Accounting for Language Effect in the Evaluation of Cross-lingual AMR Parsers

Shira Wein
Georgetown University
sw1158@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

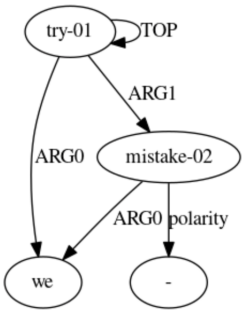
Abstract

Cross-lingual Abstract Meaning Representation (AMR) parsers are currently evaluated in comparison to gold English AMRs, despite parsing a language other than English, due to the lack of multilingual AMR evaluation metrics. This evaluation practice is problematic because of the established effect of source language on AMR structure. In this work, we present three multilingual adaptations of monolingual AMR evaluation metrics and compare the performance of these metrics to sentence-level human judgments. We then use our most highly correlated metric to evaluate the output of state-of-the-art cross-lingual AMR parsers, finding that Smatch may still be a useful metric in comparison to gold English AMRs, while our multilingual adaptation of S2match (XS2match) is best for comparison with gold in-language AMRs.

1 Introduction

The Abstract Meaning Representation (AMR; Banarescu et al., 2013) formalism captures the meaning of a sentence or phrase as a rooted, directed acyclic graph. Nodes correspond to concepts and the labeled edges reflect the relations between concepts. For example, the annotation in Figure 1 features the AMR annotation of the sentence "we will try not to make a mistake" in both PENMAN (text-based) and graph form. The edge labels can be arguments (core roles, denoted as `:argN`), or one of a number of non-core roles such as `:location` or `:manner`.

Cross-lingual AMR parsers convert non-English text to (English-focused) AMR graphs. As there are no existing multilingual AMR evaluation metrics and due to the limited availability of non-English gold AMR annotations, these cross-lingual AMR parsers have only ever been evaluated in comparison to gold *English* AMRs. This established approach of comparison to English AMRs using the monolingual Smatch metric needs to be considered more carefully because previous work has established that the source language has a dramatic effect on the cross-lingual AMRs (Damonte, 2019; Wein and Schneider, 2021, 2022).




```
(v2 / try-01
    :ARG0 (v1 / we)
    :ARG1 (v3 / mistake-02
        :polarity -
        :ARG0 v1))
```

Figure 1: An AMR annotation for the sentence "We will try not to make a mistake," in PENMAN text-based notation and as a rooted graph. The diagram was made on the AMREager website (Damonte et al., 2017).

We argue that cross-lingual AMR parsing should represent the *semantics*, beyond the lexicon, faithfully to the source language. Enabled by work developing sizable Spanish (Wein et al., 2022) and Chinese (Li et al., 2016) gold AMR corpora, we propose that cross-lingual AMR parsers should be evaluated on gold AMRs that match the language being parsed. In this work, we adapt three monolingual AMR evaluation metrics to a multilingual setting, and evaluate the performance of these metrics in comparison to human judgments of cross-lingual sentence similarity. We show that our adaptation of S2match (Opitz et al., 2020) which leverages LaBSE (Feng et al., 2022) embeddings is highly correlated with, and most correlated to, human judgment of similarity.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3824–3834
October 12–17, 2022.

Additionally, using our new cross-lingual evaluation metric, we evaluate the performance of cross-lingual AMR parsers and compare that with evaluations using the previously-used Smatch metric. This provides a new, informative ranking of existing cross-lingual AMR parsers and offers insight into the applicability of the monolingual Smatch metric for cross-lingual evaluation.

Ours is the first work to address the evaluation of cross-lingual AMR parsers with respect to the language being parsed. Our contributions include:

- Three multilingual adaptations of monolingual AMR evaluation metrics.
- Human judgments on Spanish-English and Chinese-English sentence pairs, corresponding to gold AMR pairs.
- An assessment of the effectiveness of existing monolingual and proposed multilingual metrics to evaluate similarity between cross-lingual AMR pairs, and correlations with human judgments for these metrics.
- An evaluation of state-of-the-art cross-lingual AMR parsers using both our new XS2match metric and Smatch, in comparison to gold English *and* gold Spanish AMRs.

Our code is available online at `https://github.com/shirawein/Crossling-AMR-Eval` to promote ease of cross-lingual AMR evaluation using our metrics.

