

Cross-lingual AMR Aligner: Paying Attention to Cross-Attention

Abelardo Carlos Martínez Lorenzo^{1,2*} Pere-Lluís Huguet Cabot^{1,2*}

Roberto Navigli²

¹ Babelscape, Italy

² Sapienza NLP Group, Sapienza University of Rome

{martinez, huguetcabot}@babelscape.com

navigli@diag.uniroma1.it

Abstract

This paper introduces a novel aligner for Abstract Meaning Representation (AMR) graphs that can scale cross-lingually, and is thus capable of aligning units and spans in sentences of different languages. Our approach leverages modern Transformer-based parsers, which inherently encode alignment information in their cross-attention weights, allowing us to extract this information during parsing. This eliminates the need for English-specific rules or the Expectation Maximization (EM) algorithm that have been used in previous approaches. In addition, we propose a guided supervised method using alignment to further enhance the performance of our aligner. We achieve state-of-the-art results in the benchmarks for AMR alignment and demonstrate our aligner’s ability to obtain them across multiple languages. Our code will be available at github.com/Babelscape/AMR-alignment.

1 Introduction

At the core of Natural Language Understanding lies the task of Semantic Parsing, aimed at translating natural language text into machine-interpretable representations. One of the most popular semantic formalisms is the Abstract Meaning Representation (Banarescu et al., 2013, AMR), which embeds the semantics of a sentence in a directed acyclic graph, where concepts are represented by nodes, such as *time*, semantic relations between concepts by edges, such as *:beneficiary*, and the co-references by reentrant nodes, such as *r* representing *rose*. In cross-lingual AMR, the English AMR graph represents the sentence in different languages (see Figure 1). To date, AMR has been widely used in Machine Translation (Song et al., 2019), Question Answering (Lim et al., 2020; Kapanipathi et al., 2021), Human-Robot Interaction (Bonial et al., 2020), Text Summarization (Hardy and Vlachos, 2018;

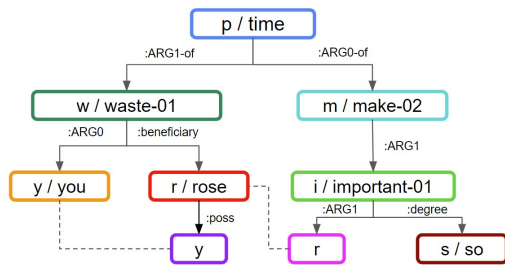
Liao et al., 2018) and Information Extraction (Rao et al., 2017), among other areas.

The alignment between spans in text and semantic units in AMR graphs is an essential requirement for a variety of purposes, including training AMR parsers (Zhou et al., 2021), cross-lingual AMR parsing (Blloshmi et al., 2020), downstream task application (Song et al., 2019), or the creation of new semantic parsing formalisms (Navigli et al., 2022; Martínez Lorenzo et al., 2022). Despite the emergence of various alignment generation approaches, such as rule-based methods (Liu et al., 2018) and statistical strategies utilizing Expectation Maximization (EM) (Pourdamghani et al., 2014; Blodgett and Schneider, 2021), these methods rely heavily on English-specific rules, making them incompatible with cross-lingual alignment. Furthermore, even though several attempts extend the alignment to non-English sentences and graphs (Damonte and Cohen, 2018; Uhrig et al., 2021), these efforts are inherently monolingual and therefore lack the connection to the richer AMR graph bank available in English, which can be exploited as a source of interlingual representations.

On the other hand, current state-of-the-art AMR parsers are auto-regressive neural models (Bevilacqua et al., 2021; Bai et al., 2022) that do not generate alignment when parsing the sentence to produce the graph. Therefore, to obtain both, one needs to i) predict the graph and then ii) generate the alignment using an aligner system that is based on language-specific rules.

This paper presents the first AMR aligner that can scale cross-lingually by leveraging the implicit information acquired in Transformer-based parsers (Bai et al., 2022). We propose an approach for extracting alignment information from cross-attention, and a guided supervised method to enhance the performance of our aligner. We eliminate the need for language-specific rules and enable simultaneous generation of the AMR graph and

* Equal contributions.



It is the **time** **you** have wasted for **your** **rose** that **makes** your **rose** **so** **important**

El **tiempo** que **perdiste** por **tu** **rosa** **hace** que tu **rosa** sea **tan** **importante**

È il **tempo** che **tu** hai **perduto** per la **tua** **rosa** che **ha fatto** la tua **rosa** **così** **importante**

Die **Zeit**, die **du** für **deine** **Rose** **verloren** hast, sie **macht** deine **Rose** **so** **wichtig**.

正因为**你**为**你的****玫瑰**花费了**时间**，这才**使****你的****玫瑰**变得**如此****重要**

Figure 1: AMR graph (left) and its corresponding sentences in several languages (right). Colors represent alignment.

alignment. Our approach is efficient and robust, and is suitable for cross-lingual alignment of AMR graphs.

Our main contributions are: (i) we explore how Transformer-based AMR parsers preserve implicit alignment knowledge and how we can extract it; (ii) we propose a supervised method using cross-attention to enhance the performance of our aligner, (iii) we achieve state-of-the-art results along different alignment standards and demonstrate the effectiveness of our aligner across languages.

2 Related Work

AMR alignment Since the appearance of AMR as a Semantic Parsing formalism, several aligner systems have surfaced that provide a link between the sentence and graph units. JAMR (Flanigan et al., 2014) is a widely used aligner system that employs an ordered list of 14 criteria, including exact and fuzzy matching, to align spans to subgraphs. However, this approach has limitations as it is unable to resolve ambiguities or learn novel alignment patterns. TAMR (Liu et al., 2018) extends JAMR by incorporating an oracle parser that selects the alignment corresponding to the highest-scored candidate AMR graph. ISI (Pourdamghani et al., 2014) aligner utilizes an EM algorithm to establish alignment between words and graphs’ semantic units. First, the graph is linearized, and then the EM algorithm is employed with a symmetrized scoring function to establish alignments. This method leads to more diversity in terms of alignment patterns, but fails to align easy-to-recognize patterns that could be aligned using rules. LEAMR (Blodgett and Schneider, 2021) is another aligner system that combines rules and EM. This approach aligns all the subgraph structures to any span in the sentence. However, it is based on language-specific rules, making it unsuitable for cross-lingual settings. Moreover, despite several attempts to extend the alignment to non-English languages An-

chiêta and Pardo (2020); Oral and Eryigit (2022), these efforts are still monolingual since they rely on language-specific strategies. Consequently, in this paper we present an approach that fills this gap.

