Effect of Source Language on AMR Structure

Shira Wein, Wai Ching Leung, Yifu Mu, Nathan Schneider
Georgetown University, USA
{sw1158, wl607, ym431, nathan.schneider}@georgetown.edu

Abstract

The Abstract Meaning Representation (AMR) annotation schema was originally designed for English. But the formalism has since been adapted for annotation in a variety of languages. Meanwhile, cross-lingual parsers have been developed to derive English AMR representations for sentences from other languages—implicitly assuming that English AMR can approximate an interlingua. In this work, we investigate the similarity of AMR annotations in parallel data and how much the language matters in terms of the graph structure. We set out to quantify the effect of sentence language on the structure of the parsed AMR. As a case study, we take parallel AMR annotations from Mandarin Chinese and English AMRs, and replace all Chinese concepts with equivalent English tokens. We then compare the two graphs via the Smatch metric as a measure of structural similarity. We find that source language has a dramatic impact on AMR structure, with Smatch scores below 50% between English and Chinese graphs in our sample—an important reference point for interpreting Smatch scores in *cross-lingual* AMR parsing.

Keywords: Abstract Meaning Representation (AMR), Chinese language resources, annotation

1 Introduction

Though the Abstract Meaning Representation (AMR; Banarescu et al., 2013) framework was originally designed for annotating English sentences, and not intended as an interlingua, it has since been adapted to a number of other languages (§2.1), raising the question of how well it abstracts away from the particularities of individual languages. To investigate AMR's ability to serve as an interlingua, previous work has explored methods of characterizing the types of differences between parallel AMR graphs (AMRs annotating parallel sentences in different languages; §2.2). However, there has not yet been an effort to *systematically quantify* the effect on AMR structure of the language of the sentence being parsed (hereafter, the *source language*). We hypothesize that regardless of any language-specific information in the AMR (i.e. if the labels are made to be in the same language), the structure of AMRs across language pairs will likely differ because of the linguistic properties of the source sentence. To better understand the impact of language on AMR structure in the pursuit of effective evaluation of cross-lingual AMR pair similarity, we aim to quantify the amount of impact in parallel AMRs.

Here we explore the effect of source language on AMR structure in the large annotated parallel corpus of Mandarin Chinese and English AMRs (Li et al., 2016). To quantify the impact of source language on the AMR, we eliminate the measurable impact of lexical divergence and focus solely on structural divergences. To do this, we take a pair of parallel English and Chinese AMRs and manually translate every word in the Chinese graph into its English equivalent. Structural elements of the AMR are largely unchanged (§3.2). We then evaluate via Smatch (Cai and Knight, 2013), which is an algorithm to compare AMR graphs and calculate similarity. Ultimately, we have a Smatch score quantifying the effect of source language on AMR structure.

From these Smatch scores, we are able to demonstrate that the source language has a dramatic effect on the structure of an AMR, even if the AMR is a gold annotation with no noise introduced by automatic parsing. This result has important implications for (1) identifying cross-linguistic inconsistencies in the AMR schema, and (2) interpreting scores in cross-lingual AMR parsing evaluations (Damonte, 2019).[1]

Our primary contributions include:

- a novel approach to quantifying effect of source language on AMR structure;
- a small dataset of 120 Chinese AMRs with English concept labels, following our approach;[2] and
- an analysis of the Smatch score differences between our Chinese AMRs with English concept labels and the corresponding gold English AMRs.

2 Related Work

2.1. Abstract Meaning Representation

The Abstract Meaning Representation (AMR) formalism is a graph-based representation of the meaning of a sentence or phrase. In AMR annotations, nodes reflect entities and events, and the edges are labeled with semantic roles. AMR aims to abstract away from surface details of morphology and syntax in favor of core elements of meaning, such as predicate-argument structure and coreference. With that in mind, sentences with the same meaning (and content word vocabulary) should be represented by the same AMR. English AMR annotations are unanchored—the nodes are not explicitly

[1] "Cross-lingual AMR parsing" typically refers to parsing a sentence from a language other than English into a standard English AMR.

[2] Our annotations can be found at https://github.com/shirawein/effect-language-amr-structure

mapped to tokens in the sentence—but the concepts (semantic node labels) largely consist of lemmatized words from the sentence.