2 Background

Abstract Meaning Representation parsers produce AMR annotations from natural text. "Cross-lingual AMR parsing" refers to parsing a non-English sentence into a standard English AMR (Damonte, 2019). Damonte and Cohen (2018) introduced the task of cross-lingual AMR parsing and developed non-English parsers by projecting from English annotations to their non-English counterparts through the use of parallel corpora. Current approaches to cross-lingual AMR parsing evaluate AMRs produced from text in four languages: Mandarin Chinese, Spanish, Italian, and German (§5.1). Language-specific AMR parsers have also been developed to parse from Mandarin Chinese (Wang et al., 2018), Portuguese (Anchiêta and Pardo, 2018), and Indonesian (Roaffa Ilmy and Leylia Khodra, 2021).

AMR parsers have traditionally been evaluated using the Smatch metric (Cai and Knight, 2013). In order to compare a pair of semantic graphs (system and gold), Smatch aligns their nodes, searching for a maximal alignment via hill climbing. With these alignments, triples representing edges of the two graphs are compared to compute an F-score.

Cross-lingual AMR parsers have been evaluated on gold English AMR graphs created from English sentences, paired with sentences that were manually translated from those English sentences into other languages. The English AMRs come from the AMR 2.0 dataset (Knight et al., 2017), and the non-English sentence translations from the AMR 2.0 - Four Translations dataset (Damonte and Cohen, 2020) with translations into Mandarin Chinese, Italian, Spanish, and German. Thus, cross-lingual AMR parsers have been evaluated by comparing the system-produced AMR parsed from the non-English sentence to the gold English AMR corresponding to the translated sentence—which does not take into account any effect the source language might have on AMR structure. We discuss this concern in §3.

In addition to the resources noted above, recent work has produced gold Spanish AMRs for AMR 2.0 - Four Translations Spanish sentences (Wein et al., 2022). As a result, we can compare the system output to in-language (Spanish) gold AMRs, and develop metrics to enable this comparison.

3 Developing Cross-lingual AMR Evaluation Metrics

When comparing English and non-English AMRs, the concepts themselves are in different languages and the structure of the AMR will also differ, as it is affected by the syntax and semantics of the language being parsed from (Damonte, 2019; Wein and Schneider, 2021; Blloshmi et al., 2020). Therefore, to be able to evaluate the similarity of AMRs in two different languages, there are likely changes that need to be made to the monolingual metric.

A naive assumption is that AMR should be structurally the same for parallel sentences regardless of language, because AMR encodes meaning and translation preserves meaning. However, previous work has demonstrated that this is not the case, and that even when lexical items are made to be monolingual, the source language has a marked effect on the AMR structure itself (for at least English and Chinese): Wein and Schneider (2022) reported Smatch scores consistently below 50% between English and Chinese parallel gold AMR graphs, even when all Chinese tokens are replaced by their

Original gold English AMR:

```
(s / surge-01
  :ARG1 (a / and
          :op1 (s2 / speed-01)
          :op2 (a2 / accident))
  :mod (a3 / as-well))
```

Original gold Spanish AMR:

```
(c0 / aumentar-01
    :manner (c1 / también)
    :ARG1 (c2 / y
              :op1 (c3 / exceder-01
                       :ARG1 (c4 / velocidad))
              :op2 (c5 / accidente)))
```

Our translated version of the gold Spanish AMR:

```
(c0 / increase
    :manner (c1 / also)
    :ARG1 (c2 / and
              :op1 (c3 / exceed
                       :ARG1 (c4 / speed))
              :op2 (c5 / accident)))
```

Figure 2: Parallel gold English and Spanish AMRs for the sentence "Speeding and accidents have surged as well" from Knight et al. (2017) and Wein et al. (2022) respectively, followed by our translated version of the gold Spanish AMR per the cross-lingual Smatch (XSmatch) method.

corresponding English AMRs. Therefore, we alter existing metrics to be able to compare (English) AMRs parsed from non-English sentences to gold in-language AMRs.

We consider the applicability of existing AMR metrics in cross-lingual parser evaluation and adapt them to function multilingually. Cross-lingual AMR parsers are currently evaluated via Smatch. Here we consider three metrics: Smatch, SemBleu, and S2match.

As mentioned in §2, Smatch aligns the semantic graphs via hill climbing. S2match (Opitz et al., 2020) incorporates word embeddings into Smatch to account for similarity of concept nodes without the same token being used. SemBleu (Song and Gildea, 2019) is based on the machine translation metric BLEU (Papineni et al., 2002). SemBleu does not involve variable alignment and instead converts the graph to a bag of k-grams.

Broadly, our approach to adapting these monolingual metrics is that we alter Smatch (§3.1) and SemBleu (§3.2) by translating the lexical material in the AMR graphs into English, and S2match (§3.3) by using cross-lingual embeddings.