Cross-attention Most state-of-the-art systems for AMR parsing are based on Encoder-Decoder Transformers, specifically on BART (Lewis et al., 2020). These models consist of two stacks of Transformer layers, which utilize self- and cross-attention as their backbone. The popularity of Transformer models has led to increased interest in understanding how attention encodes information in text and relates to human intuition (Vashishth et al., 2019) and definitions of explainability (Bastings and Filippova, 2020; Bibal et al., 2022). Research has been conducted on how attention operates, relates to preconceived ideas, aggregates information, and explains model behavior for tasks such as natural language inference (Stacey et al., 2021), Translation (Yin et al., 2021; Zhang and Feng, 2021; Chen et al., 2021), Summarization (Xu et al., 2020; Manakul and Gales, 2021) or Sentiment Analysis (Wu et al., 2020). Furthermore, there have been attempts to guide attention to improve interpretability or performance in downstream tasks (Deshpande and Narasimhan, 2020; Sood et al., 2020). However, to the best of our knowledge, there has been no prior study on attention for AMR parsing. This paper fills this gap by investigating the role of attention in AMR parsing.

3 Method

Originally described by Vaswani et al. (2017) as “multi-head attention over the output of the Encoder”, and referred to as cross-attention in Lewis et al. (2020), it enables the Decoder to attend to the output of the Encoder stack, conditioning the hidden states of the autoregressive component on the input text. Self-attention and cross-attention

modules are defined as:

$$\text{Attention}(Q, K, V) = \text{att}(Q, K)V$$

$$\text{att}(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

$$\text{CrossAtt}(Q, K, V) =$$

$$\text{Concat}(\text{head}_1, \dots, \text{head}_H)W^O$$

$$\text{head}_h = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V)$$

where $K, V = E^\ell \in \mathbb{R}^{n_e \times d_k H}$ and $Q = D^\ell \in \mathbb{R}^{n_d \times d_k H}$ are the Encoder and Decoder hidden states at layer ℓ , n_e and n_d are the input and output sequence lengths, H is the number of heads, W_h^Q, W_h^K and $W_h^V \in \mathbb{R}^{d_k H \times d_k}$ are learned weights that project the hidden states to the appropriate dimensions, d_k , for each head and $W^O \in \mathbb{R}^{d_k H \times d_k H}$ is a final learned linear projection. Therefore in each head h and layer ℓ we define the attention weights as $\text{att}_h^\ell = \text{att}(D^\ell W_h^Q, E^\ell W_h^K) \in \mathbb{R}^{n_d \times n_e}$.

3.1 Unguided Cross-Attention

We argue that there is an intuitive connection between cross-attention and alignments. Under the assumption the Decoder will attend to the parts in the input that are more relevant to predicting the next token, we infer that, when decoding the tokens for a certain node in the graph, attention should focus on related tokens in the input, and therefore the words that align to that node. We will use the cross-attention matrices (att_h^ℓ) to compute an alignment between the input and the output.

3.2 Guided Cross-Attention

We also aim to explore whether cross-attention can be guided by the alignment between the words of the sentence and the nodes of the graph. To this end, we construct a sparse matrix $\text{align} \in \mathbb{R}^{n_d \times n_e}$ from the automatically-generated alignments:

$$\text{align}(i, j) = \begin{cases} 1 & \text{if } x_i \sim y_j \\ 0 & \text{if } x_i \not\sim y_j \end{cases}$$

where \sim indicates alignment between subword token x_i and graph token y_j .

However, even though there are sparse versions of attention (Martins and Astudillo, 2016), these did not produce successful alignments in our experiments. Hence we choose to alleviate the constraint

of imposing sparsity by employing the scalar mixing approach introduced in ELMo (Peters et al., 2018). We learn a weighted mix of each head and obtain a single attention matrix:

$$\text{att}^\ell = \gamma \sum_{h=0}^{H-1} s_h^\ell \text{att}_h^\ell \in \mathbb{R}^{n_d \times n_e} \quad (1)$$

where $\mathbf{s} = \text{softmax}(\mathbf{a})$ with scalar learnable parameters γ, a_0, \dots, a_H .

The model has the flexibility to learn how to distribute weights such that certain heads give sparser attention similar to alignment, while others can encode additional information that is not dependent on alignment. In our experiments, we use the implementation of Bevilacqua et al. (2021, SPRING) to train our parser but add an extra cross-entropy loss signal:

$$\mathcal{L} = - \sum_{j=1}^{n_d} \log p_{BART}(y_j | y_{<j}, x) - \sum_{\substack{j=1 \\ \sum_i \text{align}(i,j) > 0}}^{n_d} \sum_{i=1}^{n_e} \log \left(\frac{e^{\text{att}^\ell(i,j)} \text{align}(i,j)}{\sum_{k=1}^{n_d} e^{\text{att}^\ell(i,k)} \sum_{k=1}^{n_e} \text{align}(k,j)} \right)$$

3.3 Saliency Methods

A theoretical alternative to our reasoning about cross-attention is the use of input saliency methods. These methods assign higher importance to the input tokens that correspond to a particular node in the graph or were more important in their prediction during decoding. To obtain these importance weights, we employ Captum (Kokhlikyan et al., 2020), an open-source library for model interpretability and understanding, which provides a variety of saliency methods, including gradient-based methods such as Integrated Gradients (IG), Saliency (Simonyan et al., 2014), and Input X Gradient (IxG), backpropagation-based methods such as Deeplift (Shrikumar et al., 2017) and Guided Backpropagation (GB) (Springenberg et al., 2015), and finally occlusion-based methods (Zeiler and Fergus, 2014).

We obtain a weight matrix $\text{sal} \in \mathbb{R}^{n_d \times n_e}$ with the same size as the cross-attention matrix and use it to extract alignments in the same fashion as the unguided cross-attention method. This approach allows us to explore the input tokens that have a greater impact on the decoding process and can aid in understanding the reasoning behind the alignments made by the model.

3.4 Alignment Extraction

Our algorithm¹ to extract and align the input-output spans is divided into six steps:

1. **Alignment score matrix:** we create a matrix $M \in \mathbb{R}^{n_d \times n_e}$, where n_e is the number of tokens in the sentence and n_d is the number of tokens in the linearized graph, using the cross-attention weights (att_h^ℓ or att^ℓ) as described in Section 3.
2. **Span segmentation:** For each sentence word, we sum the scores of tokens that belong to the same word column-wise in M . Then, for LEAMR alignments (see Section 4.2), the sentence tokens are grouped into spans using their span segmentation (see Appendix A).
3. **Graph segmentation:** We sum the score of tokens that belong to the same graph’s semantic unit row-wise in M .
4. **Sentence graph tokens map:** We iterate over all the graph’s semantic units and map them to the sentence span with highest score in M .
5. **Special graph structures:** We revise the mapping by identifying subgraphs that represent literal or matching spans – e.g., named entities, dates, specific predicates, etc. – and align them accordingly.
6. **Alignment formatting:** We extract the final alignments to the appropriate format using the resulting mapping relating graph’s semantic units to sentence spans.

