Measuring Fine-Grained Semantic Equivalence with Abstract Meaning Representation

Shira Wein

Zhuxin Wang

Nathan Schneider

GEORGETOWN UNIVERSITY

nert.georgetown.edu

IWCS 2023
Nancy, France ▪ 22 June 2023

Semantically 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.

Semantically Equivalent?

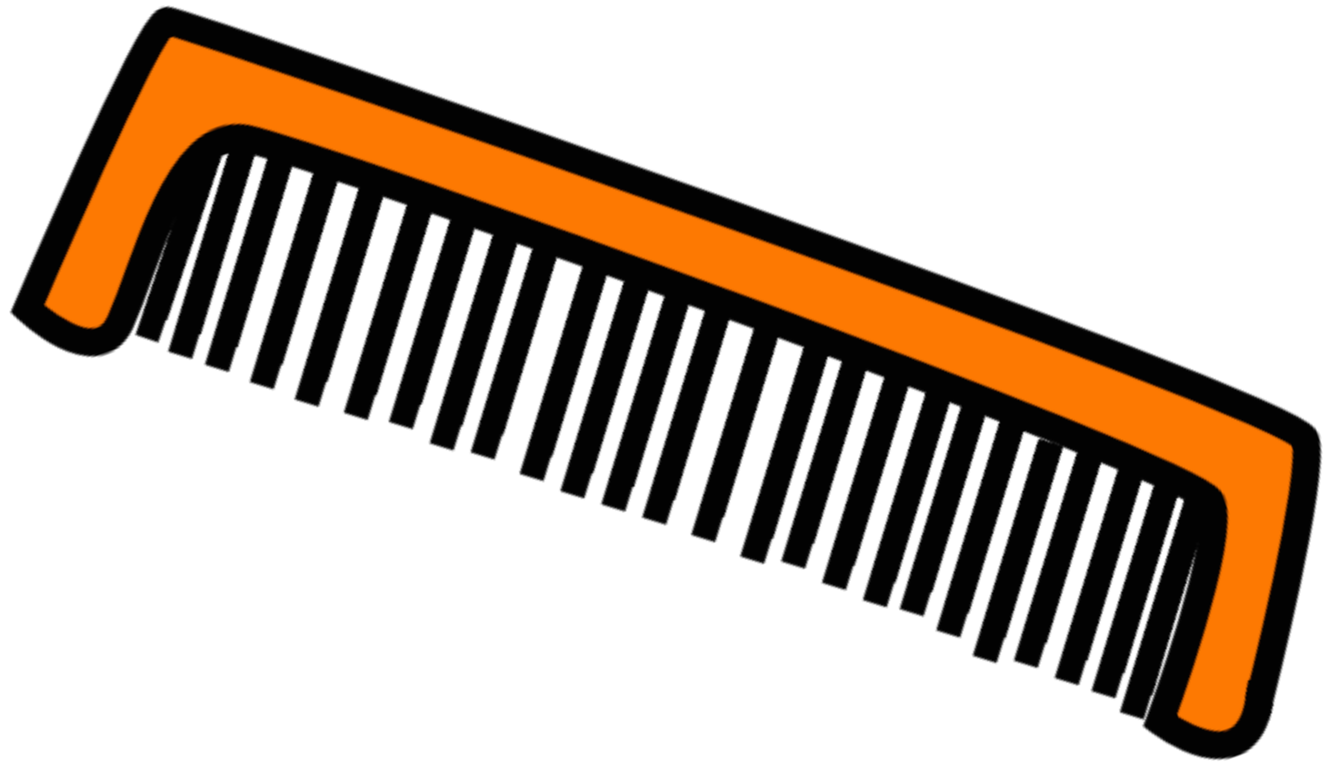
- Although the sales were slow (admittedly, according to the band), the second single from the album, “Sweetest Surprise” reached No. 1 in Thailand within a few weeks of release.
- Même si les exemplaires ont du mal à partier (comme l’admet le groupe), le second single de l’album, Sweetest Surprise, atteint la première place en Thaïlande la première semaine de sa sortie.

