Testing AMR Universality: Translating Reduced AMR Structures of The Little Prince in English, Chinese, and Persian

Daniel Stuhlinger
daniel.stuhlinger
@student.uni-tuebingen.de

Darja Jepifanova
darja.jepifanova
@student.uni-tuebingen.de

Abstract

Abstract Meaning Representation (AMR) is a semantic framework designed to capture sentence meaning through a rooted, directed acyclic graph of concepts and relations. Despite its focus on language-agnostic meaning, AMR was originally designed for English, raising concerns about how well it generalizes across languages. This study investigates cross-linguistic structural divergences in AMR by comparing reduced AMR structures of English, Chinese, and Persian sentences, using translations from The Little Prince. By stripping AMRs of language-specific labels and training bidirectional LSTM encoder-decoder models, we assess how structural differences impact translation performance. BLEU scores suggest that English and Persian AMRs align more closely than English and Chinese, though results are complicated by the limitations of using BLEU for AMR data and the small dataset size. Future work is required to explore alternative metrics, such as XSmatch and XS2match, which may offer a more robust assessment of cross-linguistic AMR compatibility.

1 Introduction

One of the most used and well-known semantic representation models is Abstract Meaning Representation (AMR). It is a semantic representation framework that captures the meaning of a sentence or phrase as a rooted, directed acyclic graph, where nodes represent concepts and labeled edges denote the relationships between these concepts (Banarescu et al., 2013). AMR abstracts away from syntactic aspects such as word order, morphological variation, and function words, focusing instead on semantic concepts and relations between them.

This abstraction makes AMR a versatile tool for representing meaning in a way that can be applied to various NLP tasks. In machine translation, AMR's abstract representation aids in transferring meaning across languages, potentially improving translation quality (Li and Flanigan, 2022), (Song et al., 2019). In summarization, AMR helps generate concise and coherent summaries by focusing on the core meaning of texts (Kouris et al., 2022), (Lima Inácio and Pardo, 2021). It is also used in detecting toxic content by analyzing the underlying meaning of the text (Elbasani and Kim, 2022).

Though AMR was originally designed for annotating English sentences and not intended as an interlingua,[1] it has since been adapted to several other languages. In these cross-lingual adaptations, it is often assumed that AMRs for non-English languages should mirror the structure of English AMRs. This assumption is problematic because it leads to the evaluation of cross-lingual AMR parsing against English gold-standard data. Consequently, AMRs for other languages are frequently adjusted to fit an English-like structure, which may not accurately reflect the syntactic or semantic characteristics of those languages (Wein and Schneider, 2022). This bias undermines our ability to assess how effectively AMR can abstract from the particularities of individual languages.

In practice, comparing gold-standard English AMRs with those from other languages reveals differences in both concepts and structure due to the syntactic and semantic influences of the source language (Damonte and Cohen, 2019; Wein and Schneider, 2021; Blloshmi et al., 2020). A naive assumption might be that AMRs for parallel sentences should be structurally identical across languages, based on the idea that AMR encodes meaning and that translation preserves meaning perfectly. However, research has shown that this is actually not the case. Even when lexical items are aligned to a common language, such as English, the source language still impacts the structure of the AMR.

Recently (Wein and Schneider, 2024) attempted to quantify this impact by manually translating Chi-

[1] Interlingua – an artificial language proposed for use as an auxiliary international language

nese concepts into their English equivalents. Their method involved replacing Chinese labels with English ones and checking for corresponding synonyms of concepts in the English gold-standard AMRs. If a suitable synonym was found, it was used; otherwise, the manually translated term was retained. Despite these efforts to minimize lexical divergence, the resulting Smatch scores between the gold English AMRs and the translated Chinese AMRs were consistently below 50%. This suggests that significant differences remain – likely due to deeper structural divergences between languages. However, it is crucial to note that these divergences are not solely due to structural or syntactic differences but also arise from translation choices.

