Measuring Fine-Grained Semantic Equivalence with Abstract Meaning Representation

Shira Wein
Georgetown University
sw1158@georgetown.edu

Zhuxin Wang
Georgetown University
zw85@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

Identifying semantically equivalent sentences is important for many NLP tasks. Current approaches to semantic equivalence take a loose, sentence-level approach to "equivalence," despite evidence that fine-grained differences and implicit content have an effect on human understanding and system performance. In this work, we introduce a novel, more sensitive method of characterizing cross-lingual semantic equivalence that leverages Abstract Meaning Representation graph structures. We find that parsing sentences into AMRs and comparing the AMR graphs enables finer-grained equivalence measurement than comparing the sentences themselves. We demonstrate that when using gold or even automatically parsed AMR annotations, our solution is finer-grained than existing corpus filtering methods and more accurate at predicting strictly equivalent sentences than existing semantic similarity metrics.

1 Introduction

Translation between two languages is not always completely meaning-preserving, and information can be captured by one sentence which is not captured by the other. Semantic divergence (or conversely, semantic equivalence) detection aims to pick out parallel texts which have less than equivalent meaning. Though semantic divergence across sentences in parallel corpora has been well-studied, current detection methods fail to capture the full scope of semantic divergence. State-of-the-art semantic divergence systems rely on perceived *sentence-level divergences*, which do not entirely encapsulate all semantic divergences.

For example, consider the parallel French and English sentences from the REFreSD dataset (Briakou and Carpuat, 2020) shown in Figure 1. The French sentence says "tous les autres édifices" (*all other buildings*) while the English specifies "all other *religious* buildings." Because the sentence goes on to list religious buildings, it could be inferred from context that the French is describing other *religious* buildings despite being omitted; the sentences thus convey the same overall meaning but are not *exactly* parallel. Under a strict or close analysis of the translation, these sentences could be considered divergent—because the meanings are not identical—but at the sentence-level they are essentially equivalent.

All other *religious* buildings are mosques or Koranic schools founded after the abandonment of Old Ksar in 1957.

Tous les autres édifices sont des mosquées ou des écoles coraniques fondées à l'époque postérieure à l'abondance du vieux ksar en 1957.

Figure 1: Two parallel sentences from the REFreSD dataset marked as having no meaning divergence, for which the AMRs diverge.

Fine-grained semantic equivalence detection is not widely studied—in spite of evidence that: (1) implicit information can be critical to the understanding of the sentence (Roth and Anthonio, 2021), (2) fine-grained divergences in parallel training data have a negative effect on neural machine translation system performance (Briakou and Carpuat, 2021), and finally, that (3) fine-grained semantic equivalence detection holds promise for a number of applications. Most notably, translation studies, semantic analyses, and language learning contexts could all benefit from the distinction between semantically equivalent sentence pairs and sentence pairs which have subtle or implicit differences (Bassnett, 2013). A fine-grained divergence detection system would enable the probing of machine translation models for semantic equivalence (Mallinson et al., 2017) and could point to areas where the source language itself affects semantics in parallel sentences (Taguchi, 2005). Other potential uses include: reducing the workload of human

Proceedings of the 15th International Conference on Computational Semantics pages 144–154
June 21–23, 2023. ©2023 Association for Computational Linguistics

translators in post-editing of machine translation output by filtering out exactly semantically equivalent sentence pairs (Green et al., 2013) and cross-lingual text reuse detection (plagiarism detection) (Potthast et al., 2011).

Given the wide-ranging motivation for the development of a fine-grained equivalence detection system, coupled with the notable gap in research on this task, we argue that a finer-grained measure of semantic equivalence is needed: a way to detect *strictly* semantically equivalent sentence pairs. We leverage explicit semantic information in the form of Abstract Meaning Representation (AMR; Banarescu et al., 2013) to fill this gap. In this work, we demonstrate that parsing sentences into AMR graphs and comparing those graphs enables a finer-grained semantic comparison than simply comparing the sentences. We suspect that AMR may be useful in this case because it makes explicit every concept and relationship between those concepts present in the sentence, taxonomically categorizing each concept's role and argument.

With analysis of data in two language pairs (English-French and English-Spanish), we demonstrate that sentence-level divergence annotations can be coarse-grained, neglecting slight differences in meaning (§3). We find that comparing two AMR graphs is an effective way to characterize meaning in order to uncover finer-grained divergences (§4), and this can be achieved even with automatic AMR parsers (§5). Finally, in §6 we evaluate our AMR-based metric on a cross-linguistic semantic textual similarity dataset, and show that for detecting semantic equivalence, it is more precise than a popular existing model, multilingual BERTScore (Zhang et al., 2020).

