different training strategies under the setting of §3.2 in order to investigate which combination leads to stronger multilingual AMR-to-text generation. Besides training models using SILVERAMR or SILVERSENT, we investigate different combinations of multi-source training also using GOLDAMR.

Main Results. Table 1 shows our main results.⁸ First, SILVERAMR substantially outperforms Fan and Gardent (2020) despite being trained on the same amount of silver AMR data. We believe this is because we utilize mT5, whereas Fan and Gardent (2020) use XLM (Conneau et al., 2020), and our parallel data may contain different domain data.

SILVERSENT considerably outperforms SILVERAMR in all metrics, despite SILVERAMR consisting of two orders of magnitude more data. We believe the reasons are twofold: Firstly, the correct semantic structure of gold AMR annotations is necessary to learn a faithful realization; Secondly, SILVERSENT provides examples of the same domain as the evaluation test set. We observe similar performance to SILVERSENT when training on both GOLDAMR and SILVERAMR, indicating that the combination of target domain data and gold AMR graphs are

⁴<https://tatoeba.org/>

⁵<https://github.com/UKPLab/sentence-transformers/tree/master/docs/datasets>

⁶The English sentences of the parallel corpus are parsed using a state-of-the-art AMR parser (Cai and Lam, 2020a).

⁷Hyperparameter details are in the appendix A.

⁸METEOR and chrF++ results can be found in Appendix Table 6.

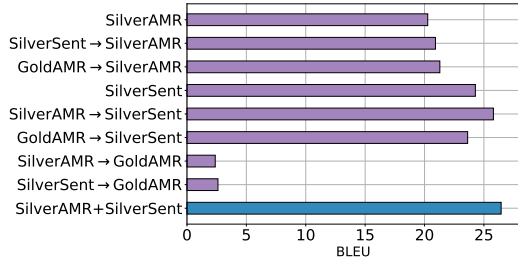


Figure 2: Order impact of sequential fine-tuning for IT.

necessary for downstream task performance. However, training on both GOLDAMR and SILVERSENT yields small gains, indicating that the respective information is adequately encoded within the silver standard dataset.

We observe similar patterns when combining the silver standard datasets. While SILVERAMR+SILVERSENT complement each other, resulting in the overall best performance, adding GOLDAMR does not yield any notably gains. These results demonstrate that *both* gold AMR structure and gold sentence information are important for training multilingual AMR-to-text models, while SILVERSENT are seemingly more important.

Effect of the Fine-tuning Order. In Figure 2 we illustrate the impact of different data source orderings when fine-tuning in a two-phase setup for IT.⁹ Firstly, we observe a decrease in performance for all sequential fine-tuning settings, compared to our proposed mixed multi-source training, which is likely due to *catastrophic forgetting*.¹⁰ Secondly, training on SILVERAMR and subsequently on SILVERSENT (or vice versa), improves performance over only using either, again demonstrating their complementarity. Thirdly, SILVERSENT continues to outperform SILVERAMR as a second task. Finally, GOLDAMR is not suitable as the second task for multilingual settings as the model predominantly generates English text.

Impact of Sentence Length and Graph Size. As silver annotations potentially lead to noisy inputs, models trained on SILVERAMR are potentially less capable of encoding the AMR semantics correctly, and models trained on SILVERSENT potentially generate fluent sentences less reliably. To analyze the advantages of the two forms of data, we measure the performance against the sentence lengths and

⁹Other languages follow similar trends and are presented in Figure 4 in the Appendix.

¹⁰The model trained on the second task forgets the first task.

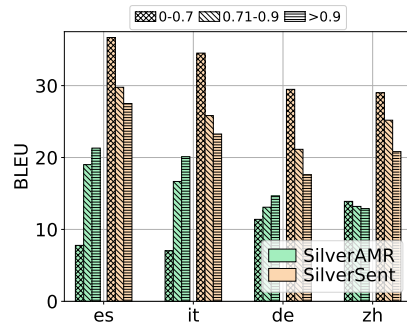


Figure 3: Impact of the sentence length and graph size ratio γ on the LDC2020T07 multilingual test set.

| | ES | IT | DE | ZH |
|------------------------|------|------|------|------|
| SILVERAMR | 19.3 | 16.5 | 11.8 | 11.9 |
| SILVERSENT | 22.3 | 17.3 | 12.7 | 11.9 |
| SILVERAMR + SILVERSENT | 23.5 | 19.2 | 15.0 | 13.0 |

Table 2: BLEU results for out of domain evaluation.

graph sizes.¹¹ We define γ to be a ratio of the sentence length, divided by the number of AMR graph nodes. In Figure 3 we plot the respective results for SILVERAMR and SILVERSENT, categorized into three bins. We find that almost all SILVERAMR’s BLEU increases for longer sentences, suggesting that training with longer gold *sentences* improves performance. In contrast, with larger *graphs*, the BLEU performance improves for SILVERSENT, indicating that large gold AMR graphs are also important. SILVERAMR and SILVERSENT present relative gains in performance on opposite ratios of sentence length and graph size, suggesting that they capture distinct aspects of the data.