Xue et al., 2014 also note that differences in lexicalization between languages introduce mismatches that cannot easily be resolved without moving to a higher level of abstraction than AMR currently provides. It raises concerns that some of these differences might stem from translation effects, meaning that the act of translating itself introduces divergences that obscure the true cross-lingual compatibility of AMRs.

To address these concerns, in this work we investigate how well AMR captures cross-linguistic semantics by studying the translation between reduced AMR structures of English, Chinese, and Persian sentences. To quantify the impact of the source language on the AMR, we try to eliminate the impact of lexical divergence and focus solely on structural divergences. To achieve this, we reduce the AMRs by retaining only the bracketing structure and common labels across parallel English, Chinese, and Persian sentences of "The Little Prince". This approach ensures that any quantifiable differences between the AMRs of these languages are attributed to structural elements rather than lexical variations.

We then train identical LSTM bidirectional encoder-decoder models for the English-Chinese and English-Persian AMR pairs, maintaining the same architecture and hyperparameters. By evaluating and comparing the performance of these models, we aim to assess how structural differences between AMRs in different languages impact translation accuracy. We hypothesize that greater structural divergence will result in lower accuracy of the models. There is excessive previous work about neural machine translation working better for more similar, related languages (Dabre et al., 2020). This even leads to a point, where machine translation models are trained on a high resource language and then deployed for translations to low resource languages, optionally being fine tuned on a small amount of training data for this language (Lakew et al., 2020).

2 Data

In selecting the data for our experiments, we chose to use gold AMR annotations rather than parser-generated AMRs to minimize errors stemming from parsing inaccuracies. Prior research has shown that data points classified as having no meaning difference but displaying extremely low F1 scores are often affected by parser errors, rather than genuine structural divergences (Wein and Schneider, 2024). By using gold-standard AMRs, we ensure that any observed differences are more likely to reflect true structural divergences between languages rather than effects of faulty parsing.

For our analysis, we use the AMR corpus derived from "The Little Prince", a widely used text in AMR research. The corpus includes translations into English, Chinese, Spanish and Vietnamese. This text was chosen due to its open-source availability in multiple languages. It is important to note that "The Little Prince" was originally written in French, and as such, the datasets reflect translations that may exhibit features of translations or differences introduced by French serving as a pivot language (Koppel and Ordan, 2011). Additionally, the story's literary style can lead to more pronounced translation variations, which may introduce greater variance in the translations and potentially reduce comparability between sentences.

In this study, we focus on English, Chinese and Persian parallel corpora. The language pair English-Chinese is typologically more diverse, representing different language families and structural characteristics. English and Persian are members of the Indo-European language family. While English is a Germanic language with analytic features and a relatively fixed word order, Persian displays agglutinative features and a more flexible word order. Chinese on the other hand is a Sino-Tibetan language and exhibits isolating traits and a topic-prominent structure. This typological diversity provides a broad spectrum of linguistic characteristics, allowing us to investigate how AMR structures handle different language types. By examining these

languages, we aim to explore whether AMR structures can generalize effectively across such diverse linguistic contexts. For the analysis, we utilize English-Chinese parallel AMR data from (Li et al., 2016) and Persian AMR data from (Takhshid et al., 2022).

3 Preprocessing

We began by preprocessing and aligning the AMR data for all three languages. The initial step involved cleaning the files by extracting only IDs, sentences, and AMR graphs. While the English and Chinese files were perfectly aligned due to their shared source, the Persian data required additional adjustments due to swapped or duplicated sentences. These discrepancies were resolved by manually aligning the Persian sentences with their English and Chinese counterparts.

After aligning the datasets, we cleaned the AMR graphs by removing language-specific labels and retaining only their structural elements. By stripping away non-common labels and preserving key elements like parentheses and colons, we ensured that only the shared core structures of the AMRs remained. This approach minimizes language-specific bias in the annotations, as discussed in Section 1. With this process we retrieved 1561 sentence pairs and corresponding AMR representations. 90% was used for training and 10% was left for evaluation.