Our primary contributions include:

- Our novel approach to the identification of semantic divergence which uses AMR to move beyond perceived sentence-level divergences
- A simple pipeline algorithm (which modifies Smatch (Cai and Knight, 2013)) to automate the detection of AMR-level divergence in cross-lingual pairs
- Studies demonstrating that our AMR-based approach accurately captures a finer-grained degree of semantic equivalence than both the state-of-the-art corpus filtering method and a semantic textual metric

We will release the code and dataset for this work upon publication to enable the use of AMR for semantic divergence detection.

2 Background on Semantic Divergence

Semantic divergences can arise when translating from one language to another. These divergences can arise due to different language structure, syntactic differences in the language, or translation choices (Dorr, 1994, 1990). Additional divergences can be introduced when automatically extracting and aligning parallel resources (Smith et al., 2010; Zhai et al., 2018; Fung and Cheung, 2004).

To address these divergences, a number of systems have been developed to automatically identify divergences in parallel texts (Carpuat et al., 2017; Vyas et al., 2018; Briakou and Carpuat, 2020, 2021; Zhai et al., 2020). The approach taken by Briakou and Carpuat (2020) to detecting sentence-level semantic divergences involves training multilingual BERT (Devlin et al., 2018) to rank sentences diverging to various degrees. They introduced a novel dataset called Rational English-French Semantic Divergences (REFreSD). REFreSD is a subset of the French-English WikiMatrix (Schwenk et al., 2021) with crowdsourced annotations classifying the sentences as having no meaning divergence, some meaning divergence, or being unrelated.

Recent work has investigated the differences in cross-lingual (English-Spanish) AMR pairs within the framework of translation divergences (Wein and Schneider, 2021). Specifically, this work developed an annotation schema to classify the types and causes of differences between cross-lingual AMR pairs. We use this dataset to test the performance of our system on English-Spanish gold AMR pairs. (For English-French, we produce our own gold judgments of AMR divergence to test our algorithm.) Additional prior work has explored the role of structural divergences in cross-lingual AMR parsing (Blloshmi et al., 2020; Damonte, 2019).

The relationship between Abstract Meaning Representation metrics and measures of semantic similarity has been explored in (Leung et al., 2022). Recent work has also integrated sentence-level embeddings and comparison of AMR graphs (Opitz et al., 2021; Wein and Schneider, 2022; Zeidler et al., 2022).

3 AMR for Identification of Semantic Equivalence

Semantic representations are designed to capture and formalize the meaning of a sentence. In partic-

He later scouted in Europe for the Montreal Canadiens.

```
(s / scout-02
      :ARG0 (h / he)
      :ARG1 (c / continent
            :wiki "Europe"
            :name "Europe")
      :ARG2 (c2 / canadiens
            :mod "Montreal")
      :time (a / after))
```

Il a plus tard été dépisteur du Canadiens de Montréal en Europe. (*He later scouted for the Montreal Canadiens in Europe.*)

```
(d / dépister-02
      :ARG0 (i / il)
      :ARG1 (c / continent
            :wiki "Europe"
            :name "Europe")
      :ARG2 (c2 / canadiens
            :mod "Montreal")
      :time (p / plus-tard))
```

Figure 2: A pair of sentences and their human annotated AMRs, for which the sentences receive a "no meaning divergence" judgment in the REFreSD dataset, and are also equivalent per AMR divergence.

ular, the Abstract Meaning Representation (AMR) framework aims to formalize sentence meaning as a graph in a way that is conducive to broad-coverage manual annotation (Banarescu et al., 2013, 2019). These semantic graphs are rooted and labeled, such that each node of the graph corresponds to a semantic unit. AMR does not capture nominal or verbal morphology or many function words, abstracting away from the syntactic features of the sentence.

We leverage the semantic information captured by AMR to recognize semantic equivalence or divergence across parallel sentences. Figure 2, for example, illustrates a strictly meaning-equivalent sentence pair along with the AMRs. Though the sentences differ with respect to syntax and lexicalization, the AMR graphs are structurally isomorphic. If the AMR structures were to differ, that would signal a difference in meaning.

Two particularly beneficial features of the AMR framework are the rooted structure of each graph, which elucidates the semantic focus of the sentence, as well as the concrete set of specific non-core roles, which are useful in classifying the specific relation between concepts/semantic units in the sentence. For example, in Figure 3, the emphasis on the English sentence is on possession—*your* planet—but the emphasis on the Spanish sentence is on place of origin, asking, which planet are you *from?* This difference in meaning is reflected in the diverging roots of the AMRs.