Out of Domain Evaluation. To disentangle the effects of in-domain sentences and gold quality AMR graphs in SILVERSENT, we evaluate both silver data approaches on the **Weblog and WSJ** subset of the LDC2020T07 dataset; The domain of this subset is *not included* in the LDC2017T10 training set. We present the BLEU results in Table 2.¹² While we find that SILVERSENT prevails in achieving better performance — demonstrating that AMR gold structures are an important source for training multilingual AMR-to-text models — SILVERAMR and SILVERSENT perform more comparably than when evaluated on the full LDC2020T07 test set. This demonstrates that the domain transfer factor plays an important role in the strong performance of SILVERSENT. Overall, SILVERAMR+SILVERSENT outperforms both single source settings, establishing the

¹¹Sentence lengths were measured using subwords.

¹²BERTscore results can be found in Appendix Table 5.

| Model | Examples |
|----------------------|--|
| AMR | (m / multi-sentence :snt1 (w2 / wish-01 :ARG0 (i2 / i) :ARG1 (p / possible-01 :ARG1 (w3 / wipe-out-02 :ARG0 i2 :ARG1 (s / she) :source (l / live-01 :ARG0 i2)))) :snt2 (g / good-02 :ARG1 (t / thing) :degree (m2 / more :degree (m3 / much :degree (s2 / so))) :prep-without (s3 / she))) |
| SILVERAMR | Con ella, las cosas son mucho mejor. Deseo que pudiera eliminarla de mi vida. |
| SILVERSENT | Desearía que podía eliminarla de mi vida. Las cosas serían mucho mejor sin ella. |
| SILVERAMR+SILVERSENT | Desearía poder eliminarla de mi vida, las cosas serían mucho mejor sin ella. |
| Reference | Ojalá pudiera borrarla de mi vida, las cosas hubieran sido mucho mejor sin ella. |
| English Reference | I wish I could wipe her out of my life - things would be so much better without her. |

Table 3: Example of an AMR, generated texts in ES by the different models, and its ES and EN references. We indicate in **red** errors (unfaithfulness in SILVERAMR and incorrect grammar in SILVERSENT) that are not present in SILVERAMR+SILVERSENT and in the human-written reference.

complementarity of both silver sources of data.

Case Study. Table 3 shows an AMR, its reference sentences in ES and EN, and sentences generated in ES by SILVERAMR, SILVERSENT, and their combination. The incorrect verb tense is due to the lack of tense information in AMR. SILVERAMR fails in capturing the correct concept *prep-without* generating an unfaithful first sentence. This demonstrates a potential issue with approaches trained with silver AMR data where the input graph structure can be noisy, leading to a model less capable of encoding AMR semantics. On the other hand, SILVERSENT correctly generates sentences that describe the graph, while it still generates a grammatically incorrect sentence (wrongly generating *que podía* after *desearía*). This highlights a potential problem with approaches that employ silver sentence data where sentences used for the training could be ungrammatical, leading to models less capable of generating a fluent sentence. Finally, SILVERAMR+SILVERSENT produces a more accurate output than both silver approaches by generating grammatically correct and fluent sentences, correct pronouns, and mentions when control verbs and reentrancies (nodes with more than one entering edge) are involved.

5 Conclusion

The unavailability of gold training data makes multilingual AMR-to-text generation a challenging topic. We have extensively evaluated data augmentation methods by leveraging existing resources, namely a set of gold English AMR-to-text data and a corpus of multilingual parallel sentences. Our experiments have empirically validated that both sources of silver data — silver AMR with gold sentences and gold AMR with silver sentences — are complementary, and a combination of both leads to state-of-the-art performance on multilingual AMR-to-text generation tasks.

Acknowledgments

We would like to thank Gözde Gül Sahin, Ji-Ung Lee, Kevin Stowe, Kexin Wang and Nandan Thakur for their feedback on this work. Leonardo F. R. Ribeiro is supported by the German Research Foundation (DFG) as part of the Research Training Group “Adaptive Preparation of Information from Heterogeneous Sources” (AIPHES, GRK 1994/1) and as part of the DFG funded project UKP-SQuARE with the number GU 798/29-1. Jonas Pfeiffer is supported by the LOEWE initiative (Hesse, Germany) within the emergenCITY center.