Key Idea

- A sentence and its translation can convey *essentially the same information overall* despite *slight semantic differences at the word/phrase level*.

Key Idea

- A sentence and its translation can convey *essentially the same information overall* despite *slight semantic differences at the word/phrase level*.
- We say a translation pair exhibits **fine-grained semantic divergence** if there is any difference in semantics (even if the overall meaning is understood to be the same).
  - **Equivalence** = lack of divergence



Semantically 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.

DIVERGENT

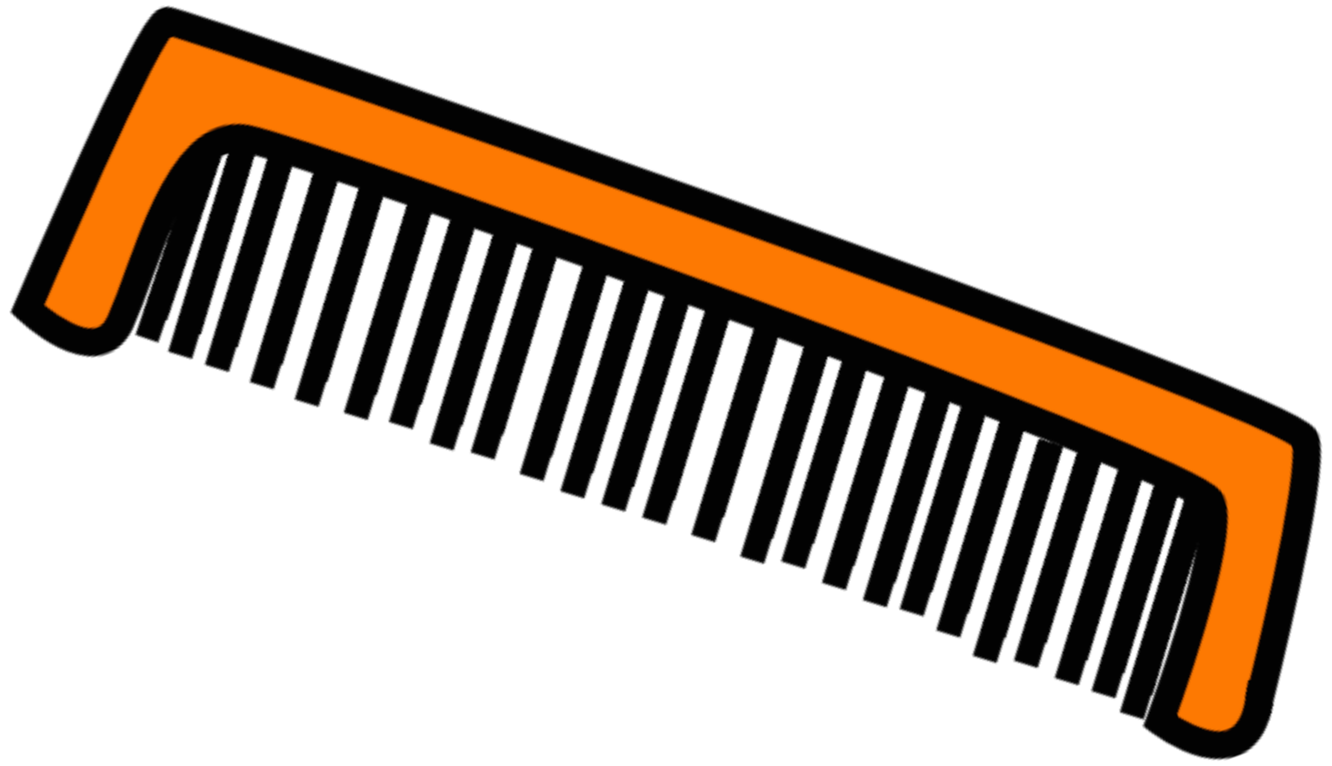
Semantically Equivalent?

- 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 partier (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 [the first week of its release].