AMR was designed exclusively for English and was not intended to be an interlingua (Banarescu et al., 2013), but has now been extended to multiple languages. AMR has been adapted to Chinese (Li et al., 2016), Portuguese (Anchiêta and Pardo, 2018; Sobrevilla Cabezudo and Pardo, 2019), Spanish (Migueles-Abraira et al., 2018; Wein et al., 2022), Vietnamese (Linh and Nguyen, 2019), Turkish (Azin and Eryiğit, 2019; Oral et al., 2022), Korean (Choe et al., 2020), and Persian (Takhshid et al., 2022).

A multilingual adaptation of AMR, the Uniform Meaning Representation (Van Gysel et al., 2021), was developed to incorporate linguistic diversity into the AMR annotation process.

2.2. Differences in Cross-lingual AMR Pairs

AMR has been assessed as an interlingua, considering the types of differences which appear across AMR language pairs, for Czech (Hajič et al., 2014), Chinese (Xue et al., 2014), and Spanish (Wein and Schneider, 2021), in comparison to English.

Xue et al. (2014) explore the adaptability of English AMR to Czech and Chinese. They suggest that AMR may be cross-linguistically adaptable because it abstracts away from morpho-syntactic differences. Cross-linguistic comparisons between English/Czech and English/Chinese AMR pairs indicate that most pairs align well. Also, the compatibility is higher for English and Chinese than for English and Czech.

Hajič et al. (2014) describe the types of differences between AMRs for parallel English and Czech sentences, and find that the differences may be either due to convention/surface-level nuances which could be changed in the annotation guidelines, or may be due to inherent facets of the AMR annotation schema. One notable cross-lingual AMR difference is from the appearance of language-specific idioms and phrases.

Wein and Schneider (2021) define the types and causes of divergences between cross-lingual AMR pairs for English-Spanish parallel sentences. The causes of structural differences between parallel AMRs are identified as being due to semantic divergences, syntactic divergences, or annotation choices.

Though previous work has explored methods of characterizing the differences between pairs of cross-lingual AMRs, in this work, we aim to quantify the impact of the source language on AMR structure.

3 Annotation

3.1. Dataset

For our annotation and analysis, we make use of parallel gold Chinese and English AMR annotations of the novel *The Little Prince*—the Chinese AMRs from the CAMR dataset (Li et al., 2016)[3] and their parallel English AMR annotations (Banarescu et al., 2013).[4] We were interested in using this set of parallel data because of the notable divergence in linguistic properties between Chinese and English, as well as the prominence of Chinese sentence–to–English AMR parsing (Damonte and Cohen, 2020). The 100 AMRs used are the first 100 annotations of both development sets, corresponding to the first 100 sentences of *The Little Prince*.[5] The average sentence length is 15.3 tokens for the 100 English sentences and 19.5 tokens for the 100 Chinese sentences. Since the Chinese AMRs do not include `:wiki` tags, we remove all `:wiki` tags from the gold English AMRs.

Note that *The Little Prince* was originally written in French, so both the English and Chinese versions are translations and may exhibit features of translationese and/or may be subject to differences due to French serving as a third pivot language (Koppel and Ordan, 2011).

3.2. Approach

Our broad approach to annotation consists of taking the CAMR annotation and replacing the Chinese concepts with English tokens. We want to replace the Chinese concepts with English tokens so that we do not penalize lexical differences (which are apparent as the words are originally in different languages), but rather, exclusively measure the structural differences between the AMRs. Specifically, this consists of a three-step process:

1. Manually translate the Chinese concepts to equivalent English tokens.
2. Check the parallel gold English AMR to identify synonyms of the manually generated translations of the Chinese concepts.
3. If a synonym (close enough in meaning such that faithfulness to the Chinese sentence is not lost) of the manually generated translation appears in the gold English AMR, the term from the English AMR is used to replace the manually generated translation. Otherwise, the manually generated translation is used.

Additionally, there are some terms that appear in the CAMR annotations which would not appear in English AMR annotations. For example, functional particles such as 就 (a central particle with a multitude of uses) appear in the CAMR annotation schema but prepositions and other morphosyntactic details do not appear in the English AMR annotation schema. We remove these functional particles from the Chinese annotations rather than attempt to translate them into English. No other structural changes are made to the Chinese AMR.