References

- 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.
- Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. [Graph-to-sequence learning using gated graph neural networks](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, Melbourne, Australia. 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.
- Deng Cai and Wai Lam. 2020a. [AMR parsing via graph-sequence iterative inference](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online. Association for Computational Linguistics.
- Deng Cai and Wai Lam. 2020b. [Graph transformer for graph-to-sequence learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7464–7471.
- 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.
- Marco Damonte, Rahul Goel, and Tagyoung Chung. 2019. [Practical semantic parsing for spoken language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 16–23, Minneapolis, Minnesota. Association for Computational Linguistics.
- Michael Denkowski and Alon Lavie. 2014. [Meteor universal: Language specific translation evaluation for any target language](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. 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.
- Qiankun Fu, Linfeng Song, Wenyu Du, and Yue Zhang. 2021. [End-to-end AMR coreference resolution](#). 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 4204–4214, Online. Association for Computational Linguistics.
- Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019. [Densely connected graph convolutional networks for graph-to-sequence learning](#). *Transactions of the Association for Computational Linguistics*, 7:297–312.
- Jan Hajič, Ondřej Bojar, and Zdeňka Urešová. 2014. [Comparing Czech and English AMRs](#). In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 55–64, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.
- Hardy Hardy and Andreas Vlachos. 2018. [Guided neural language generation for abstractive summarization using Abstract Meaning Representation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773, Brussels, Belgium. Association for Computational Linguistics.
- Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. [Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Philipp Koehn. 2005. [Europarl: A Parallel Corpus for Statistical Machine Translation](#). In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

- Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. [Neural AMR: Sequence-to-sequence models for parsing and generation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.
- Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md. Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. [Gpt-too: A language-model-first approach for amr-to-text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1846–1852.
- Noelia Migueles-Abraira, Rodrigo Agerri, and Arantza Diaz de Ilarraza. 2018. [Annotating Abstract Meaning Representations for Spanish](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. [The first multilingual surface realisation shared task \(SR'18\): Overview and evaluation results](#). In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.
- Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. [The second multilingual surface realisation shared task \(SR'19\): Overview and evaluation results](#). In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 1–17, Hong Kong, China. Association for Computational Linguistics.
- 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.
- 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.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *arXiv e-prints*.
- Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. 2019. [Enhancing AMR-to-text generation with dual graph representations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3183–3194, Hong Kong, China. Association for Computational Linguistics.
- Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020a. [Investigating pretrained language models for graph-to-text generation](#). *arXiv e-prints*.
- Leonardo F. R. Ribeiro, Yue Zhang, Claire Gardent, and Iryna Gurevych. 2020b. [Modeling global and local node contexts for text generation from knowledge graphs](#). *Transactions of the Association for Computational Linguistics*, 8:589–604.
- Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021. [Structural adapters in pretrained language models for amr-to-text generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Punta Cana, November 7-11, 2021*.
- Martin Schmitt, Leonardo F. R. Ribeiro, Philipp Dufer, Iryna Gurevych, and Hinrich Schütze. 2021. [Modeling graph structure via relative position for text generation from knowledge graphs](#). In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 10–21, Mexico City, Mexico. Association for Computational Linguistics.
- Noam Shazeer and Mitchell Stern. 2018. [Adafactor: Adaptive learning rates with sublinear memory cost](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.
- 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.
- Marco Antonio Sobrevilla Cabezedo and Thiago Pardo. 2019. [Towards a general Abstract Meaning Representation corpus for Brazilian Portuguese](#). In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 236–244, Florence, Italy. Association for Computational Linguistics.
- Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. [A graph-to-sequence model for AMR-to-text generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. [Um-corpus: A large english-chinese parallel corpus for statistical machine translation](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 1837–1842.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Tianming Wang, Xiaojun Wan, and Hanqi Jin. 2020. [AMR-to-text generation with graph transformer](#). *Transactions of the Association for Computational Linguistics*, 8:19–33.

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. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. 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.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Uřešová, and Xiuhong Zhang. 2014. [Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1765–1772, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Huaiyu Zhu, Yunyao Li, and Laura Chiticariu. 2019. [Towards universal semantic representation](#). In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 177–181,

Florence, Italy. Association for Computational Linguistics.