Which is your planet?

```
(p / planet
    :poss (y / you)
    :domain (a / amr-unknown))
```

¿ De qué planeta eres ? (*Which planet are you from?*)

```
(s / ser-de-91
   :ARG1 (t / tú)
   :ARG2 (p / planeta
       :domain (a / amr-desconocido)))
```

Figure 3: Two parallel sentences and AMRs from the Migueles-Abraira et al. English-Spanish AMR dataset, which diverge in meaning. The Spanish role labels are translated into English here for ease of comparison.

Finally, we identify the fact that non-core roles (such as `:manner`, `:degree`, and `:time`) are particularly helpful in identifying parallelism or lack of parallelism between the sentences. This is because AMR abstracts away from the syntax (so that word order and part of speech choices do not affect equivalence), but instead explicitly codes relationships between concepts via semantic roles. Furthermore, AMRs use special frames for certain relations, such as `have-rel-role-91` and `include-91`, which can be useful in enforcing parallelism when the meaning is the same but the specific token is not the same. For example, if the English and French both have a concession which the English marks via "although" and the French marks with "*mais*" (*but*), the AMR special frame role will still preserve parallelism by indicating them both as a concession.

Granularity of the REFreSD dataset. Sentence-level divergences (as annotated in REFreSD) do not capture all meaning differences. Another example of this surface-level divergence adjudication, using sentences from the REFreSD dataset, is shown in Figure 4. These sentences are marked as having no meaning divergence in the REFreSD dataset but do have diverging AMR pairs. The difference highlighted by the AMR pairs is the `:time` role of reach / *atteindre*. The English sentence says that no. 1 is reached "within a few weeks" of the release, while the French sentence says that no. 1 is reached the first week of the release (*la première semaine*).

We explore the ability to discover semantic divergences in sentences either with gold parallel AMR annotations or with automatically parsed AMRs using a multilingual AMR parser, in order to enable the use of this approach on large corpora (considering that AMR annotation requires training).

We propose that an approach to detecting di-

Although the sales were slow (admittedly, according to the band), the second single from the album, "Sweetest Surprise" reached No. 1 in Thailand *within a few weeks* of release.

Même si les exemplaires ont du mal à partir (comme l'admet le groupe), le second single de l'album, Sweetest Surprise, atteint la première place en Thaïlande *la première semaine* de sa sortie.

Figure 4: Two parallel sentences from the REFreSD dataset (Briakou and Carpuat, 2020) marked as having no meaning divergence, but for which the AMRs diverge. Italicized spans indicate the cause of the AMR divergence.

vergences using AMR will be a stricter, finer-grained measurement of semantic divergence than perceived sentence-level judgments.

4 Examining and Automatically Detecting Differences in Gold AMRs

In this section, we **evaluate the ability of AMR to expose fine-grained differences in parallel sentences** and how to **automatically detect those differences**. In order to do so, we produce and examine English-French AMR pairs, which is the first annotated dataset of French AMRs; we also examine a number of English-Spanish AMR pairs.

This is a relatively small dataset (100 English-French items and 50 English-Spanish items) because it serves as a manually annotated precursor to validate our hypothesis, ahead of our extensive automatically-produced AMR experimentation (§5) which uses 1033 items.

4.1 Examination of Gold AMR Data

We focus on French for effective comparison with sentence-level semantic divergence models (because of the available resources), though it also makes for ideal candidates in a cross-lingual AMR comparison, as it is broadly syntactically similar to English. This suggests that the AMRs could be expected to look similar (though not exactly the same) as inflectional morphology and function words are not represented in AMR. Prior work has investigated the transferability of AMR to languages other than English, and has found that it is not exactly an interlingua, but in some cases cross-lingual AMRs align well. Additionally, some languages are more compatible (Chinese) with English AMR than other languages (Czech) (Xue et al., 2014).

English-French AMR Parallel Corpus In investigating the differences between the degree of divergence captured by AMR and sentence-level divergence, we aim to compare quantitative measures of AMR similarity with corresponding sentence-level judgments of similarity. In order to compare human judgments and AMR judgments, we develop the first French-English AMR parallel corpus, which represents the first application of AMR to French. We produce gold AMR annotations for 100 sentences, which were randomly sampled, from the REFreSD dataset (Briakou and Carpuat, 2020; Linh and Nguyen, 2019). We also test our system on the full REFreSD dataset, using an automatic AMR parser (described in §5).

For the French AMR annotation process, the role/argument labels were added in English as has been done in related non-English AMR corpora (Sobrevilla Cabezudo and Pardo, 2019), and the concept (node) labels were in French. The specific concept sense numbers were based on English PropBank frames (Kingsbury and Palmer, 2002; Palmer et al., 2005).