We trained two linguistics students bilingual in English and Chinese in our approach. Approximately 4 hours were spent per annotator to produce the annotations and no annotation tool was used.

[3] `https://www.cs.brandeis.edu/~clp/camr/res/blj_dev.txt`

[4] `https://amr.isi.edu/download/amr-bank-struct-v1.6-dev.txt`

[5] 20 sentences were double-annotated: see §4.

4 Results & Analysis

We collect 60 annotations from each annotator, with 20 sentences overlapping so that we can calculate inter-annotator agreement (120 annotations total, on 100 unique sentences). We calculate the Smatch scores between the annotations (Chinese AMR with English concepts) and the corresponding gold English AMR.

4.1. Inter-Annotator Agreement

English translation: Nothing about him gave any suggestion of a child lost in the middle of the desert, a thousand miles from any human habitation.
Annotation 1:

```
(x0 / look-02
   :polarity  (x2 / -)
   :degree  (x3 / slightest)
   :arg0  (x4 / he)
   :arg1  (x5 / child
      :quant  (x6 / 1)
      :arg0-of  (x7 / lose-02
         :location  (x8 / desert
            :mod  (x9 / large)
            :mod  (x10 / uninhabited)))))
```

Annotation 2:

```
(x0 / seem-01
   :polarity  (x2 / -)
   :degree  (x3 / remote)
   :arg0  (x4 / he)
   :arg1  (x5 / child
      :quant  (x6 / 1)
      :arg0-of  (x7 / lose-02
         :location  (x8 / desert
            :mod  (x9 / huge)
            :mod  (x10 / uninhabited)))))
```

Figure 1: Both annotations (from Annotator 1 and Annotator 2) for one of the sentences in our dataset. Note that the annotators provided the English concepts and the structure of the annotation is derived from the parallel Chinese annotation.

We find that the average inter-annotator agreement (calculated by Smatch) is 0.8645, on a scale from 0 to 1, with 1 being exactly the same. Inter-annotator agreement here measures lexical agreement between the translators. The reason IAA would not be 1 is because translation choices are being made when producing the annotations. For example, in figure 1, one annotator felt that a more faithful translation of 像 is `seem`, while the other annotator decided that a more accurate translation would be `look`. The same is true for the difference between `slightest` and `remote`, as well as between `huge` and `large`. None of those terms (either item of any of the three pairs) are captured in the parallel gold English AMR, so these differences reflect translation choices and not errors in annotation. This pair of annotations received an IAA score of 0.85.

4.2. Annotations versus Gold English AMRs

English sentence: “It has horns.”
Gold English annotation:

```
(h / have-03
   :arg0 (i / it)
   :arg1 (h2 / horn))
```

Chinese sentence: “ 还有犄角呢。”
Annotation (Chinese AMR with English concept labels):

```
(x0 / say
   :arg1  (x2 / have-03
      :manner  (x3 / even)
      :arg1  (x4 / horn)))
```

Figure 2: Gold English AMR and our annotation for parallel sentences.

English sentence: “Boa constrictors swallow their prey whole , without chewing it.
Gold English annotation:

```
(s2 / say-01
   :arg0 (b2 / book)
   :arg1 (s / swallow-01
      :arg0 (b / boa)
      :arg1 (p / prey
         :mod (w / whole)
         :poss b)
      :manner (c2 / chew-01 :polarity -
         :arg0 b
         :arg1 p)))
```

Chinese sentence: 这本书中写道：“这些蟒蛇把它们的猎获物不加咀嚼地囫囵吞下
Annotation (Chinese AMR with English concept labels):

```
(x11 / writes-01
   :arg0  (x13 / book-01)
   :arg1  (x14 / swallow-01
      :arg0  (x15 / boa
         :mod  (x16 / these))
      :arg1  (x17 / prey
         :poss  (x25 / x15))
      :manner  (x19 / whole)
      :manner  (x21 / chew-01 :polarity -))))
```

Figure 3: Gold English AMR and our annotation for parallel sentences (some roles removed for brevity of presentation).

The production of our annotations is motivated by the ability to then quantify the amount of difference between our annotations and the gold English AMRs. We

English sentence: And after some work with a colored pencil I succeeded in making my first drawing.
Chinese sentence: 于是，我也用彩色铅笔画出了我的第一副图画。
Literal English translation of Chinese sentence: So, I also drew my first drawing with colored pencils.