3.1 XSmatch

In order to make Smatch (Cai and Knight, 2013) multilingual, we translate individual tokens within the non-English AMR to English. We use the EasyNMT package[1] for translation, which was also the translation package used in the cross-lingual parser of Uhrig et al. (2021). Specifically, we employ the Opus-MT model. Recall that Smatch (like the other evaluation metrics) compares AMR graphs rather than strings. Therefore, we are translating individual elements of the AMR and not the sentence itself. The elements of the AMR which we translate are the words in the instance and attribute triples. We also remove the word senses (numeric affixes to the concepts) for ease of translation and comparison. An example parallel gold English and gold Spanish AMR, plus our corresponding translated version of the gold Spanish AMR, can be seen in Figure 2.

We also developed a version of Smatch that aligns concepts across AMRs in different languages via fast_align (Dyer et al., 2013), and found that using machine translation was more reliable.

3.2 XSemBleu

To adapt SemBleu to function cross-lingually, we again translate the tokens in one of the AMRs, and additionally truncate the tokens (truncate after translation for the non-English AMR, and also truncate for the English AMR). SemBleu does not break the AMR into triples, so we instead translate the entire non-English AMR to an English AMR by iterating token by token over the AMR and determining whether the current token needs to be translated. For example, parentheses, digits, and roles starting with a colon do not need to be translated. This approach to translation is more intensive than the translation required for individual tokens in XSmatch. Therefore, we aim to account for translation discrepancies and errors (e.g. part-of-speech discrepancies) by truncating the translations to the first `n` tokens. In this case we use `n=5`. We use the default weights and smoothing function.

We suspected that SemBleu may be a better fit for cross-lingual AMR comparison because SemBleu prioritizes content over graph structure. One potential issue with SemBleu is that the nodes with higher connectivity are disproportionately weighted (Opitz et al., 2020).

[1] https://github.com/UKPLab/EasyNMT

3.3 XS2match

The current implementation of S2match relies on an external text file of embeddings, with the token being paired to an embedding in the file, and the embedding being retrieved from the text file for each token. To transport S2match to a multilingual format, we make use of the LaBSE (Feng et al., 2022) preprocesser and encoder. Where the existing S2match approach retrieved the embedding for a token from a text file, we elicit a constant tensor of the word, preprocess it, and encode it to a LaBSE embedding. Finally, we normalize the embedding and convert it to a numpy vector.

Consequently, we adjust the similarity computation when comparing the individual units of the two AMRs to matrix multiplication of two vectors, with one vector transposed.

We also trialed our approach with multilingual BERT (Devlin et al., 2018) and found that LaBSE was a more effective solution, though it takes longer to run than using multilingual BERT.

A benefit of XS2match is that, unlike XSmatch and XSemBleu, it does not rely on neural machine translation practices that could unduly benefit a parser using the same translation tool (e.g. Uhrig et al. (2021)) through exact lexical matching. When using XSmatch or XSemBleu to evaluate cross-lingual parser performance, it is worth verifying whether the translation approach is the same for the metric and the parser.

4 Analysis of Metrics

To compare cross-lingual AMR metrics, we need a source of ground truth about how the sentences in a translation pair relate to one another. For this we utilize human judgments of cross-lingual similarity. Because AMR is a meaning representation, the similarity scores of cross-lingual AMR pairs ideally should correlate with the similarity scores of their associated sentence pair. In line with previous work (Opitz et al., 2020), we determine the accuracy of our AMR metrics by calculating Pearson's correlation between system output and human judgments of cross-lingual sentence similarity. Specifically, we use gold AMRs as input to the metrics and calculate how correlated the metric-based similarity scores are to the human similarity ratings for the corresponding sentence pair. We normalize the AMRs by removing all wikification (links to the associated Wikipedia pages for entities in the AMR).

4.1 Collection of Human Judgments

We collect human judgments for 100 Spanish-English sentence pairs and 150 Mandarin Chinese-English sentence pairs which have associated gold AMRs. Both sets of data are doubly annotated by speakers fluent in both English and Chinese / Spanish.

We use both language pairs because Spanish and Chinese are notably syntactically distinct languages, and vary noticeably in cross-lingual AMR performance (§5.1). We also only use sentences which have associated gold AMRs, as opposed to existing sentence similarity metrics (Agirre et al., 2016), because we want to avoid introducing noise by relying on automatic parsers when comparing the AMR similarity with sentence similarity, or biasing our later assessment of cross-lingual parsers towards the parsers being used.