4 Experimental Setup

4.1 Graph inventory

AMR 3.0 (LDC2020T02) consists of 59,255 sentence-graph pairs that are manually annotated. However, it lacks alignment information between the nodes in the graphs and the concepts in the sentences. We use the train split for the guided approach and use the respective validation and test splits from the alignment systems. Additionally, to evaluate cross-lingual performance, we use the gold German, Italian, and Spanish sentences of “AMR 2.0 – Four Translation” (LDC2020T07) which are human parallel translations of the test set

in AMR 2.0², paired with their English graphs from the AMR 3.0 test set. Despite this, as the graph inventory does not contain alignment information, it becomes necessary to access other repositories in order to obtain the alignment.

4.2 Alignment Standards

We propose an approach that is agnostic to different alignment standards and we evaluate it on two standards that are commonly used: ISI and LEAMR.

ISI The ISI standard, as described in (Pourdamghani et al., 2014), aligns single spans in the sentence to graphs’ semantic units (nodes or relations), and aligns relations and reentrant nodes when they appear explicitly in the sentence. The alignments are split into two sets of 200 annotations each, which we use as validation and test sets, updated to the AMR 3.0 formalism. For the cross-lingual alignment setup, we project English ISI graph-sentence alignments to the sentences in other languages, using the machine translation aligner (Dou and Neubig, 2021). This involves connecting the nodes in the graph to the spans in non-English sentences using the projected machine translation alignments between the English spans and the corresponding non-English sentence spans. By leveraging this, we are able to generate a silver alignment for cross-lingual AMR, which enables us to validate the model’s performance in a cross-lingual setup and determine its scalability across-languages.

LEAMR The LEAMR standard differentiates among four different types of alignment: i) Subgraph Alignments, where all the subgraphs that explicitly appear in the sentence are aligned to a list of consecutive spans, ii) Duplicate Subgraph, where all the subgraphs that represent omitted repeated concepts in the sentence are aligned, iii) Relation Alignments, where all the relations that were not part of a previous subgraph structure are aligned, and iv) Reentrancy Alignments, where all the reentrant nodes are aligned. In contrast to ISI, all the semantic units in the graph are aligned to some list of consecutive spans in the text. We use 150 alignments as the validation set and 200 as the test set, which includes sentence-graph pairs from *The Little Prince* Corpus (TLP) complemented with randomly sampled pairs from AMR 3.0.

¹The pseudo-algorithm is described in the Appendix C.

²The sentences of AMR 2.0 are a subset of AMR 3.0.

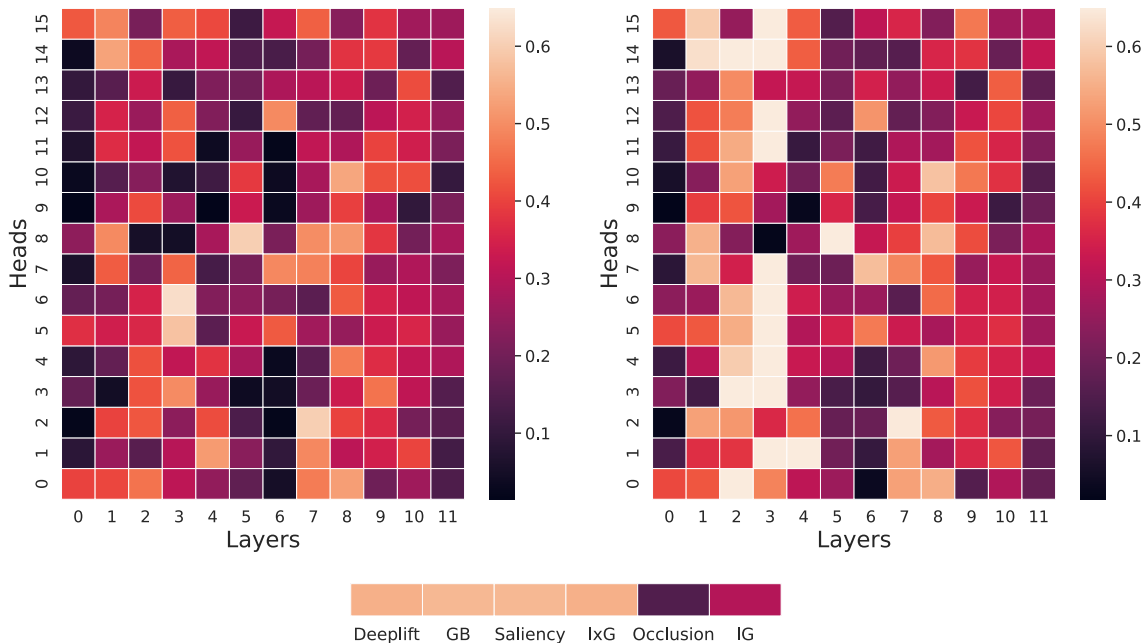


Figure 2: Heatmap of Pearson’s r correlation to LEAMR validation set for unguided (left) and guided on half the heads in layer 3 (right) cross-attention weights, as well as saliency methods (bottom).

4.3 Model

We use SPRING (Bevilacqua et al., 2021) as our parsing model based on the BART-large architecture (Lewis et al., 2020) for English and SPRING based on mBART for non-English languages mBART (Liu et al., 2020) for the multilingual setting. We extract all att_h^ℓ matrices from a model trained on AMR 3.0 as in Blloshmi et al. (2021) in order to perform our unguided cross-attention analysis. For the guided approach we re-train using the same hyperparameters as the original implementation, but with an extra loss signal as described in Section 3.2 based on either LEAMR or ISI. When using LEAMR alignments, we restructure the training split in order to exclude any pair from their test and validation sets.

5 Experiments

5.1 Correlation

In this study, we investigate the correlation between cross-attention and alignment by computing the Pearson’s r correlation coefficient between the att_h^ℓ matrix and the LEAMR alignment matrix $align$. To do so, we first flatten the matrices and remove any special tokens that are not relevant for alignment. As shown in Figure 2, there is a clear positive correlation between the two.

While we do not have a clear explanation for

why certain heads have a higher correlation than others, it is evident that there is a connection between cross-attention and alignment. For example, head 6 in layer 3 (i.e., att_6^3) has a correlation coefficient of 0.635, approximately the same as the sum of the entire layer.

With regard to the saliency methods described in Section 3.3, the two most highly correlated methods were Saliency and GB, with a correlation coefficient of 0.575. Despite this result, we observe that saliency methods tend to focus more on essential parts of the sentence, such as the subject or predicate. These parts are usually aligned to more nodes and relations, which explains the high correlation, but they lack nuance compared to cross-attention.