	AMR Div.	AMR Equi.
Sentence-Level Div.	57	0
Sentence-Level Equi.	26	17

Table 1: Comparison between AMR Divergence annotations and Sentence-Level Divergence REFreSD annotations for 100 French-English sentences.

Findings from Corpus Annotation In light of our research question considering whether AMR can serve as a proxy of fine-grained semantic divergence, we consider both qualitative and quantitative evidence. While producing this small corpus of French-English parallel AMRs, our suspicions that AMR would be able to more fully capture semantic divergence than perceived sentence-level divergence were confirmed. We uncovered a number of ways in which perceived sentence-level equivalence is challenged by the notion of AMR divergence. Take the example in Figure 1. The difference between "religious" being applied in the French sentence and appearing in the English sentence is not captured by perceived sentence-level divergence, but is captured by AMR divergence.

The results in Table 1 demonstrate that when using AMR as a lens to filter meaning, the result is always stricter than when simply comparing their corresponding sentences in the form of human judgment. There are no instances where the sentence-level annotation claims that the sentences are divergent but the AMR annotations are equivalent. Conversely, there are 26 instances with AMR divergence but no perceived sentence-level semantic divergence. From this annotation we find that AMR divergence is a finer-grained measure of divergence than perceived sentence-level divergence.

4.2 Quantifying Divergence in Cross-Lingual AMR Pairs

We have shown that not all pairs that humans considered equivalent at the sentence level receive isomorphic AMRs because they actually contain low-level semantic divergences. This suggests AMRs can be useful for more sensitive automatic detection of divergence. Now, we investigate whether we can automatically detect and quantify this divergence on gold AMRs via the graph comparison algorithm Smatch. In order to quantify this divergence in cross-lingual AMR pairs, we develop a simple pipeline algorithm which is a modified version of Smatch and incorporates token alignment. We test our modified Smatch algorithm on gold English-French AMR pairs and gold English-Spanish AMR pairs in comparison to the similarity scores output by Briakou and Carpuat (2020).

Modified cross-lingual version of Smatch. Our simple pipeline algorithm extends Smatch, a measurement of similarity between two (English) AMRs (Cai and Knight, 2013). Smatch quantifies the similarity of two AMRs by searching for an alignment of nodes between them that maximizes the F_1-score of matching (*node1*, *role*, *node2*) and (*node1*, `instance-of`, *concept*) triples common between the graphs. However, Smatch was designed to compare AMRs in the same language, with the same role and concept vocabularies.

To compare AMR nodes across languages, the nodes first need to be cross-lingually aligned. This involves translating the concept and role labels. We take a simple approach of first word-aligning the sentence pair to ascertain corresponding concepts (most of which are lemmas of content words in the sentence). Our approach is similar to that of *AMRICA* (Saphra and Lopez, 2015), but we use a different word aligner (fast_align rather than GIZA++[1]) and deterministic translation of role names if the labels are not in English. The deterministic translation is done using a mapping of the role names between Spanish and English provided in the Spanish annotation guidelines (Migueles Abraira, 2017). To align AMR graphs across languages, we word-align the sentence pairs, then map these alignments onto nodes in the graph (most concept labels on nodes correspond to lemmas of words in the sentence). Role names are mapped deterministically based on a list from Migueles Abraira (2017).

We normalize the strings and remove sense labels from the English and French/Spanish concept labels. An error that we noticed while developing the system was associated with the same concept label appearing more than once in either AMR, so we tag repeated words numerically before performing the alignment.

Finally, we run Smatch with the default number of 4 random restarts to produce an alignment. The Smatch score produced is an F1 score from 0 to 1 where 1 indicates that the AMRs are equivalent. This can be converted to a binary judgment, where all non-1 pairs are divergent, or used as a continuous value (as in §5).

Testing our Approach on Gold AMRs. One of the benefits of leveraging semantic representations in our approach to semantic divergence detection is that the identification of divergence boils down to determining whether the graphs are isomorphic or not (and accurate word alignment). This suggests that our pipeline algorithm (§4.2) should be highly effective at identifying whether AMR pairs are divergent or equivalent. In order to test our AMR-based approach to strict semantic equivalence identification, we first test on gold AMRs, which are created by humans and thus have no external noise from being automatically parsed.

We expect that our AMR divergence characterization would behave differently from a classifier of sentence-level divergence. This is because the sentence-level classification methods require specialized training data and as such learn to classify based on the perceived sentence-level judgments of semantic divergence. To test the strictness of our framing, we validate our quantification on gold English-French and gold English-Spanish cross-lingual AMR pairs.