The sentences used come from the Chinese annotations of *The Little Prince* (Li et al., 2016) and the Spanish annotations (Wein et al., 2022) of AMR 2.0 - Four Translations (Damonte and Cohen, 2020). The parallel English sentences for both the Chinese and Spanish sentences are very related in meaning to the non-English sentences, so it was necessary to construct a dataset with varying degrees of sentence similarity (with all sentences still having associated gold AMRs).

In order to construct a Spanish-English dataset of varying similarities, 100 Spanish sentences from different genres in Damonte and Cohen (2020) were chosen. Then, a portion (25%) of the sentences were paired with English (from Knight et al., 2017) sentences with minimal to no similarity. Half of the sentences were paired with English sentences having a moderate amount of similarity / some divergence, as determined by being from the same relative part of a text and discussing the same topic without being a parallel sentence. The remaining 25% of the sentences were then paired with their parallel English sentences.

A similar approach was used when construct-

Annotator	0	1	2	3	4	5
Zh-Eng Anno. 1	35	15	5	17	41	37
Zh-Eng Anno. 2	40	9	3	10	25	63
Es-Eng Anno. 1	41	17	7	7	1	27
Es-Eng Anno. 2	34	15	14	8	6	23

Table 1: Distribution of human judgments of sentence similarity from 0-5 for each annotator. Zh-Eng annotators provided 150 judgments and Es-Eng annotators provided 100 judgments.

	Smatch	XSmatch	SemBleu	XSemBleu	XS2match	BERTscore
Zh-Eng Anno. 1	0.43	0.40	0.20	0.42	0.51	0.76
Zh-Eng Anno. 2	0.38	0.40	0.21	0.40	0.50	0.72
Zh-Eng Anno. Sum	0.41	0.41	0.21	0.42	0.51	0.75
Zh-Eng BERTscore	0.46	0.39	0.25	0.38	0.52	1.00
Es-Eng Anno. 1	0.69	0.79	0.37	0.60	0.77	0.87
Es-Eng Anno. 2	0.72	0.82	0.39	0.63	0.81	0.86
Es-Eng Anno Sum	0.72	0.82	0.38	0.63	0.80	0.88
Es-Eng BERTscore	0.74	0.82	0.41	0.62	0.79	1.00

Table 2: Pearson's correlation scores between the evaluation metrics (in the columns, along with BERTscores) and the human judgments of similarity (in the rows, with BERTscore, again). For each language pair, being Chinese-English and Spanish-English, we get the correlation with each of the two annotators as well as the sum of the similarity judgments.

ing the Chinese-English dataset, with 66% of the dataset being mostly parallel and 33% of the dataset being mostly divergent.

We asked human annotators to provide a score from 0 to 5 of how similar the content of a Spanish-English or Chinese-English sentence pair is. We use the task instructions from Agirre et al. (2016) as the basis for our instructions, "where 0 represents two sentences that are unrelated in meaning, and 5 indicates that the two sentences are perfect paraphrases of each other". We also added in degrees of similarity to the instructions to add clarity:

- (0) Completely unrelated
- (1) Not equivalent but share few subjects
- (2) Not equivalent but share some details
- (3) Roughly equivalent
- (4) Equivalent except for some details
- (5) Completely equivalent

We find that agreement for our sentence similarity protocol is high, with the correlation between annotator judgments being 0.93 for both the Spanish-English annotations and the Chinese-English annotations.[2] The distribution of the sentence similarity scores is not uniform (table 1).[3]

4.2 Results of Correlation Analysis

In order to assess the applicability of our metrics for cross-lingual AMR evaluation, we calculate Pearson's correlation for the human sentence similarity judgments and the AMR metrics. We also compare with the sentence-based metric BERTscore as a point of reference. These results can be seen in table 2.

[2]Annotator agreement for the SemEval task (Agirre et al., 2016) is not reported.

[3]For the Chinese-English sentences, we initially collected judgments from a third annotator, but that annotator's interpretation of similarity was skewed towards saying most of the valid translations were completely equivalent, so the data was not informative for studying degrees of similarity. As a result we used the data from two other annotators.

First, note that the use of translation is beneficial in SemBleu for both language pairs and in Smatch for Spanish-English. Applying translation to Chinese-English data has little effect on Smatch, for reasons discussed later in this section.

Comparing the three cross-lingual metrics, the two with the highest correlation to human judgment of sentence similarity are XSmatch and XS2match. While the correlation for Spanish-English is similar for those two multilingual metrics, though slightly higher via XSmatch, the correlation for Chinese-English is substantially higher using XS2match. As a result, we recommend that XS2match is likely the best metric to use for cross-lingual AMR parser evaluation.