Our best results were obtained by supervising layer 3 during training with the approach outlined in Section 3.2, using Cross-Entropy Loss on half of the heads (i.e., 3, 4, 5, 6, 7, 11, 12, and 15) that were selected based on their correlation on the validation set. This did not affect the performance of parsing. When we looked at att^3 using the learned weighted mix from Equation 1 with LEAMR alignments, the correlation reached 0.866, which is significantly higher than any other method. Figure 2 shows the impact of supervising half the heads on layer 3 and how it influences heads in other layers.

To gain a better understanding of these results, we present an example from the TLP corpus in Fig-



Figure 3: Unguided (left), saliency (center-left) and guided (center-right) alignment weights and LEAMR (right) gold alignment for lpp_1943.1209. To explore all cross-attention weights interactively, please go [here](#).

Figure 3 to illustrate the different methods, including cross-attention and saliency methods. The left image shows the cross-attention values for att_6^3 . Despite not having seen any alignment information, the model is able to correctly match non-trivial concepts such as "merchant" and "person". The center-left image illustrates how saliency methods focus on essential parts of the sentence, but lack nuance compared to cross-attention. The center-right image shows that supervising learning on layer 3 results in more condensed attention, which is associated with the improvement in correlation. However, it is important to note that the model can reliably attend to incorrect positions, such as aligning "pointer" to "merchant" instead of "sold".

5.2 Results

LEAMR Table 1 shows the performances of our two approaches on the LEAMR gold alignments compared to previous systems. We use the same evaluation setup as [Blodgett and Schneider \(2021\)](#), where the partial match assigns a partial credit from Jaccard indices between nodes and tokens. In both guided and unguided methods, we extract the score matrix for Algorithm 1 from the sum of the cross-attention in the first four layers. We use a Wilcoxon signed-rank test ([Wilcoxon, 1945](#)) on the alignment matches per graph to check for significant differences. Both our approaches are significantly different compared to LEAMR ($p=0.031$ and $p=0.007$ respectively). However, we find no statistical difference between our unguided and guided approaches ($p=0.481$).

Our guided attention approach performs best, improving upon LEAMR on Subgraph (+0.5) and Relation (+2.6). For Reentrancy, performance is relatively low, and we will explore the reasons for this in Section 7. Perhaps most interesting is the performance of the unguided system using raw cross-attention weights from SPRING. The system remains competitive against the guided model without having access to any alignment information. It outperforms LEAMR which, despite being unsupervised with respect to alignments, relies on a set of inductive biases and rules based on alignments. While we also draw on specific rules related to the graph structure in post-processing, we will need to investigate their impact in an ablation study.

Relations that are argument structures (i.e., $:ARG$ and $:ARG-of$) usually depend on the predictions for their parent or child nodes; hence their improvement would be expected to be tied to the Subgraph Alignment. The results in Table 2 reassure us that this intuition is correct. Notice how for Single Relations (such as $:domain$ or $:purpose$ in Figure 3) the performance by LEAMR was much lower, even worse than that of ISI: [Blodgett and Schneider \(2021\)](#) argued that this was due to the model being overeager to align to frequent prepositions such as *to* and *of*. On the other hand, our unguided method achieves 15 points over ISI and 20 over LEAMR, which hints at the implicit knowledge on alignment that cross-attention encodes. Our guided approach experiences a considerable drop for Single Relations since it was trained on data generated by LEAMR, replicating its faulty behavior albeit

		Exact Alignment			Partial Alignment			Spans	Coverage
		P	R	F1	P	R	F1	F1	
Subgraph Alignment (1707)	ISI	71.56	68.24	69.86	78.03	74.54	76.24	86.59	78.70
	JAMR	87.21	83.06	85.09	90.29	85.99	88.09	92.38	91.10
	TAMR	85.68	83.38	84.51	88.62	86.24	87.41	94.64	94.90
	LEAMR	93.91	94.02	93.97	95.69	95.81	95.75	96.05	100.00
	LEAMR †	93.74	93.91	93.82	95.51	95.68	95.60	95.54	100.00
	Ours - Unguided	94.11	94.49	94.30	96.03	96.42	96.26	95.94	100.00
	Ours - Guided - ISI	89.87	91.97	90.91	92.11	94.27	93.18	93.69	100.00
	Ours - Guided - LEAMR	94.39	94.67	94.53	96.62	96.90	96.76	96.40	100.00
Relation Alignment (1263)	ISI	59.28	8.51	14.89	66.32	9.52	16.65	83.09	9.80
	LEAMR	85.67	87.37	85.52	88.74	88.44	88.59	95.41	100.00
	LEAMR †	84.63	84.85	84.74	87.77	87.99	87.88	91.98	100.00
	Ours - Unguided	87.14	87.59	87.36	89.87	90.33	90.10	91.03	100.00
	Ours - Guided - ISI	83.82	83.39	83.61	86.45	86.00	86.22	87.30	100.00
	Ours - Guided - LEAMR	88.03	88.18	88.11	91.08	91.24	91.16	91.87	100.00
Reentrancy Alignment (293)	LEAMR	55.75	54.61	55.17	—	—	—	—	100.00
	LEAMR †	54.61	54.05	54.33	—	—	—	—	100.00
	Ours - Unguided	44.75	44.59	44.67	—	—	—	—	100.00
	Ours - Guided - ISI	42.09	39.35	40.77	—	—	—	—	100.00
	Ours - Guided - LEAMR	56.90	57.09	57.00	—	—	—	—	100.00
Duplicate Subgraph Alignment (17)	LEAMR	66.67	58.82	62.50	70.00	61.76	65.62	—	100.00
	LEAMR †	68.75	64.71	66.67	68.75	64.71	66.67	—	100.00
	Ours - Unguided	77.78	82.35	80.00	77.78	82.35	80.00	—	100.00
	Ours - Guided - ISI	63.16	70.59	66.67	65.79	73.53	69.44	—	100.00
	Ours - Guided - LEAMR	70.00	82.35	75.68	72.50	85.29	78.38	—	100.00

Table 1: LEAMR alignment results. Column blocks: models; Exact and Partial alignment scores; Span and Coverage measures. Row blocks: alignment types, number of instances in brackets. † indicates our re-implementation. Guided versions using ISI/LEAMR silver alignments. Bold is best.

	AMR parser	P	R	F1
ALL	ISI	59.3	08.5	14.9
	LEAMR †	84.6	84.9	84.7
	Ours - Unguided	87.1	87.6	87.4
	Ours - Guided - LEAMR	88.0	88.2	88.1
Single Relations (121)	ISI	82.9	52.1	64.0
	LEAMR †	64.8	55.7	59.9
	Ours - Unguided	79.5	79.5	79.5
	Ours - Guided - LEAMR	77.5	64.8	70.5
Argument Structure (1042)	ISI	39.6	03.5	06.4
	LEAMR †	86.6	88.2	87.4
	Ours - Unguided	87.9	88.4	88.2
	Ours - Guided - LEAMR	89.0	90.8	89.9

Table 2: LEAMR results breakdown for Relation Alignment. Column blocks: relation type; models; scores. Bold is best. † indicates our re-implementation.

being slightly more robust.