Results on Gold English-French AMR Pairs We test our pipeline algorithm on the 100 English-French annotated AMR pairs described in §4.1. As expected, the simple pipeline algorithm is very accurate at correctly predicting whether the cross-lingual pairs do or do not diverge according to the

[1] fast_align has been shown to produce more accurate word alignments, such as in the case for Latvian-English translation (Girgzdis et al., 2014).

	Equivalent (17)			**Divergent** (83)			**All**
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1**
Ours	1.00	0.82	0.90	0.97	1.00	0.98	0.97
BC'20	0.39	0.82	0.53	0.95	0.73	0.83	0.75

Table 2: FR-EN: Binary divergence classification on on 100 gold French-English AMR pairs, annotated for sentences from the REFreSD dataset. Precision (P), Recall (R), and F1 scores are reported for the equivalent, divergent, and all AMR pairs. We compare the performance of our model with the performance of the (Briakou and Carpuat, 2020) model, referenced as BC'20, on our finer-grained measure of divergence for the same English-French parallel sentences.

stricter criterion.

Table 2 showcases the ability of our pipeline system and the (Briakou and Carpuat, 2020) system (described in §2) to identify these finer-grained semantic divergences. On these English-French AMR pairs, the F1 score for our system is 0.97 overall and 1.00 for equivalent AMR pairs. This high level of accuracy indicates we can reliably predict cross-lingual AMR divergence.

The (Briakou and Carpuat, 2020) system performs worse when using our finer-grained delineation of semantic divergence, achieving an F1 score of 0.75.[2] Unsurprisingly, the precision, recall, and F1 for their system is lower than the performance of our system, because theirs is not trained to pick up on these more subtle divergences. Note that on their own measure of divergence (perceived sentence-level divergence), the system achieves an F1 score of 0.85 on these same 100 sentences.

Of the 3 errors made by our algorithm (in all cases, classifying equivalent AMR pairs as divergent), 2 of the 3 are caused by word alignment errors. Named entities seem to pose an issue with fast_align for our use case.

	Equivalent (13)			**Divergent** (37)			**All**
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1**
Ours	1.00	0.92	0.96	0.97	1.00	0.99	0.98
BC'20	0.24	0.38	0.29	0.72	0.57	0.64	0.52

Table 3: EN-ES: Binary divergence classification with gold parallel AMRs. Included are Precision (P), Recall (R), and F1 for the Equivalent, Divergent, and All AMR pairs for our pipeline algorithm compared to the system by Briakou and Carpuat (2020), referenced as BC'20, on the same English-Spanish parallel sentences.

Results on Gold English-Spanish AMR Pairs. In addition to testing our system on our English-French AMR annotations, we test our system on the 50 English-Spanish AMRs and sentences released by Migueles-Abraira et al. (2018), who collected sentences from *The Little Prince* and altered them to be more literal translations; recent work classified these AMRs according to a structural divergence schema (Wein and Schneider, 2021).

In Table 3, we measure the ability of our pipeline system and the (Briakou and Carpuat, 2020) system to detect semantic divergences at a stricter level, as picked up by the AMR divergence schema.

Our system performs similarly well on Spanish-English pairs as it did on the English-French pairs, described in Table 2. This demonstrates that our pipeline algorithm is not limited to success on only one language pair, and we further affirm that the simple pipeline algorithm is a reliable way to predict cross-lingual AMR divergence.

5 Strictness Results Using Automatic English-French AMR Parses

In §4, we confirmed our hypothesis by demonstrating that we are able to use gold (human annotated) AMRs to capture a finer-grained level of semantic divergence, quantifiable via Smatch. We extend this further by determining whether fine-grained semantic divergences can be detected well even when using noisy automatically parsed AMRs. To do so, we compare the Smatch scores of automatically parsed AMR pairs with the human judgments output on the corresponding sentences by Briakou and Carpuat (2020).

To take the expensive human annotation piece out of the process, we show that automatic AMR parses can be used instead of gold annotations by establishing a threshold, instead of via binary classification. Therefore, we use the F1 score output by our pipeline algorithm as a *continuous score* and establish thresholds (described later in this section) to divide the data between divergent and equivalent.

We automatically parse cross-lingual AMRs for the entirety of the English-French parallel REFreSD dataset (1033 pairs). The REFreSD dataset is parsed using the `mbart-st` version of SGL, a state-of-the-art multilingual AMR parser (Procopio et al., 2021). The (monolingual) Smatch score for the SGL parser, comparing our gold AMRs with the automatically parsed AMRs, is 0.41 for the 100 French sentences using Smatch (0.43 using our

[2] The Briakou and Carpuat (2020) system does not take AMRs as input, so we use the corresponding sentences as input for their system.

pipeline algorithm)[3] and 0.52 for the 100 parallel English sentences using Smatch.