Notably, though perhaps unsurprisingly, correlation with the Chinese-English human annotations is lower for all metrics than correlation with the Spanish-English human annotations. This is likely not due to any issues with the human annotation itself, because the annotations still correlate well with BERTscore judgment of similarity, as seen in the final column of table 2. Nonetheless, the Chinese-English human annotations are less correlated with BERTscore than the Spanish-English human annotations. Instead, the lower correlation with the Chinese-English annotations is likely due to lower performance on Chinese for the automatic machine translation systems and embeddings, as well as a greater degree of dissimilarity between the Chinese and English parallel AMRs than between the Spanish and English parallel AMRs. This greater degree of dissimilarity for certain AMR pairs has been studied previously (Xue et al., 2014) and is also evidenced here by the difference in the Smatch column in table 2. The baseline Smatch similarity, with no multilingual component, is already much more correlated with human judgments for Spanish-English than for Chinese-English.

The monolingual Smatch score is already highly correlated with sentence similarity (for English-Spanish in particular, but for both language pairs) because of structural similarity between the AMRs and matching between a subset of non-lexical nodes. For example, the Smatch scores aren't relying on lexical items as much as they are relying on the entities, e.g. shared name entities. This presence of names and named entities may also affect these correlation scores across languages because the Spanish-English text is from the news domain, which includes many country and person names, whereas the Chinese-English text is *The Little Prince*, which includes fewer of these named entities. This finding is a benefit of our approach to consider two different languages and text domains in our correlation analysis; as a result our recommendation to use XS2match for cross-lingual AMR evaluation is a more robust one.

Even with the translation and truncation practices, XSemBleu correlation does not exceed XS2match correlation for either language pair. We hoped that SemBleu might be able to overcome structural differences between cross-lingual AMR pairs, but the undesirable presence of bias in the metric, which cannot be overcome without introducing a different bias (Opitz et al., 2020), likely led to the consequence of correlating less with the human annotations than the other metrics. Still, XSemBleu correlates fairly well with both language pairs.

We also measure correlation with scores from the BERTscore metric (Zhang et al., 2020), which uses the sentences directly and not the AMR graphs. BERTscore uses BERT-based models to compare embeddings of the words in the candidate and reference sentence via cosine similarity. We use BERTscore with the `bert-base-multilingual-cased` model as is the default for multilingual pairs. The last column of table 2 shows that BERTscore achieves very strong correlations with human judgments, which can be interpreted as validating those judgments. Recall that our ultimate goal is to arrive at a cross-lingual AMR metric to compare AMR parsers, not to compare the raw sentences, so we use BERTscore here to validate the human judgments. Reassuringly, the AMR metrics are not too far behind BERTscore.[4] Rows 4 and 8 compare the AMR metrics with BERTscore, showing that they are about as well correlated with each other as the metrics are with human judgments.

We also verify that sentence length is not a confounding variable in these judgments, with the correlation between average sentence length and human similarity score being only 0.07.

5 Evaluating Cross-lingual AMR Parsers

Now that we have assessed the metrics discussed in §4 on gold AMRs in comparison to human judgments, we are interested in seeing how existing cross-lingual AMR parsers perform on our recommended cross-lingual metric versus on monolingual Smatch.

5.1 Approach to Parser Evaluation

We compare the performances of four state-of-the-art cross-lingual AMR parsers: SGL (Procopio et al., 2021), Bilingual Information for Cross-lingual AMR Parsing ("BI") (Cai et al., 2021), XLPT-AMR (Xu et al., 2021), and Translate then Parse ("TP") (Uhrig et al., 2021).

The SGL semantic parser (Procopio et al., 2021) is a seq2seq architecture trained for neural machine translation. SGL as a cross-lingual AMR parser works well in a zero-shot setting (without seeing any non-English AMR examples in training). Using their mBART + AP (where AP stands for annotation projection) model, SGL reports Smatch scores of 73.3, 73.9, 73.4, and 64.9 for German, Spanish, Italian, and Mandarin Chinese respectively on machine translations of the test set.

Cai et al.'s (2021) AMR parser (which here we call "BI" because of its use of bilingual information) introduces translated and non-English texts into the training of a seq2seq parser, to better predict non-English concepts. BI reports Smatch scores of 64.0, 65.4, 67.3, and 56.5 for German, Spanish, Italian, and Chinese respectively.

XLPT-AMR (Xu et al., 2021) approaches zero-shot AMR parsing via multi-task learning. XLPT-AMR reports Smatch scores of 70.5, 71.8, and 70.8 for German, Spanish, and Italian respectively; XLPT-AMR was not evaluated on Chinese data.