ISI When we test our systems against the ISI alignments, both our models achieve state-of-the-

art results, surpassing those of previous systems, including LEAMR. This highlights the flexibility of cross-attention as a standard-agnostic aligner (we provide additional information in Appendix B). Table 3 shows the performance of our systems and compares ones with the ISI alignment as a reference. We omit relations and Named Entities to focus solely on non-rule-based alignments and have a fair comparison between systems. Here, our aligner does not rely on any span-segmentation, hence nodes and spans are aligned solely based on which words and nodes share the highest cross-attention values. Still, both our alignments outperform those of the comparison systems in English.

Moreover, only our approach achieves competitive results in Spanish, German and Italian – obtaining 40 points more on average above the second best model – while the other approaches are hampered by the use of English-specific rules. However, we found two reasons why non-English systems

	EN			DE			ES			IT			AVG		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
JAMR	92.7	80.1	85.9	75.4	6.6	12.1	84.4	16.1	27.1	64.8	13.2	21.9	79.3	29.0	36.8
TAMR	92.1	84.5	88.1	73.7	6.4	11.8	84.0	16.4	27.5	64.3	13.2	21.9	78.5	30.1	37.3
LEAMR	85.9	92.3	89.0	8.4	9.3	8.8	8.1	9.0	8.5	9.0	9.5	9.3	27.9	30.0	28.9
Unguided	95.4	93.2	94.3	64.0	74.4	68.85	67.9	77.1	72.2	67.4	75.5	71.2	73.7	80.1	76.6
Guided	96.3	94.2	95.2	—	—	—	—	—	—	—	—	—	—	—	—

Table 3: ISI results. Column blocks: models, language.

	GOLD			Without Rules			Layers										
	LEAMR †	Ung.	Guided	LEAMR †	Ung.	Guided	Unguided				Guided						
							Sal.	[0:4]	[4:8]	[8:12]	[0:12]	[0:4]	[4:8]	[8:12]	[0:12]	[3]	[3]*
Sub.	96.5	96.7	97.0	87.6	88.6	93.4	62.2	94.3	69.8	63.3	87.7	94.5	74.4	66.3	93.2	93.7	93.7
Rel.	87.1	89.2	90.3	26.6	60.1	83.4	50.0	87.7	72.7	61.6	84.5	88.1	73.8	62.5	87.9	86.2	85.9
Reen.	56.8	46.7	59.0	15.2	38.6	57.0	34.5	44.7	41.1	36.1	41.9	57.0	39.2	33.0	51.0	52.7	53.4
Dupl.	62.9	80.0	75.7	40.0	71.8	73.7	9.5	80.0	11.1	27.3	64.3	75.9	30.0	27.3	66.7	70.3	70.3

Table 4: F1 results on Exact Alignment on ablation studies. Column blocks: alignment types; using gold spans; removing rules from the models; by layers. Guided approach using LEAMR silver alignments. † indicates our re-implementation. [x:y] indicates sum from layer x to y. * indicates weighted head sum. Bold is best.

perform worse than in English: 1) linguistic divergences (as explained in Wein and Schneider (2021)), and ii) the machine translation alignment error.

6 Ablation Study

Gold spans LEAMR relies on a span segmentation phase, with a set of multiword expressions and Stanza-based named entity recognition. We use the same system in order to have matching sentence spans. However, these sometimes differ from the gold spans, leading to errors. Table 4 (left) shows performance using an oracle that provides gold spans, demonstrating how our approach still outperforms LEAMR across all categories.

Rules All modern alignment systems depend on rules to some degree. For instance, we use the subgraph structure for Named Entities, certain relations are matched to their parent or child nodes, etc. (see Appendix A for more details). But what is the impact of such rules? As expected, both LEAMR and our unguided method see a considerable performance drop when we remove them. For Relation, LEAMR drops by almost 60 points, since it relies heavily on the predictions of parent and child nodes to provide candidates to the EM model. Our unguided approach also suffers from such dependency, losing 25 points. However, our guided model is resilient to rule removal, dropping by barely one point on Subgraph and 5 points on Relation.

Layers Figure 2 shows how alignment acts

differently across heads and layers. We explore this information flow in the Decoder by extracting the alignments from the sum of layers at different depths. The right of Table 4 shows this for both our unguided and guided models, as well as the Saliency method. [3] indicates the sum of heads in the supervised layer, while [3]* is the learned weighted mix. From our results early layers seem to align more explicitly, with performance dropping with depth. This corroborates the idea that Transformer models encode basic semantic information early (Tenney et al., 2019). While layers 7 and 8 did show high correlation values, the cross-attention becomes more disperse with depth, probably due to each token encoding more contextual information.

7 Error Analysis

We identify two main classes of error that undermine the extraction of alignments.

Consecutive spans Because each subgraph in LEAMR is aligned to a list of successive spans, the standard cannot deal correctly with transitive phrasal verbs. For example, for the verb "take off" the direct object might appear in-between ("He took his jacket off in Málaga"). Because these are not consecutive spans, we align just to "take" or "off".

Rules We have a few rules for recognizing subgraph structures, such as Named Entities, and align them to the same spans. However, Named Entity structures contain a placeholder node indicating the entity type; when the placeholder node appears explicitly in the sentence, the node should not be

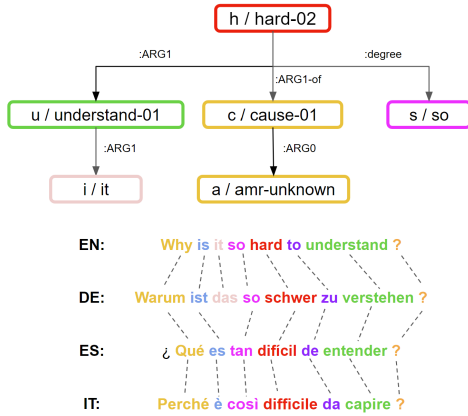


Figure 4: AMR graph and its sentence from "AMR 2.0 – Four Translation". Color represents the alignment.

part of the Named Entity subgraph. For example, when aligning 'Málaga', the city, the placeholder node should be aligned to *city* while our model aligned it to *Málaga*.