In doing error analysis, we find that the data points which are classified as having no meaning divergence but have extremely low F1 scores are largely suffering from parser error. We do find that there are pairs classified in REFreSD as having no meaning divergence at the sentence-level that do correctly receive low F1 scores. For example, the sentence pair in Figure 4, which has a REFreSD annotation of sentence-level equivalence and a gold AMR-level annotation of divergence, was assigned an F1 score of 0.3469.

Despite Smatch scores of 0.5 between the gold and automatic parses, both are usable for the task of detecting finer-grained semantic equivalence. To demonstrate the usefulness of our continuous metric of semantic divergence using automatically parsed AMR pairs, we develop potential thresholds at which you could separate data as being equivalent vs. divergent.

Because our metric is more sensitive, a practitioner could choose their own threshold by determining appropriate precision (how semantically equivalent they wanted a subset of filtered data to be) and recall (how much data they are willing to filter out) needs. This tradeoff is depicted in Figure 5. For example, if all pairs are marked as equivalent, precision would be approximately 40% on the REFreSD dataset if considering solely the "no meaning divergence" pairs equivalent.

Comparing with model probabilities. Though it is reasonable to assume that if the gold AMR annotations provide a distinctly finer-grained measure of divergence than sentence-level divergence then this would also be the case when using automatically parsed AMRs, we want to ensure the continued strictness of our methodology. To do this, we compare the values of our continuous metric and the probabilities produced by the (Briakou and Carpuat, 2020) system.

Because the probabilities produced by the system described in (Briakou and Carpuat, 2020) are always very close to 1 (equivalent) or very close to 0 (divergent) and there are far more divergent instances than equivalent instances, median and mode serve as a more effective form of comparison than mean between our F1 score and their probability score. Above the 0.7 threshold, the median F1 for our system is 0.7869 and mode is 0.8; the median probability for the Briakou and Carpuat (2020) system is 0.9990 and the mode is 1.0. For the 0.6 threshold, our median is 0.6667 and our mode is 0.6667; their median is 0.9871 and mode is 1.0. Above the 0.5 threshold, our median is 0.5814 and our mode is 0.5; their median is 0.8907 and mode is 1.0. Because these numbers are lower for our system than their system, we confirm that our measure is a stricter measure of equivalence even when using the automatically parsed AMRs.

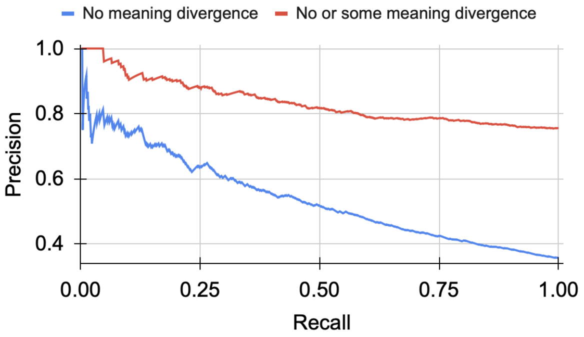

Figure 5: Precision / recall curve for equivalence detection in the 1033 sentence pairs in the full REFreSD dataset (English-French) using automatic AMR parses. Precision reflects the percent of sentences in which REFreSD human annotation was equivalent (as labeled as no meaning divergence in the blue/bottom curve, or as labeled as having either no or some meaning divergence in the red/top curve).

If the goal is to prioritize items for a human to look at on a fixed budget, the absolute scores may matter less than rankings, though the rankings additionally differ drastically. Of the top 50 sentences ranked by AMR divergence (which range in AMR similarity score from 0.96 to 0.67), only 19 of the 50 appear in the 166 sentences scored 1.0 by the Briakou and Carpuat (2020) system.

6 Sentence Similarity Evaluation with Automatically Parsed English-Spanish AMRs

As we have shown in previous sections, our AMR-focused approach in general is stricter than sentence-based measures of equivalence, in particular corpus filtering methods. Because our system is a stricter measure of semantic equivalence, it may be the case that our system can more precisely identify the most similar sentences than existing

[3] The SGL parser approaches cross-lingual parsing as the task of recovering the AMR graph for the English translation of the sentence, as defined in prior work (Damonte and Cohen, 2018). The result is that the parses of French sentences are largely in English, and default to French concepts only for out-of-vocabulary French words. The alignments in our pipeline account for this to better reward the native French concepts.