DIVERGENT

Key Questions

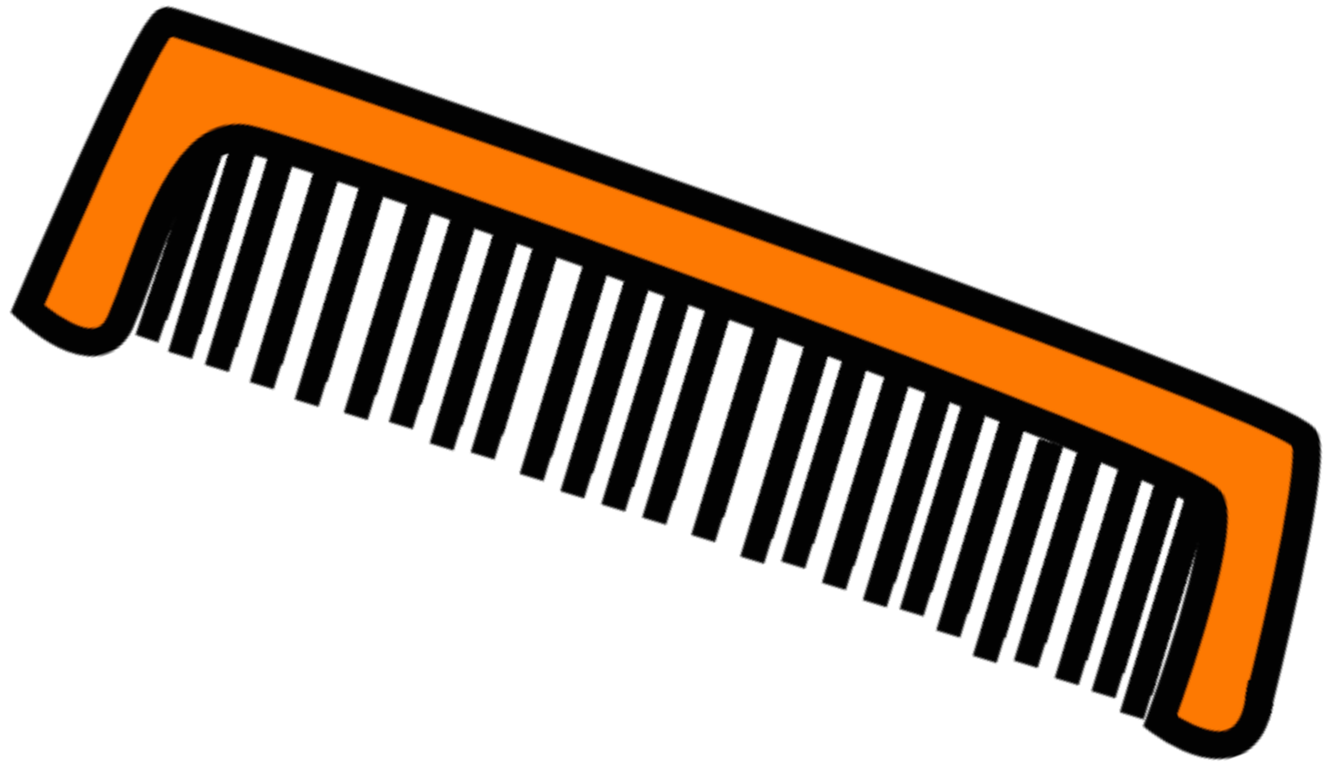
- Can we develop an algorithm to **predict** fine-grained divergence vs. equivalence?
- Can a semantic representation (AMR) help?



This talk

We explore these questions with two language pairs: English-French and English-Spanish.

- Background
- Sentence-level vs. fine-grained judgments
- Annotation
- Automatic detection using Smatch
- Gold vs. automatic AMR parses
- Sentence similarity evaluation



Translation Divergences in CL

- **Syntactic divergences:** Two languages conventionally use different constructions to express the same meaning ("I like Mary" vs. "María me gusta à mi") (Dorr, 1994; Deng & Xue, 2017)
- **Semantic divergences:** The source sentence and its translation differ in meaning (Carpuat et al., 2017; Vyas et al., 2018)