8 Cross-lingual Analysis: A Case Study

To investigate the potential causes of misalignment between English and non-English languages, we conduct a case study that qualitatively examines the differences in alignment generated by different systems and languages. Figure 4 illustrates the sentence, "why is it so hard to understand?" with its human translations in German, Spanish, and Italian, and its AMR. In the Italian translation, the subject of the verb is omitted, while in the Spanish translation the focus of the question is modified from asking the reason *why something is difficult to understand* to asking directly *what is difficult to understand*, making "qué" the subject. As a consequence, in both cases making it impossible to align "it" with any word in either the Italian or Spanish sentence by Machine Translation Alignment. Table 5 presents the alignments generated for the sentence in each language and with each model in ISI format. Although our model was able to align the node "it" by aligning it with the conjugated verb in the Italian sentence and with the word "qué" in the Spanish sentence, which serves as the subject, this resulted in an error in our evaluation since the alignment of "it" was not projected in either Italian or Spanish. In addition, we also observed the performance of jamr and tamr, which are rule-based systems, and found that they were only able to align the word "so" in the German translation, as it shares the same lemma in English. In contrast, LEAMR was able to detect more alignments

En	Why	is	it	so	hard	to	understand	?
ref	1.2 1.2.1	—	1.1.1	1.3	1	—	1.1	—
ours	1.2 1.2.1	—	1.1.1	1.3	1	—	1.1	—
jamr	—	—	1.1.1	1.3	1	—	1.1	—
tamr	—	—	1.1.1	1.3	1 1.2	—	1.1	—
leamr	1.2	1.1.1	1.1.1	1.3	1	—	1.1	1.2.1
De	Warum	ist	das	so	schwer	zu	verstehen	?
ref	1.2 1.2.1	—	1.1.1	1.3	1	—	1.1	—
ours	1.2 1.2.1	—	1.1.1	1.3	1	—	1.1	—
jamr	—	—	—	1.3	—	—	—	—
tamr	—	—	—	1.3	—	—	—	—
leamr	1	—	1.2	1.2	1.3	1.1.1	1.2.1	—
Es	Qué	es	—	tan	difícil	de	entender	?
ref	1.2 1.2.1	—	—	1.3	1	—	1.1	—
ours	1.1.1 1.2.1	—	—	1.2 1.3	1	—	1.1	—
jamr	—	—	—	—	—	—	—	—
tamr	—	—	—	—	—	—	—	—
leamr	1.1.1 1	1.2	—	—	1.1	—	1.3	1.2.1
It	Perché	é	—	cosí	difficile	da	capire	?
ref	1.2 1.2.1	—	—	1.3	1	—	1.1	—
ours	1.2 1.2.1	1.1.1	—	1.3	1	—	1.1	—
jamr	—	—	—	—	—	—	—	—
tamr	—	—	—	—	—	—	—	—
leamr	1	1.1	—	1.3	1.1.1	—	1.2.1 1.2	—

Table 5: Alignments between sentences and graph from Figure 4 across different system. "ref" is the reference alignment obtained by Machine Translation Alignment.

due to its requirement to align all nodes to a corresponding word in the target language. However, the alignments generated by LEAMR appeared to be almost entirely random.

9 Conclusion

In this paper, we have presented the first AMR aligner that can scale cross-lingually and demonstrated how cross-attention is closely tied to alignment in AMR Parsing. Our approach outperforms previous aligners in English, being the first to align cross-lingual AMR graphs. We leverage the cross-attention from current AMR parsers, without overhead computation or affecting parsing quality. Moreover, our approach is more resilient to the lack of handcrafted rules, highlighting its capability as a standard- and language-agnostic aligner, paving the way for further NLP tasks. As a future direction, we aim to conduct an analysis of the attention heads that are not correlated with the alignment information in order to identify the type of information they capture, such as predicate identification, semantic relations, and other factors. Additionally, we plan to investigate how the alignment information is captured across different NLP tasks and languages in the cross-attention mechanism of sequence-to-sequence models. Such analysis can provide insights into the inner workings of the models and improve our understanding of how to enhance their performance in cross-lingual settings.

10 Limitations

Despite the promising results achieved by our proposed method, there are certain limitations that need to be noted. Firstly, our approach relies heavily on the use of Transformer models, which can be computationally expensive to train and run. Additionally, the lower performance of our aligner for languages other than English is still a substantial shortcoming, which is discussed in Section 5.2.

Furthermore, our method is not adaptable to non-Transformer architectures, as it relies on the specific properties of Transformer-based models to extract alignment information.

Lastly, our method is based on the assumption that the decoder will attend to those input tokens that are more relevant to predicting the next one. However, this assumption may not always hold true in practice, which could lead to suboptimal alignments.

In conclusion, while our proposed method presents a promising approach for cross-lingual AMR alignment, it is important to consider the aforementioned limitations when applying our method to real-world scenarios. Future research could focus on addressing these limitations and exploring ways to improve the performance of our aligner for languages other than English.

11 Ethics Statement


While our approach has shown itself to be effective in aligning units and spans in sentences of different languages, it is important to consider the ethical and social implications of our work.

One potential concern is the use of Transformer-based models, which have been shown to perpetuate societal biases present in the data used for training. Our approach relies on the use of these models, and it is therefore crucial to ensure that the data used for training is diverse and unbiased. Furthermore, the use of cross-attention in our approach could introduce new ways to supervise a model in order to produce harmful or unwanted model predictions. Therefore, it is crucial to consider the ethical implications of any guidance or supervision applied to models and to ensure that any training data used to guide the model is unbiased and does not perpetuate harmful stereotypes or discrimination.

Additionally, it is important to consider the potential impact of our work on under-resourced languages. While our approach has shown to be ef-

fective in aligning units and spans in sentences of different languages, it is important to note that the performance gap for languages other than English still exists. Further research is needed to ensure that our approach is accessible and beneficial for under-resourced languages.

Acknowledgments

The authors gratefully acknowledge the support of the European Union's Horizon 2020 research project *Knowledge Graphs at Scale* (Knowledge Graphs) under the Marie Marie Skłodowska-Curie grant agreement No 860801. 

The last author gratefully acknowledges the support of the PNRR MUR project PE0000013-FAIR.

References

- Rafael Anchieta and Thiago Pardo. 2020. [Semantically inspired AMR alignment for the Portuguese language](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1595–1600, Online. Association for Computational Linguistics.
- Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. [Graph pre-training for AMR parsing and generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Dis-course*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Jasmijn Bastings and Katja Filippova. 2020. [The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?](#) In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. 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](#). In *Proceedings of AAAI*.
- Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Souza Wilkens, Xiaou Wang, Thomas François, and Patrick Watrin. 2022. [Is attention explanation? an](#)

- introduction to the debate. In *Association for Computational Linguistics. Annual Meeting. Conference Proceedings*.
- Rexhina Blloshmi, Michele Bevilacqua, Edoardo Fabiano, Valentina Caruso, and Roberto Navigli. 2021. [SPRING Goes Online: End-to-End AMR Parsing and Generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 134–142, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- Austin Blodgett and Nathan Schneider. 2021. [Probabilistic, structure-aware algorithms for improved variety, accuracy, and coverage of AMR alignments](#). 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 3310–3321, Online. Association for Computational Linguistics.
- Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. [Dialogue-AMR: Abstract Meaning Representation for dialogue](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.
- Chi Chen, Maosong Sun, and Yang Liu. 2021. [Mask-align: Self-supervised neural word alignment](#). 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 4781–4791, 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.
- Ameet Deshpande and Karthik Narasimhan. 2020. [Guiding attention for self-supervised learning with transformers](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4676–4686, Online. Association for Computational Linguistics.
- Zi-Yi Dou and Graham Neubig. 2021. [Word alignment by fine-tuning embeddings on parallel corpora](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.
- Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. [A discriminative graph-based parser for the Abstract Meaning Representation](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.
- 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.
- Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernández 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.
- Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. [Captum: A unified and generic model interpretability library for pytorch](#).
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- 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.