Figure 4: An English and Chinese sentence pair from the dataset, displaying slight variation in the translation.

use Smatch to quantify this difference as the standard similarity evaluation technique for AMR pairs.

The Smatch score for the gold English AMRs in comparison to the annotations is 41% for those produced by Annotator 1 and 44% for those produced by Annotator 2. These Smatch scores are over 60 sentence pairs each. This indicates that there is a sizable effect of source language on the structure of the AMR even with the Chinese labels being replaced, raising questions for how we evaluate cross-lingual AMR parsers.

We expect that some of the differences we capture in our approach are due to translation, and some differences are due to syntactic and semantic properties, as established by previous work comparing more similar languages (Spanish and English) (Wein and Schneider, 2021). One example of a syntactic effect on AMR structure can be seen in figure 2.

This divergence arises out of the ability in Chinese to omit sentence subjects when they can be understood from context, which explains why the Chinese graph is missing an `:arg0` argument. It is likely that there are differences in meaning in parallel sentences as caused by the translation process, though there are also observed syntactic differences as noted in the example in figure 2.

A more subtle effect of source language on AMR structure can be seen in figure 3 relating to the `:arg1 prey`. In English, we have “swallow their prey whole,” such that “whole” is a semantic modifier of “prey,” denoted by `:mod`. In Chinese, the equivalent is 囫囵 (wholly, possibly barbarically) 吞下 (swallow). Wholly (囫囵) is annotated as `:manner` to the swallowing (吞下), instead of as the `:mod` of `prey`. We consider this a faithful and standard translation reflective of cross-linguistic differences between the “swallow whole” construction in English and the “wholly swallow” construction in Chinese. This difference is reflected in the AMR.

One example of sentences being slight variants of each other rather than literal translations is the sentence pair seen in figure 4. The annotation (same for both annotators) received a Smatch score of 0.43 similarity with the gold English AMR. The majority of the sentences are closely parallel, so we expect that the difference we are quantifying is an effect of syntactic and semantic divergence between Chinese and English.[6]

4.3. Accounting for Design Differences

A few relatively superficial differences in annotation guidelines between Chinese and English need to be accounted for, as they may impact the Smatch score without being a direct reflection of source language impact. We found four types of differences which have an impact on AMR structure:

- CAMR uses the concept `mean` for elaboration/further explanation of another concept/structure, which is often included in parentheses/colon (present in 3 AMR pairs)
- CAMR uses the concept `cause` instead of `cause-01` to refer to the cause of an event, which is considered a non-core role (in 4 AMR pairs)
- CAMR occasionally uses `:beneficiary` instead of `:arg2` to refer to indirect object (in 5 AMR pairs)
- While English AMR does not account for the sentence being a quotation, CAMR roots all quotations with `say` (in 13 AMR pairs)[7]

Removed Diff.	Anno.1/Gold	Anno.2/Gold
None	41%	44%
Mean	43%	44%
Cause	41%	44%
Beneficiary	41%	43%
Quotation	41%	43%
All	42%	45%

Table 1: Smatch scores without each of the four design differences.

As can be seen in table 1, even when removing all AMR pairs noticeably affected by schema differences, the Smatch score similarity between our annotations and the gold English AMRs only increases incrementally, and a large effect of source language remains. This indicates that the dissimilarity we measure in AMR structure is not due to differences in annotation schema.

5 Conclusion

Our case study between Chinese and English serves as an analysis of the impact of linguistic divergence between those two languages on AMR structure. Through our annotation process of translating Chinese concepts to English, we find that there is a dramatic impact on AMR structures, with Smatch scores between our annotations and the gold English AMRs falling below 50%. For comparison, inter-annotator Smatch scores within a single language (Chinese) in the same domain have been reported at 83% (Li et al., 2016).

This substantive impact on AMR structure motivates further consideration for source language when working with AMR cross-lingually—either in evaluating cross-lingual AMR parsers or when developing and comparing AMR schema in new languages.

[6] If Chinese and English gold AMRs are released in different domains in future work, it would be interesting to repeat this analysis on those texts and compare our findings.

[7] In English AMR, only the first sentence in the quotation, starting with open quotes, is rooted with `say`. In Chinese AMR, any sentence containing quotes is rooted with `say`.