- Divergences cause difficulties for MT and other uses of parallel texts

Prior Approaches to Identifying Semantic Divergence

- Prior work identifying and classifying sentence-level divergences (Carpuat et al., 2017; Vyas et al., 2018)
- **REFreSD dataset** of English-French sentence pairs annotated with three types of divergences (Briakou and Carpuat, 2020)
- Fine-tuning to account for non-literal translations in the pre-training of cross-lingual language models (Zhai et al., 2020)

Semantic Divergence Detection

- Aims to pick out parallel texts which have less than equivalent meaning
- Current detection methods do not capture the full scope of semantic divergence
  - Rely on perceived *sentence-level divergences*

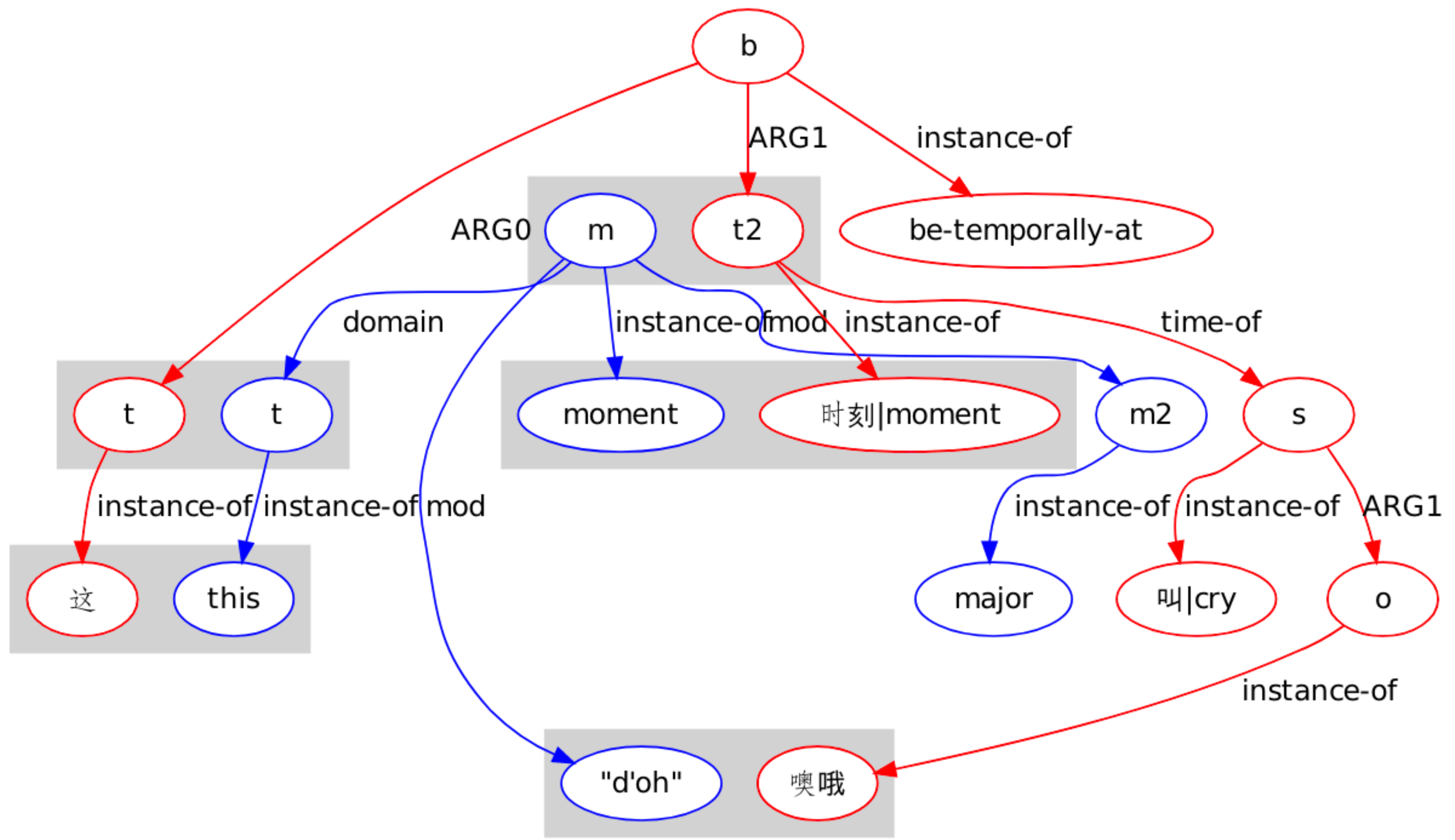
Although the sales were slow (admittedly, according to the band), the second single from the album, "Sweetest Surprise" reached No. 1 in Thailand *within a few weeks* of release.