The obtained data was tokenized by treating each character as a separate token, except labels which were kept as a single token. Each AMR representation corresponds to a natural language sentence. For the source language (English) we inserted special <BOS> tokens at the beginning and <EOS> tokens at the end. For the target AMRs only the <EOS> token was appended.

4 Model

To translate from source to target AMRs we utilized a LSTM (Sundermeyer et al., 2012) with a bidirectional encoder and decoder component. This follows the idea that the encoder creates a hidden representation of the source language's AMR, which can then in turn be decoded into the AMR of the target language (Sutskever et al., 2014). The code can be found in our GitHub repository [2]. We found that applying teacher-forcing yields the best results. The following hyperparameters were applied for both models:

Parameter	Value
max. number of epochs	100
hidden size	256
learning rate	0.00001
teacher forcing ratio	1.0
number of layers	2
dropout	0.2

These parameters were chosen based on the best performance for English-Chinse translation. We continued to train two models given the described architecture. One to translate from English to Chinese and another one from English to Persian AMRs.

5 Results

The evaluation metric of our choice was the BLEU score (Papineni et al., 2002). It is based on the precision of overlapping n-grams. We tested the our models with different numbers of n-grams between predicted and ground truth AMRs.

N-gram	English-Chinese	English-Farsi
Unigrams	0.60	0.78
Bigrams	0.83	0.82
Trigrams	0.80	0.80

Table 1: BLEU score for the AMR pairs based on different n-grams

Looking only at unigrams these results suggest that there is a significant difference in performance between English-Chinese and English-Persian. However, considering bigrams and trigrams as well, the models perform similar on both language pairs.

6 Limitations

The unigram results support our hypothesis that English AMRs are structurally more similar to Persian than they are to Chinese. However, the BLEU score was developed to assess the quality of natural language language translations. Since we are dealing with AMRs instead of natural language and our vocabulary consists of only 112 tokens, the choice of overlapping unigrams as a measure of similarity might not be very meaningful.

Additionally, it could be the case that this specific model architecture just happens to work better

[2] https://github.com/ydarja/MR_Project

for English-Chinese than for English-Persian. To arrive at a reliable conclusion further experiments across different model architectures would have to be conducted.

Obviously, training data consisting of only around 1400 sentence pairs is not enough to draw a valid conclusion. Further research would be needed. Similarly, for this project we had limited computational resources available. A more complex architecture combined with more training instances might lead to more insightful results.

7 Future Work

Given the limitations as described in the previous section, it would be beneficial to address these issues by future studies.

Wein and Schneider, 2022 introduce more suitable metrics to evaluate the similarity between cross-linguistic AMR, which are kept mostly intact in contrast to our reduced representations. However, we should note that these approaches lead away from our original idea to assess similarities based on the performance on LSTM translations.

- Xsmatch
  This metric is based on the Smatch score (Cai and Knight, 2013) but translates words in the AMR instance and attribute triples. The translation process is automated using machine translation. Word senses are removed for better comparison.

- XSemBleu
  This approach is based on SemBleu (Song and Gildea, 2019) but involves a translation step similar to XSmatch. Again, the translation process is automated.

- XS2match
  S2match (Opitz et al., 2020) utilizes word embeddings for concepts. XS2match adapts this idea to a cross-linguistical context by creating the token embedding with LaBSE (Feng et al., 2022), which is language agnostic.

  It might as well be interesting to test how the models would perform on languages that are even more similar like English and German.

8 Conclusion

While the performance based on the BLEU score of unigrams might suggest that there is some evidence for our hypothesis, we don't feel confident to make a strong claim about its validity. Even with more resources such as more training data and computaional capacity, other methods might generally work better to assess the simiarity between cross-liguistic AMR.

References

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider (2013, August). Abstract Meaning Representation for sembanking. In A. Pareja-Lora, M. Liakata, and S. Dipper (Eds.), *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, pp. 178–186. Association for Computational Linguistics.

Blloshmi, R., R. Tripodi, and R. Navigli (2020, November). XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In B. Webber, T. Cohn, Y. He, and Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 2487–2500. Association for Computational Linguistics.