- Jungwoo Lim, Dongsuk Oh, Yoonna Jang, Kisu Yang, and Heuseok Lim. 2020. [I know what you asked: Graph path learning using AMR for commonsense reasoning](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2459–2471, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Yijia Liu, Wanxiang Che, Bo Zheng, Bing Qin, and Ting Liu. 2018. [An AMR aligner tuned by transition-based parser](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2430, Brussels, Belgium. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Potsawee Manakul and Mark Gales. 2021. [Long-span summarization via local attention and content selection](#). 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 6026–6041, Online. Association for Computational Linguistics.
- Abelardo Carlos Martínez Lorenzo, Marco Maru, and Roberto Navigli. 2022. [Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1727–1741, Dublin, Ireland. Association for Computational Linguistics.
- André F. T. Martins and Ramón F. Astudillo. 2016. [From softmax to sparsemax: A sparse model of attention and multi-label classification](#). In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, page 1614–1623. JMLR.org.
- Roberto Navigli, Rexhina Blloshmi, and Abelardo Carlos Martinez Lorenzo. 2022. [BabelNet Meaning Representation: A Fully Semantic Formalism to Overcome Language Barriers](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36.
- K. Elif Oral and Gülşen Eryiğit. 2022. [AMR alignment for morphologically-rich and pro-drop languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 143–152, Dublin, Ireland. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Nima Pourdamghani, Yang Gao, Ulf Hermjakob, and Kevin Knight. 2014. [Aligning English strings with Abstract Meaning Representation graphs](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 425–429, Doha, Qatar. Association for Computational Linguistics.
- Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. [Biomedical event extraction using Abstract Meaning Representation](#). In *BioNLP 2017*, pages 126–135, Vancouver, Canada,. Association for Computational Linguistics.
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. [Learning important features through propagating activation differences](#). In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 3145–3153. JMLR.org.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. [Deep inside convolutional networks: Visualising image classification models and saliency maps](#). *CoRR*, abs/1312.6034.
- Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. [Semantic neural machine translation using AMR](#). *Transactions of the Association for Computational Linguistics*, 7:19–31.
- Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. [Improving natural language processing tasks with human gaze-guided neural attention](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 6327–6341. Curran Associates, Inc.
- Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. [Striving for simplicity: The all convolutional net](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.
- Joe Stacey, Yonatan Belinkov, and Marek Rei. 2021. [Supervising model attention with human explanations for robust natural language inference](#).
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. [BERT rediscovered the classical NLP pipeline](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Sarah Uhrig, Yoalli Garcia, Juri Opitz, and Anette Frank. 2021. [Translate, then parse! a strong baseline for cross-lingual AMR parsing](#). In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 58–64, Online. Association for Computational Linguistics.

- Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. [Attention interpretability across nlp tasks](#). *CoRR*, abs/1909.11218.
- 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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Shira Wein and Nathan Schneider. 2021. [Classifying divergences in cross-lingual AMR pairs](#). In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 56–65, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Frank Wilcoxon. 1945. [Individual comparisons by ranking methods](#). *Biometrics Bulletin*, 1(6):80–83.
- Zhengxuan Wu, Thanh-Son Nguyen, and Desmond Ong. 2020. [Structured self-AttentionWeights encode semantics in sentiment analysis](#). In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 255–264, Online. Association for Computational Linguistics.
- Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. [Self-attention guided copy mechanism for abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1355–1362, Online. Association for Computational Linguistics.
- Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. [Do context-aware translation models pay the right attention?](#) 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 788–801, Online. Association for Computational Linguistics.
- Matthew D. Zeiler and Rob Fergus. 2014. [Visualizing and understanding convolutional networks](#). In *ECCV*.
- Shaolei Zhang and Yang Feng. 2021. [Modeling concentrated cross-attention for neural machine translation with Gaussian mixture model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1401–1411, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiawei Zhou, Tahira Naseem, Ramón Fernández Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021. [Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A LEAMR Alignment Rules

The LEAMR standard has some predefined strategies for alignments that were followed during their annotation, as well as fixed in their alignment pipeline along EM. We kept a few of these strategies when extracting the alignment, just those related to the structure of the graph, but not those concerning token matching between the sentence and the graph.

A.1 Subgraph

- Nodes *have-org-role-91* and *have-rel-role-91* follow a fixed structure related to a person ie. the sentence word *enemy* is represented as *person* → *have-rel-role-91* → *enemy*, therefore for such subgraphs we use the alignment from the child node.
- Similarly for Named Entities, we align the whole subgraph structure based on its child nodes which indicate its surfaceform. However this leads to some errors as described in Section 7.
- We align node *amr-unknown* to the question mark if it appears in the sentence.

A.2 Relations

- For the relation *:condition* we align it to the word *if* when it appears in the sentence.
- *:purpose* is aligned with *to* when in the sentence.
- *:ARGX* relations are aligned to the same span as the parent node, while *:ARGX-of* to that of the child, since they share the alignment of the predicate they are connected to.
- For *:mod* and *:duration* we use the alignment from the child node.
- For *:domain* and *:opX* we use the alignment from the parent node.