Translate then Parse (Uhrig et al., 2021) takes a simple approach to cross-lingual AMR parsing: translating the non-English sentence to English and then parsing with an English AMR parser (`amrlib`).[5] Translate then Parse ("TP") claims Ger-

[4] This is unsurprisingly especially true for XS2match, which uses LaBSE embeddings (BERT-based cross-lingual sentence embeddings).

[5] https://github.com/bjascob/amrlib

	BI	XLPT-AMR	SGL	TP
Consensus	0.737	0.733	0.655	0.756
DFA	0.703	0.685	0.652	0.722
Bolt	0.671	0.676	0.608	0.708
Proxy	0.776	0.785	0.737	0.808
Xinhua	0.651	0.682	0.685	0.724
Average	**0.708**	**0.712**	**0.669**	**0.744**

Table 3: XS2match scores in comparison to gold Spanish AMRs for each parser on every subset of data in the evaluation. The column labels are automatic AMR parsers and the row labels are the five genres of data in the Spanish AMR corpus.

man, Spanish, Italian, and Chinese Smatch scores of 67.6, 72.3, 70.7, and 59.1, respectively.

While these cross-lingual parsers have previously only been evaluated in comparison to gold English AMRs via Smatch, we now perform three comparisons on a substantial subset of the AMR 2.0 (Knight et al., 2017) (and AMR 2.0 - Four Translations (Damonte and Cohen, 2020)) dataset: (1) Smatch evaluation in comparison to gold English AMRs, as was used for evaluation of these parsers in previous work, (2) Smatch evaluation in comparison to gold Spanish AMRs, and (3) XS2match evaluation in comparison to gold Spanish AMRs.

In this section, we consider the subset (486 sentences) of AMR 2.0 - Four Translations that is annotated in the Spanish AMR corpus (Wein et al., 2022). We then retrieve the parser data, either by contacting the authors of the work or by running the parser ourselves, for all of those Spanish sentences. Though it is the traditional form of cross-lingual AMR parser evaluation, we compare the system output to the English gold AMRs via Smatch, so that we have a direct comparison on this subset of data with our two completely novel sets of evaluation for these parsers: in comparison to gold Spanish AMRs, via Smatch as well as XS2match.[6]

5.2 Analysis of Results

English AMRs have been viewed as a proxy for evaluating parser output for Spanish sentences, but Spanish AMRs should be the true gold standard as they are not corrupted by translation divergences. In this subsection, we empirically assess how much this difference makes for comparing and ranking parser performance. To do this, we compare cross-lingual parser performance via the traditional method of comparing output to gold English AMRs, as well as using the existing monolingual method of Smatch and our proposed multilingual method of XS2match (S2match with multilingual embeddings) in comparison to the gold Spanish AMRs.

[6]We perform this comparison exclusively with Spanish gold AMRs because we want to focus on the sentences that have been used by previous work in the evaluation of cross-lingual AMR parsing, namely, the AMR 2.0 - Four Translations dataset. Only the Spanish sentences in this dataset have gold AMRs.

Metric	BI	XLPT-AMR	SGL	TP
Eng. Smatch	0.682	0.680	0.582	0.696
Span. Smatch	0.378	0.378	0.382	0.408
Span. XS2match	0.708	0.712	0.669	0.744

Table 4: Average evaluation scores for each of the three metrics considered for the cross-lingual parsers.

Since we are comparing the system parse of a Spanish sentence to a parallel gold AMR, the higher the score (regardless of metric), the more similar the output is to the gold Spanish AMR, and thus the better the system output. We calculate the average scores for each AMR parser by retrieving the score for each of the five texts included in the evaluation dataset and averaging them with the same weight. We opt for a macro-average by text because of comparable text sizes.

Table 4 shows the average score by metric for each of the four parsers. Monolingual Smatch puts the comparison to gold Spanish AMRs at a disadvantage because the system output is parsed into an English AMR. Therefore the lexical similarity is not considered between the two AMRs, and monolingual Smatch is not an effective tool for comparing cross-lingual AMR parser output to gold AMRs of the same language as the source sentence.

With the intent to find a method to compare to gold AMRs in the source sentence, we have already found that XS2match is a good choice for this type of evaluation in §4.2. We find that the Spanish XS2match scores are slightly higher for all parsers than the English Smatch scores, which indicates that it is actually not only a more *justified* comparison than a comparison against English AMRs as it accounts for source language, but also a *fairer and more accurate* comparison because the parallel AMRs are indeed being judged as more parallel.