As a meaning representation, it is critical that an AMR graph effectively reflect the meaning of the sentence being parsed. Current cross-lingual AMR parsers evaluate accuracy of a parsed non-English sentence by comparing to the corresponding gold English AMR. Our newfound evidence that source language has a sizable effect on AMR structure should be taken into account when interpreting cross-lingual Smatch evaluations. Ideally, gold AMRs should be created in the source language for evaluating cross-lingual parsers (even if sufficient training data is only available in English). Future work might investigate steps to mitigate source language impact when evaluating cross-lingual AMR parsing, or further investigate the effect in other language pairs.

Acknowledgments

We thank anonymous reviewers for their feedback. This work is supported by a Clare Boothe Luce Scholarship.

6 Bibliographical References

References

Anchiêta, Rafael and Pardo, Thiago (2018). Towards AMR-BR: A SemBank for Brazilian Portuguese language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

Azin, Zahra and Eryiğit, Gülşen (2019). Towards Turkish Abstract Meaning Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 43–47. Association for Computational Linguistics, Florence, Italy. doi: 10.18653/v1/P19-2006.

Banarescu, Laura, Bonial, Claire, Cai, Shu, Georgescu, Madalina, Griffitt, Kira, Hermjakob, Ulf, Knight, Kevin, Koehn, Philipp, Palmer, Martha, and Schneider, Nathan (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186. Association for Computational Linguistics, Sofia, Bulgaria.

Cai, Shu and Knight, Kevin (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. Association for Computational Linguistics, Sofia, Bulgaria.

Choe, Hyonsu, Han, Jiyoon, Park, Hyejin, Oh, Tae Hwan, and Kim, Hansaem (2020). Building Korean Abstract Meaning Representation corpus. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 21–29. Association for Computational Linguistics, Barcelona Spain (online).

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

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

Hajič, Jan, Bojar, Ondřej, and Urešová, Zdeňka (2014). Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 55–64. Association for Computational Linguistics and Dublin City University, Dublin, Ireland. doi:10.3115/v1/W14-5808.

Koppel, Moshe and Ordan, Noam (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326. Association for Computational Linguistics, Portland, Oregon, USA.

Li, Bin, Wen, Yuan, Qu, Weiguang, Bu, Lijun, and Xue, Nianwen (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. Association for Computational Linguistics, Berlin, Germany. doi: 10.18653/v1/W16-1702.

Linh, Ha and Nguyen, Huyen (2019). A case study on meaning representation for Vietnamese. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 148–153. Association for Computational Linguistics, Florence, Italy. doi:10.18653/v1/W19-3317.

Migueles-Abraira, Noelia, Agerri, Rodrigo, and Diaz de Ilarraza, Arantza (2018). Annotating Abstract Meaning Representations for Spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

Oral, Elif, Acar, Ali, and Eryiğit, Gülşen (2022). Abstract meaning representation of Turkish. *Natural Language Engineering*, pages 1–30. doi:10.1017/S1351324922000183.

Sobrevilla Cabezudo, Marco Antonio and Pardo, Thiago (2019). Towards a general Abstract Meaning Representation corpus for Brazilian Portuguese. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 236–244. Association for Computational Linguistics, Florence, Italy. doi:10.18653/v1/W19-4028.

Takhshid, Reza, Shojaei, Razieh, Azin, Zahra, and Bahrani, Mohammad (2022). Persian Abstract Meaning Representation. *arXiv preprint arXiv:2205.07712*.

Van Gysel, Jens E. L., Vigus, Meagan, Chun, Jayeol, Lai, Kenneth, Moeller, Sarah, Yao, Jiarui, O'Gorman, Tim, Cowell, Andrew, Croft, William, Huang, Chu-Ren, Hajič, Jan, Martin, James H., Oepen, Stephan, Palmer, Martha, Pustejovsky, James, Vallejos, Rosa, and Xue, Nianwen (2021). Designing a Uniform Meaning Representation for natural language processing. *KI - Künstliche Intelligenz.*

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

Wein, Shira and Schneider, Nathan (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. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Xue, Nianwen, Bojar, Ondřej, Hajič, Jan, Palmer, Martha, Urešová, Zdeňka, and Zhang, Xiuhong (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. European Language Resources Association (ELRA), Reykjavik, Iceland.