measures of sentence similarity. In this final results section, we look at the most semantically equivalent sentences in the dataset (as judged by our approach and as judged by multilingual BERTscore (mBERTscore; Zhang et al., 2020)) in comparison to their human judgments of equivalence. Specifically, we aim to investigate: (1) whether the average human similarity score for the most similar n sentences is higher when ranked by our AMR-based metric versus when ranked by mBERTscore, and (2) whether human judgments of sentence similarity for the most similar sentences are more correlated with our AMR-based metric than with mBERTscore (an embedding-based automatic evaluation metric of semantic textual similarity). We compare our AMR-based metric to mBERTscore because it has been shown to work well in cross-lingual settings when comparing system output to a reference (Koto et al., 2021). Semantic textual similarity considers the question of semantic equivalence slightly differently because it rewards semantic overlap as opposed to equivalence.

Data. To perform this comparison, we use the 301 human annotated Spanish-English test sentences from the news down of the SemEval task on semantic textual similarity (Agirre et al., 2016).

6.1 Smatch with Cross-Lingual AMR parsing

For our analysis, we use the Translate-then-Parse system (T+P; Uhrig et al., 2021). Providing the Spanish sentences as input, T+P translates them into English, and then runs an AMR parser[4] on the English translation. Because the Spanish sentence was translated into English and *then* parsed, this automatic parse can be compared against the automatic parse of the original English sentence with plain Smatch (no cross-lingual alignment added).

As we have established in §5, the noise introduced by automatic parsers can be overcome in our approach. We validate that the Smatch scores retrieved after using Uhrig et al.'s (2021) parser still bears some correlation with the Smatch scores on the aligned gold AMRs.[5]

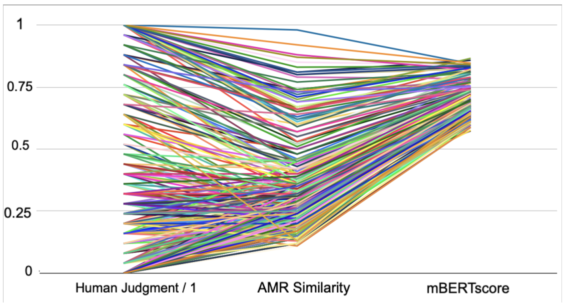

Figure 6: All data points normalized to a range of 0 to 1 for the Spanish-English sentence pairs from Agirre et al. (2016), including human judgment, AMR similarity score, and mBERTscore. This displays the decreased range of mBERTscore judgments in comparison to human judgments and AMR similarity.

6.2 Sentence Similarity Results

The average human judgment score, on a scale of 0 to 5 with 5 being exactly equivalent, for all sentence pairs which have an AMR similarity score greater than 0.8 is 4.98. The average human judgment score for all sentence pairs which have an mBERTscore similarity score greater than 0.8 is 4.89. Similarly, the average human judgment score for pairs with an AMR similarity score of greater than 0.7 is 4.86, while the average human judgment score for pairs with an mBERTscore greater than 0.7 is 3.8. This is because mBERTscore takes a much broader view of semantic equivalence. While the human judgments occupy the full range of 0 to 5, the mBERTscores of these sentences range from 0.57 to 0.87, as shown in Figure 6. The AMR similarity score ranges from 0.11 to 0.98.

This might suggest that then a higher threshold should be used for mBERTscore to achieve the same level of semantic granularity. However, our AMR similarity metric is also more correlated with human judgments for the most semantically equivalent sentences. For the top 20 items as ranked by AMR similarity, Pearson correlation with human judgments is 0.4068, while the top 20 items as ranked by mBERTScore are not correlated with human judgments (−0.0023). When looking at all items above the mBERTscore of 0.8, correlation with human judgment is 0.1645, whereas for all items above the AMR similarity score of 0.8, correlation with human judgment is 0.2675. Overall, AMR similarity score correlates with human judgment at a coefficient of 0.8367, which is slightly lower than the 0.8605 correlation between mBERTscore and human judgment. This evidence further supports that our metric is in fact a finer-

[4] Via amrlib: `https://github.com/bjascob/amrlib`

[5] On the 50 Spanish-English sentences mentioned in §4, the correlation between the Smatch scores (in comparison to the same gold AMRs) when using either the translation-then-parse method or the method of aligning concepts via fast_align is 0.31. This can be interpreted as a weak correlation. We find that both methods (translating the sentence first, or our pipeline algorithm aligning concepts in AMRs of different languages) work sufficiently well to capture the amount of divergence between cross-lingual AMR pairs.

grained measure of semantic equivalence, and is therefore better at identifying which sentences are exactly semantically equivalent.