Ultimately we find that Smatch, when comparing to English gold AMRs, provides a similar ranking and scores for the cross-lingual AMR parsers as XS2match does when comparing to Spanish AMR. Note that in table 4, the system-level comparison by average English Smatch and average Spanish XS2match is very comparable. When considering

a ranking of parser performance, our empirical results suggest that monolingual Smatch serves as a reasonable proxy of parser ranking in the absence of in-language gold AMRs. However, it is still important to note that absolute scores from monolingual Smatch are artificially depressed in the cross-lingual scenario, meaning that faithfulness to the original Spanish sentence is being rewarded. Therefore, this monolingual evaluation does not provide a sufficient substitute for the in-language comparison via XS2match due to the fact that monolingual Smatch against English AMRs does not account for the dramatic effect and importance of language on AMR structure (Wein and Schneider, 2021).

Figures 3 and 4 show scatterplots of the average performance of the four parsers. Notably, the SGL parser, which is zero-shot, performs the worst of the four parsers when using XS2match in comparison to Spanish gold AMRs (and in comparison to English gold AMRs via Smatch). XLPT-AMR is also zero-shot but performs slightly better than SGL on the English Smatch and Spanish XS2match comparisons. BI, which incorporates additional bilingual informations, performs similarly to XLPT-AMR, achieving only slightly lower scores.

Translate then Parse ("TP") performs best on all three evaluations. The XS2match scores for each parser on every text can be seen in table 3. While the four metrics achieve similarly high scores, the output from the Translate then Parse system consistently produces the highest similarity score via XS2match across all five texts. This suggests that using the highly accurate machine translation via EasyNMT as a pre-processing step, before involving any AMR parsing, is an effective way of capturing the linguistic information of the source sentence. This is perhaps surprising because the sentence is immediately translated, but less surprising due to the challenging nature of cross-lingual AMR parsing, given that none of the cross-lingual parsers are trained on gold non-English AMRs. The ability of the machine translation system to account for cross-linguistic divergence, as noted by Uhrig et al. (2021), enables an effective monolingual English AMR parser to work well in this setting.

6 Background on Other Evaluation Metrics

Other metrics which we did not adapt in this paper have been proposed for AMR evaluation.

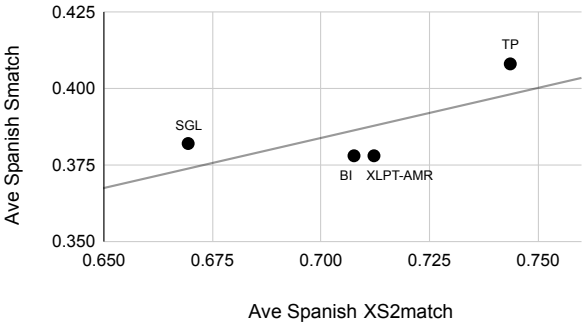


Figure 3: Average Spanish Smatch vs Average Spanish XS2match

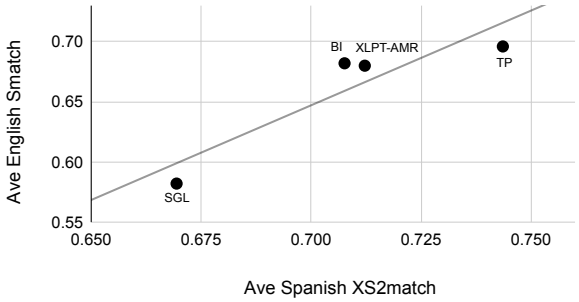


Figure 4: Average English Smatch vs Average Spanish XS2match

A document-level version of Smatch (Naseem et al., 2022) (as opposed to sentence-level) aligns the roots of its sentence-level AMR subgraphs. Similarly, Cai and Lam (2019) produces a version of Smatch designed to specifically consider core semantics, called Smatch-weighted.

SEMA (Anchiêta et al., 2019) extends Smatch by taking a breadth-first search approach to computing the maximum score; Smatch relies on one-to-one variable matching. The evaluation is limited and the metric is only shown to be *stricter* than Smatch.

Another existing monolingual AMR metric we did not consider in this work is MF_β. MF_β (Opitz and Frank, 2021) measures how easily an AMR can be reconstructed by AMR parsers and measures the grammaticality of the produced text. MF_β evaluation is more suited to AMR-to-text generation evaluation than to text-to-AMR parsing.

Goodman (2019) presents four AMR normalization techniques to ensure that isomorphic AMRs are evaluated as equivalent.