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.

Two equivalent sentences in REFresD for which the AMRs diverge

AMR for Fine-Grained Semantic Divergence

- We hypothesize that a **semantic representation** such as AMR can facilitate precise meaning comparisons for fine-grained equivalence vs. divergence detection
    - Obtain semantic graphs of the source and target sentences, then compare
- AMR attempts to abstract away from syntax, focusing attention on semantic structure in the form of a graph (Banarescu et al., 2013)
    - Previously studied as a semi-interlingua (Xue et al., 2014; Wein and Schneider, 2021; Wein et al., 2022)



A crosslinguistic comparison of parallel AMRs (Xue et al., 2014)

Annotation of 100 French-English Pairs

- Sentence pairs from REFrESD dataset, with sentence-level equivalence ratings (Briakou and Carpuat, 2020)
- Annotated both sides with AMR
- Examined each pair of AMRs, annotated whether their contents are equivalent

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))
```

Sentences and AMRs for a pair of sentences which are equivalent in REFreSD (sentence-level) and via AMR.

AMR- vs Sentence-level Divergence

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

Comparison between AMR Divergence annotations and Sentence-level Divergence REFreSD annotations for 100 French-English sentences

First indication that AMR captures finer-grained divergences

Automatic Comparison of AMRs

- The Smatch algorithm (Cai and Knight, 2012) is the most widely used metric for AMR parsing
  - It computes an F1 score based on searching for an optimal alignment of nodes
- We are aligning graphs cross-lingually: different labels. We use a word aligner (fast_align; Dyer et al., 2013) to project the labels before running Smatch

Automatic Binary Classification of AMR-Divergence

Proof of concept with gold AMRs

	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

Binary divergence classification on 100 gold French-English AMR pairs, as measured by our finer-grained measure of divergence (cross-lingual adaptation of Smatch) for the same English-French parallel sentences

	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

Binary divergence classification on 50 gold Spanish-English AMR pairs (Migueles-Abraira et al. 2018; Wein and Schneider, 2021)

Automatic Binary Classification of AMR-Divergence

Proof of concept with gold AMRs

	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

(majority baseline accuracy: 0.83)

Binary divergence classification on 100 gold French-English AMR pairs, as measured by our finer-grained measure of divergence (cross-lingual adaptation of Smatch) for the same English-French parallel sentences

	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

(majority baseline accuracy: 0.74)

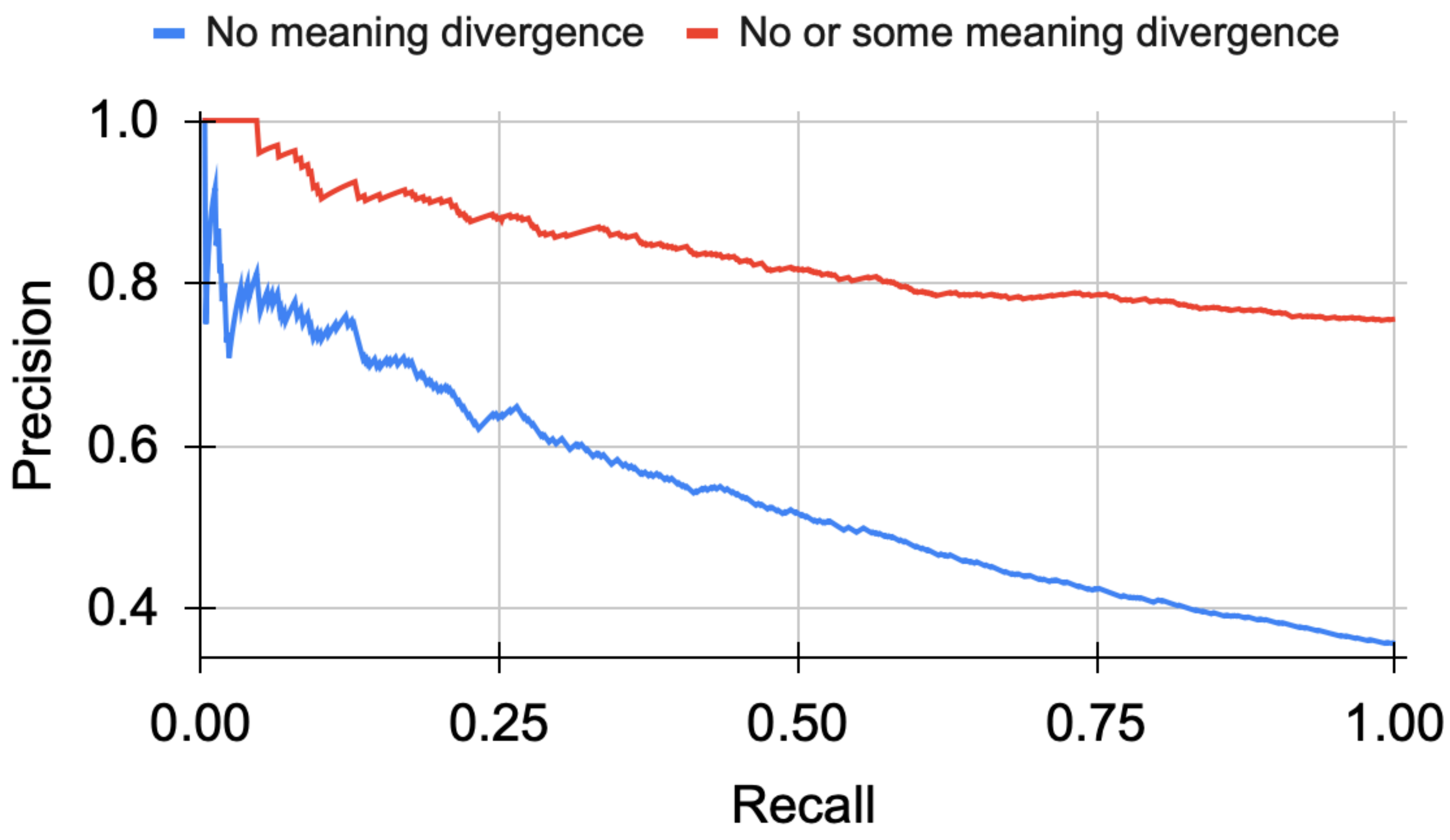
Binary divergence classification on 50 gold Spanish-English AMR pairs (Migueles-Abraira et al. 2018; Wein and Schneider, 2021)