B Extra Results

B.1 LEAMR Results

We explore the variance with different seeds when guiding cross-attention. Table 1 reports on a single seed selected at random. Table 6 shows the results for five different seeds as well as the average and standard deviation. We observe some variance, especially for those alignment types with fewer

		Exact Alignment			Partial Alignment			Spans
		P	R	F1	P	R	F1	F1
Subgraph Alignment (1707)	Run 1	94.39	94.67	94.53	96.62	96.90	96.76	96.40
	Run 2	93.79	93.85	93.82	96.22	96.27	96.25	96.05
	Run 3	94.26	94.32	94.29	96.60	96.66	96.63	96.34
	Run 4	94.20	94.26	94.23	96.47	96.53	96.50	96.22
	Run 5	93.81	94.14	93.98	95.81	96.14	95.97	95.73
	Average	94.09	94.25	94.17	96.34	96.50	96.42	96.15
Std	0.27	0.30	0.28	0.34	0.30	0.32	0.27	
Relation Alignment (1263)	Run 1	88.03	88.18	88.11	91.08	91.24	91.16	91.87
	Run 2	87.90	88.36	88.13	90.71	91.18	90.95	91.87
	Run 3	88.61	88.61	88.61	91.44	91.44	91.44	91.95
	Run 4	88.39	88.61	88.50	91.02	91.25	91.14	91.66
	Run 5	88.59	88.44	88.52	91.24	91.08	91.16	91.86
	Average	88.30	88.44	88.37	91.10	91.24	91.17	91.84
Std	0.32	0.18	0.28	0.27	0.13	0.17	0.05	
Reentrancy Alignment (293)	Run 1	56.90	57.09	57.00	—	—	—	—
	Run 2	56.23	56.42	56.32	—	—	—	—
	Run 3	57.24	57.43	57.34	—	—	—	—
	Run 4	55.56	55.74	55.65	—	—	—	—
	Run 5	55.22	55.41	55.31	—	—	—	—
	Average	56.23	56.42	56.32	—	—	—	—
Std	0.86	0.86	0.86	—	—	—	—	
Duplicate Subgraph Alignment (17)	Run 1	70.00	82.35	75.88	72.50	85.29	78.38	—
	Run 2	65.00	76.47	70.27	67.50	79.41	72.97	—
	Run 3	70.00	82.35	75.68	70.00	82.35	75.68	—
	Run 4	73.68	82.35	77.78	76.32	85.29	80.56	—
	Run 5	70.00	82.35	75.68	70.00	82.35	75.68	—
	Average	69.74	81.17	75.06	71.26	82.94	76.65	—
Std	3.09	2.63	2.82	3.33	2.46	2.90	—	

Table 6: Results on the LEAMR alignment for 5 seeds on the guided approach. Column blocks: runs; measures. Row blocks: alignment types; average and standard deviation (std). Bold is best.

elements; however, average performance is always higher than previous approaches.

C Alignment Extraction Algorithm

Algorithm 1 shows the procedure for extracting the alignment between spans in the sentence and the semantic units in the graphs, using a matrix that weights Encoder tokens with the Decoder tokens

D AMR parsing

Since our guided approach was trained with a different loss than the SPRING model, it could influence the performance in the Semantic Parsing task.

Therefore, we also tested our model in the AMR parsing task using the test set of AMR 2.0 and AMR 3.0. Table 7 shows the result, where we can observe how our model preserves the performance on parsing.

	AMR 2.0	AMR 3.0
SPRING	84.3	83.0
Ours - Guided - ISI	84.3	83.0
Ours - Guided - Leamr	84.3	83.0

Table 7: AMR parsing Results.

Algorithm 1 Procedure for extracting the alignment between spans in the sentence and the semantic units in the graphs, using a matrix that weights Encoder tokens with the Decoder tokens.

```

1: function EXTRACTALIGNMENTS(encoderTokens, DecoderTokens, scoreMatrix)
2:   alignmentMap  $\leftarrow$  dict()
3:   spansList  $\leftarrow$  SPANS(encoderTokens)            $\triangleright$  Extract sentence spans as in LEAMR
4:   spanPosMap  $\leftarrow$  TOK2SPAN(encoderTokens)        $\triangleright$  Map input tokens to spans
5:   graphPosMap  $\leftarrow$  TOK2NODE(DecoderTokens)      $\triangleright$  Map output tokens to graph unit
6:   COMBINESUBWORDTOKENS(scoreMatrix)
7:   for DecoderTokenPos, GraphUnit in graphPosMap do
8:     encoderTokensScores  $\leftarrow$  scoreMatrix[DecoderTokenPos]
9:     maxScorePos  $\leftarrow$  ARGMAX(encoderTokensScores)
10:    alignmentMap[GraphUnit]  $\leftarrow$  SELECTSPAN(spansList, maxScorePos)
11:  end for
12:  fixedMatches  $\leftarrow$  GETFIXEDMATCHES(graphPosMap)  $\triangleright$  Look for rule based matches
13:  alignmentMap  $\leftarrow$  APPLYFIXEDMATCHES(alignmentMap, fixedMatches)
14:  alignments  $\leftarrow$  FORMATALIGNMENT(alignmentMap)
15:  return alignments
16: end function

```

E Hardware

Experiments were performed using a single NVIDIA 3090 GPU with 64GB of RAM and Intel® Core™ i9-10900KF CPU.

Training the model took 13 hours, 30 min per training epoch while evaluating on the validation set took 20 min at the end of each epoch. We selected the best performing epoch based on the SMATCH metric on the validation set.

F Data

The AMR data used in this paper is licensed under the *LDC User Agreement for Non-Members* for LDC subscribers, which can be found [here](#). The *The Little Prince* Corpus can be found [here](#) from the Information Science Institute of the University of Southern California.

G Limitations

Even though our method is an excellent alternative to the current AMR aligner system, which is standard and task-agnostic, we notice some drawbacks when moving to other autoregressive models or languages:

Model In this work, we studied how Cross Attention layers retain alignment information between input and output tokens in auto-regressive models. In Section 5.1, we examined which layers in state-of-the-art AMR parser models based on BART-large best preserve this information. Unfortunately, we cannot guarantee that these layers are

optimal for other auto-regressive models, and so on. As a result, an examination of cross-attention across multiple models should be done before developing the cross-lingual application of this approach.

Sentence Segmentation It is necessary to apply LEAMR’s Spam Segmentation technique to produce the alignment in LEAMR format (Section 3.4). However, this segmentation method has several flaws: i) As stated in Section 7, this approach does not deal appropriately with phrasal verbs and consecutive segments; ii) the algorithm is English-specific; it is dependent on English grammar rules that we are unable to project to other languages. Therefore we cannot extract the LEAMR alignments in a cross-lingual AMR parsing because we lack a segmentation procedure. However, although LEAMR alignment has this constraint, ISI alignment does not require any initial sentence segmentation and may thus be utilized cross-lingually.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section 9
- A2. Did you discuss any potential risks of your work?
Section 10
- A3. Do the abstract and introduction summarize the paper’s main claims?
In the Abstract and Introduction
- A4. Have you used AI writing assistants when working on this paper?
Grammarly, we check the use of English of our paper

B Did you use or create scientific artifacts?

Section 4, 5, 6 and 7

- B1. Did you cite the creators of artifacts you used?
Section 4
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In the Appendix
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We believe these are self-explained by the licences discussed for each artifact.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We base our work on widely used datasets which already performed these steps.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 and Appendix
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4 and Appendix.

C Did you run computational experiments?

Section 5

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 5 and Appendix

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Table 5 in Appendix

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Discussed through the paper when needed.

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.