The BAMBOO suite (Opitz et al., 2021) houses various AMR similarity metrics to be able to assess the strengths and weaknesses of each metric.

7 Conclusion and Future Work

Our analysis of evaluation of cross-lingual AMR parsers indicates the usefulness of XS2match as a multilingual evaluation method. We recommend this approach as a way to compare AMRs parsed from non-English sentences to their gold non-English equivalents, while exploring additional alternatives in our work. We also find that using Smatch in comparison to gold English AMRs may be a useful tool for ranking cross-lingual AMR parser performance in the absence of in-language gold AMRs. With the future production of non-English gold AMRs, the evaluation of cross-lingual AMR parsers using our proposed metric will be more robust, accounting for the effect of source language on AMR.

8 Acknowledgements

Thank you to Yang Janet Liu, Yilun Zhu, Siyao Logan Peng, and Roy Ilany for providing human judgments. Thanks to Rexhina Blloshmi, Yitao Cai, Luigi Procopio, and Dongqin Xu for providing system results. We thank anonymous reviewers for their feedback. This work is supported by a Clare Boothe Luce Scholarship.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Rafael T Anchiêta, Marco AS Cabezudo, and Thiago AS Pardo. 2019. SEMA: an extended semantic evaluation metric for AMR. *arXiv preprint arXiv:1905.12069*.

Rafael Torres Anchiêta and Thiago Alexandre Salgueiro Pardo. 2018. A rule-based AMR parser for Portuguese. In *Ibero-American Conference on Artificial Intelligence*, pages 341–353. Springer.

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.

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. 2019. Core semantic first: A top-down approach for AMR parsing. 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 3799–3809, Hong Kong, China. Association for Computational Linguistics.

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

Yitao Cai, Zhe Lin, and Xiaojun Wan. 2021. Making better use of bilingual information for cross-lingual AMR parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1537–1547, Online. Association for Computational Linguistics.

Marco Damonte. 2019. *Understanding and Generating Language with Abstract Meaning Representation*. Ph.D. thesis, University of Edinburgh.

Marco Damonte and Shay Cohen. 2020. Abstract Meaning Representation 2.0 - Four Translations. Technical Report LDC2020T07, Linguistic Data Consortium, Philadelphia, PA.

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, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for Abstract Meaning Representation. In *Proceedings of EACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Michael Wayne Goodman. 2019. AMR normalization for fairer evaluation. *arXiv preprint arXiv:1909.01568*.

Kevin Knight, Bianca Badarau, Laura Banarescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2017. Abstract Meaning Representation (AMR) Annotation Release 2.0. Technical Report LDC2017T10, Linguistic Data Consortium, Philadelphia, PA.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating *the Little Prince* with Chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15, Berlin, Germany. Association for Computational Linguistics.

Tahira Naseem, Austin Blodgett, Sadhana Kumaravel, Tim O'Gorman, Young-Suk Lee, Jeffrey Flanigan, Radu Fernandez Astudillo, Ramón Florian, Salim Roukos, and Nathan Schneider. 2022. DocAMR: Multi-sentence AMR representation and evaluation. In *Proc. of NAACL-HLT*, Seattle, Washington.

Juri Opitz, Angel Daza, and Anette Frank. 2021. Weisfeiler-Leman in the BAMBOO: Novel AMR graph metrics and a benchmark for AMR graph similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.

Juri Opitz and Anette Frank. 2021. Towards a decomposable metric for explainable evaluation of text generation from AMR. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1504–1518, Online. Association for Computational Linguistics.

Juri Opitz, Letitia Parcalabescu, and Anette Frank. 2020. AMR similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8:522–538.

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.

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

Adylan Roaffa Ilmy and Masayu Leylia Khodra. 2021. Parsing Indonesian sentence into Abstract Meaning Representation using machine learning approach. *arXiv e-prints*, pages arXiv–2103.

Linfeng Song and Daniel Gildea. 2019. SemBleu: A robust metric for AMR parsing evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4552, 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.

Chuan Wang, Bin Li, and Nianwen Xue. 2018. Transition-based Chinese AMR parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 247–252, New Orleans, Louisiana. Association for Computational Linguistics.

Shira Wein, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, and Nathan Schneider. 2022. Spanish Abstract Meaning Representation: Annotation of a general corpus. *arXiv preprint arXiv:2204.07663*.

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.

Shira Wein and Nathan Schneider. 2022. Effect of source language on AMR structure. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW)*, Marseille, France.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2021. XLPT-AMR: Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 896–907.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, 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 *Proc. of ICLR*, Online.