Cai, S. and K. Knight (2013, August). Smatch: an evaluation metric for semantic feature structures. In H. Schuetze, P. Fung, and M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, pp. 748–752. Association for Computational Linguistics.

Dabre, R., C. Chu, and A. Kunchukuttan (2020). A comprehensive survey of multilingual neural machine translation. *CoRR abs/2001.01115*.

Damonte, M. and S. B. Cohen (2019, June). Structural neural encoders for AMR-to-text generation. In J. Burstein, C. Doran, and T. Solorio (Eds.), *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)*, Minneapolis, Minnesota, pp. 3649–3658. Association for Computational Linguistics.

Elbasani, E. and J.-D. Kim (2022, Feb.). Amr-cnn: Abstract meaning representation with convolution neural network for toxic content detection. *Journal of Web Engineering 21*(03), 677–692.

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang (2022, May). Language-agnostic BERT sentence embedding. In S. Muresan, P. Nakov, and A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 878–891. Association for Computational Linguistics.

Koppel, M. and N. Ordan (2011, 01). Translationese and its dialects. pp. 1318–1326.

Kouris, P., G. Alexandridis, and A. Stafylopatis (2022, 08). Text summarization based on semantic graphs: An abstract meaning representation graph-to-text deep learning approach.

Lakew, S. M., M. Negri, and M. Turchi (2020). Low resource neural machine translation: A benchmark for five african languages.

Li, B., Y. Wen, W. Qu, L. Bu, and N. Xue (2016, August). Annotating the little prince with Chinese AMRs. In A. Friedrich and K. Tomanek (Eds.), *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, Berlin, Germany, pp. 7–15. Association for Computational Linguistics.

Li, C. and J. Flanigan (2022, July). Improving neural machine translation with the Abstract Meaning Representation by combining graph and sequence transformers. In L. Wu, B. Liu, R. Mihalcea, J. Pei, Y. Zhang, and Y. Li (Eds.), *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, Seattle, Washington, pp. 12–21. Association for Computational Linguistics.

Lima Inácio, M. and T. Pardo (2021, September). Semantic-based opinion summarization. In R. Mitkov and G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online, pp. 619–628. INCOMA Ltd.

Opitz, J., L. Parcalabescu, and A. Frank (2020, 09). AMR Similarity Metrics from Principles. *Transactions of the Association for Computational Linguistics 8*, 522–538.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics.

Song, L. and D. Gildea (2019, July). SemBleu: A robust metric for AMR parsing evaluation. In A. Korhonen, D. Traum, and L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 4547–4552. Association for Computational Linguistics.

Song, L., D. Gildea, Y. Zhang, Z. Wang, and J. Su (2019). Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics 7*, 19–31.

Sundermeyer, M., R. Schlüter, and H. Ney (2012). Lstm neural networks for language modeling. In *Interspeech*, Volume 2012, pp. 194–197.

Sutskever, I., O. Vinyals, and Q. V. Le (2014). Sequence to sequence learning with neural networks.

Takhshid, R., R. Shojaei, Z. Azin, and M. Bahrani (2022). Persian abstract meaning representation.

Wein, S. and N. Schneider (2021, November). Classifying divergences in cross-lingual AMR pairs. In C. Bonial and N. Xue (Eds.), *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, Punta Cana, Dominican Republic, pp. 56–65. Association for Computational Linguistics.

Wein, S. and N. Schneider (2022, October). Accounting for language effect in the evaluation of cross-lingual AMR parsers. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, pp. 3824–3834. International Committee on Computational Linguistics.

Wein, S. and N. Schneider (2024, 06). Assessing the Cross-linguistic Utility of Abstract Meaning Representation. *Computational Linguistics 50*(2), 419–473.

Xue, N., O. Bojar, J. Hajič, M. Palmer, Z. Urešová, and X. Zhang (2014, May). Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, pp. 1765–1772. European Language Resources Association (ELRA).