Appendices

A Details of Models and Hyperparameters

The experiments were executed using the version 4.4.0 of the *transformers* library by Hugging Face (Wolf et al., 2020). Table 4 shows the hyperparameters used to train our models. BLEU is used for model selection using translated sentences of the LDC2017T10 development set. We train until the results on the development set BLEU have not improved for 6 epochs.

| | |
|-------------------|-------|
| learning rate | 1e-04 |
| batch size | 8 |
| beam search size | 6 |
| max source length | 350 |
| max target length | 200 |

Table 4: Hyperparameter settings for our methods.

B Main Results: Additional Metrics

In Table 6 we present additional results on the multilingual LDC2020T07 test set using METEOR (Denkowski and Lavie, 2014), chrF++ (Popović, 2015) metrics.

C Results: Out of Domain Evaluation

In Table 5 we show BERTscore (Zhang et al., 2020) results for out of domain evaluation on the **Weblog and WSJ** subset of the LDC2020T07 dataset.

| | ES | IT | DE | ZH |
|------------------------|------|------|------|------|
| SILVERAMR | 83.3 | 81.2 | 79.8 | 73.6 |
| SILVERSENT | 84.6 | 83.0 | 80.4 | 73.0 |
| SILVERAMR + SILVERSENT | 84.6 | 83.2 | 81.2 | 74.1 |

Table 5: BERT scores for out of domain evaluation.

D Results: Sequential Fine-tuning

In Figure 4 we present the impact of sequential fine-tuning strategies in the LDC2020T07 test set for ES, DE and ZH.

| | METEOR | | | | | chrF++ | | | | |
|----------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | ES | IT | DE | ZH | All | ES | IT | DE | ZH | All |
| MT | 29.9 | 27.2 | 23.2 | 25.7 | 26.5 | 54.8 | 52.0 | 47.3 | 22.3 | 44.1 |
| SILVERAMR | 28.3 | 26.0 | 22.7 | 23.3 | 25.0 | 51.3 | 49.6 | 45.9 | 19.5 | 41.5 |
| SILVERSENT | 30.6 | 27.3 | 23.0 | 24.9 | 26.4 | 55.6 | 52.2 | 47.2 | 21.7 | 44.1 |
| SILVERAMR + GOLDAMR | 29.8 | 26.9 | 23.6 | 25.2 | 26.3 | 55.9 | 51.7 | 47.5 | 22.3 | 44.3 |
| SILVERSENT + GOLDAMR | 30.4 | 27.5 | 23.3 | 24.9 | 26.5 | 55.3 | 52.3 | 47.3 | 21.8 | 44.1 |
| SILVERAMR + SILVERSENT | 31.9 | 28.7 | 24.4 | 26.4 | 27.8 | 57.2 | 54.0 | 49.4 | 23.0 | 45.9 |
| SILVERAMR + SILVERSENT + GOLDAMR | 31.7 | 28.6 | 24.2 | 25.7 | 27.5 | 57.2 | 53.6 | 48.6 | 22.5 | 45.4 |

Table 6: METEOR and chrF++ results on the multilingual LDC2020T07 test set.

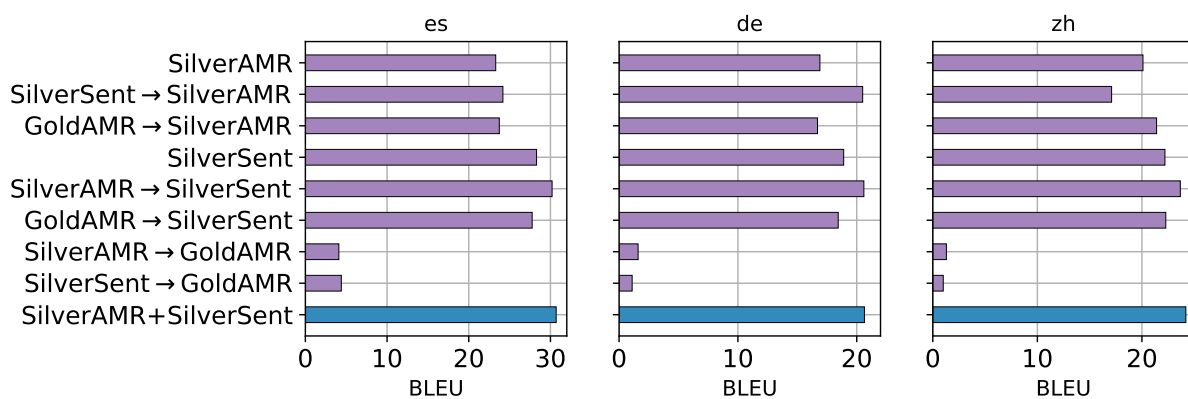


Figure 4: Order impact of sequential fine-tuning in the LDC2020T07 test set for ES, DE and ZH.