7 Conclusion

In this work, we have proposed a stricter measure of semantic divergence than existing systems which rely on perceived differences at the sentence level. We have effectively demonstrated that parsing sentences into Abstract Meaning Representations and comparing those graphs facilitates a more detailed semantic comparison, when using either gold *or* automatically parsed AMR pairs.

We are excited by the numerous possible applications of this finer-grained measure of meaning (mentioned in §1), both from an engineering standpoint and the potential it has in translation and language-learning environments to highlight specific differences in language pairs.

Limitations

As the first work exploring the use of AMR for fine-grained semantic equivalence assessment, our work faces a few limitations. First, our results were limited to the language pairs we work with. In the three languages pairs, we claim that our approach is a more fine-grained measure of semantic equivalence than existing approaches. Future work on other language pairs would provide further insight into its applicability to languages less syntactically similar to English. Second, it may be worth considering the use of other semantic representations in addition to AMR. Though our results confirm that AMR captures many aspects of meaning that are important to human judgments of cross-lingual similarity, AMR does not capture all aspects of semantics. Finally, our system is limited by the performance of automatic AMR parsers. In §5, we show that, despite Smatch scores of 0.5 between the gold and automatic parses, both are usable for the task of detecting finer-grained semantic equivalence. Still, it is reasonable to expect that better parsers would lead to better performance by our system, and thus our results currently suffer due to less-than-perfect performance.

Acknowledgements

Thank you to Rexhina Blloshmi, Eleftheria Briakou, and Luigi Procopio for providing system results, as well as to anonymous reviewers for their feedback. This work was supported in part by a Clare Boothe Luce Scholarship to the first author and NSF award IIS-2144881.

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.

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.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2019. Abstract Meaning Representation (AMR) 1.2.6 specification. https://github.com/amrisi/amr-guidelines/blob/master/amr.md.

Susan Bassnett. 2013. *Translation studies*. Routledge.

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.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. 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 7236–7249, Online. 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.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. 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 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.

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.

Bonnie Dorr. 1990. Solving thematic divergences in machine translation. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL '90, page 127–134, USA. Association for Computational Linguistics.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, Geneva, Switzerland. COLING.

Valdis Girgzdis, Maija Kale, Martins Vaicekauskis, Ieva Zarina, and Inguna Skadiņa. 2014. Tracing mistakes and finding gaps in automatic word alignments for Latvian-English translation. In Andrius Utka, Gintarė Grigonytė, Jurgita Kapočiūtė-Dzikienė, and Jurgita Vaičenonienė, editors, *Human Language Technologies – The Baltic Perspective*, volume 268 of *Frontiers in Artificial Intelligence and Applications*, pages 87–94. IOS Press.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proc. of CHI*, pages 439–448, New York, NY, USA.

Paul Kingsbury and Martha Palmer. 2002. From Tree-Bank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Online. Association for Computational Linguistics.

Wai Ching Leung, Shira Wein, and Nathan Schneider. 2022. Semantic similarity as a window into vector- and graph-based metrics. In *Proc. of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 106–115, Abu Dhabi, United Arab Emirates (Hybrid).

Ha Linh and Huyen Nguyen. 2019. A case study on meaning representation for Vietnamese. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 148–153, Florence, Italy. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Noelia Migueles Abraira. 2017. A study towards Spanish Abstract Meaning Representation. Master's thesis, University of the Basque Country, Donostia-San Sebastián, Spain, June.

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

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.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62.

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.

Michael Roth and Talita Anthonio. 2021. UnImplicit shared task report: Detecting clarification requirements in instructional text. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 28–32, Online. Association for Computational Linguistics.

Naomi Saphra and Adam Lopez. 2015. AMRICA: an AMR inspector for cross-language alignments. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 36–40, Denver, Colorado. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo and Thiago Pardo. 2019. Towards a general Abstract Meaning Representation corpus for Brazilian Portuguese. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 236–244, Florence, Italy. Association for Computational Linguistics.

Naoko Taguchi. 2005. Comprehending implied meaning in English as a foreign language. *The Modern Language Journal*, 89(4):543–562.

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.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. 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 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

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. Accounting for language effect in the evaluation of cross-lingual AMR parsers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3824–3834, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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).

Laura Zeidler, Juri Opitz, and Anette Frank. 2022. A dynamic, interpreted CheckList for meaning-oriented NLG metric evaluation – through the lens of semantic similarity rating. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 157–172, Seattle, Washington. Association for Computational Linguistics.

Yuming Zhai, Gabriel Illouz, and Anne Vilnat. 2020. Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5944–5956, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuming Zhai, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proc. of ICLR*, Online.