Using Automatic AMR Parses

- Larger-scale experiment with 1033 pairs, automatic parses (SGL; Procopio et al., 2021)
  - Crosslingual parsing for French (predict English-style AMRs)
  - Parser correctness via monolingual Smatch: 0.52 (English), ≈0.42 (French)
- We don't have fine-grained equivalence annotations for this larger set, so we evaluate using REFreSD annotations
- Need to decide AMR similarity threshold
  - Various thresholds will result in higher precision/recall

Using Automatic AMR Parses

- Clear precision/recall tradeoff when evaluated on different criteria in REFreSD
- We further compare probabilities of our model to BC’20. BC’20 probabilities tend to be toward the extremes (near 0 or 1)—our approach has more flexibility in tuning the threshold.



Precision / recall curve for equivalence detection in the 1033 sentence pairs in the full REFreSD dataset (English-French) using automatic AMR parses.

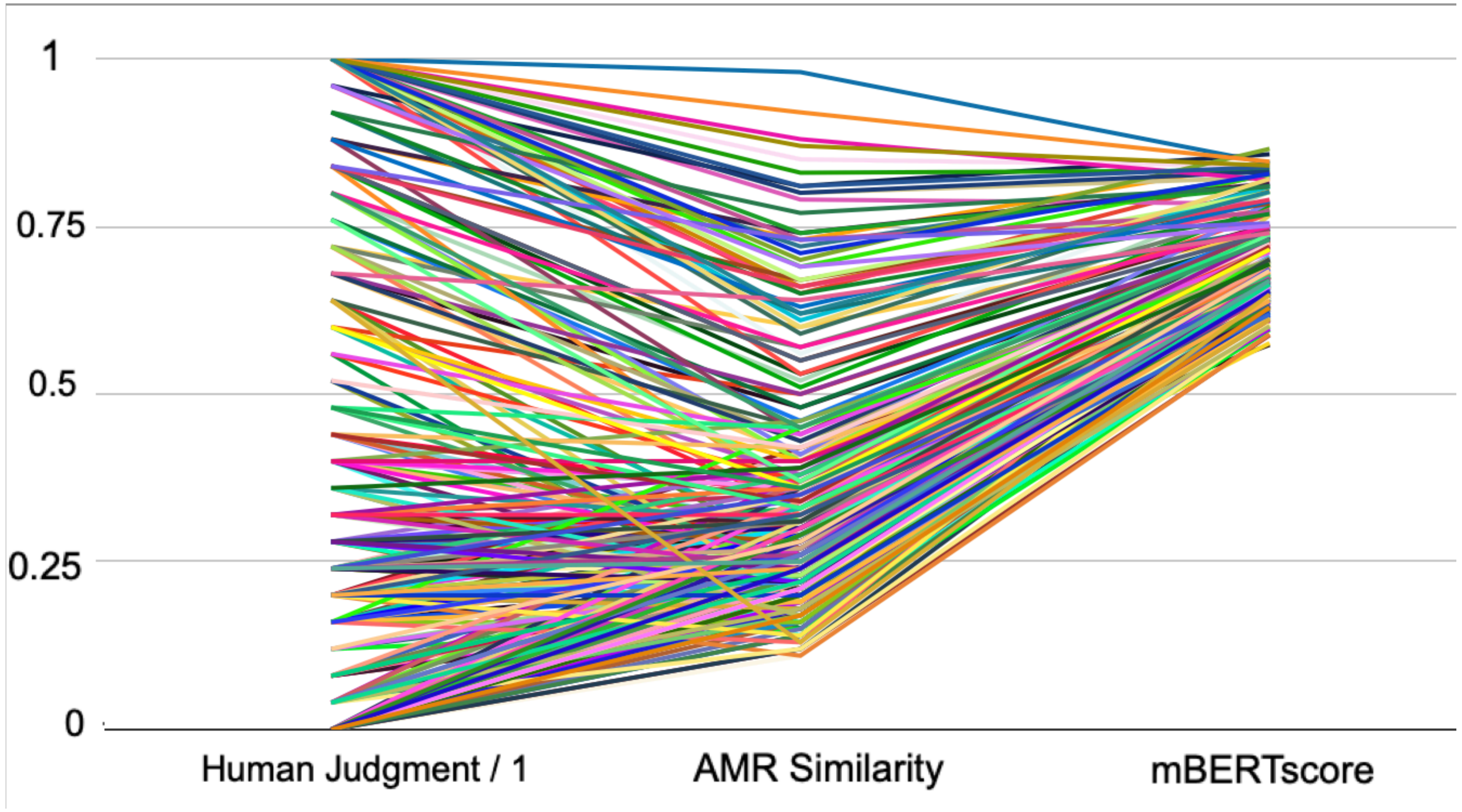
Semantic Textual Similarity Comparison

- Compare multilingual BERTscore (Zhang et al., 2020) to AMR-level divergence for semantic textual similarity in 301 Spanish-English sentence pairs
  - Translate-then-Parse system (Uhrig et al., 2021)

AMR vs mBERTscore

- At any high threshold of similarity, sentences ranked highly via AMR are judged to be more similar by humans
  - mBERTscore's overall correlation is slightly higher

→ AMR is better at identifying which sentences are exactly semantically equivalent



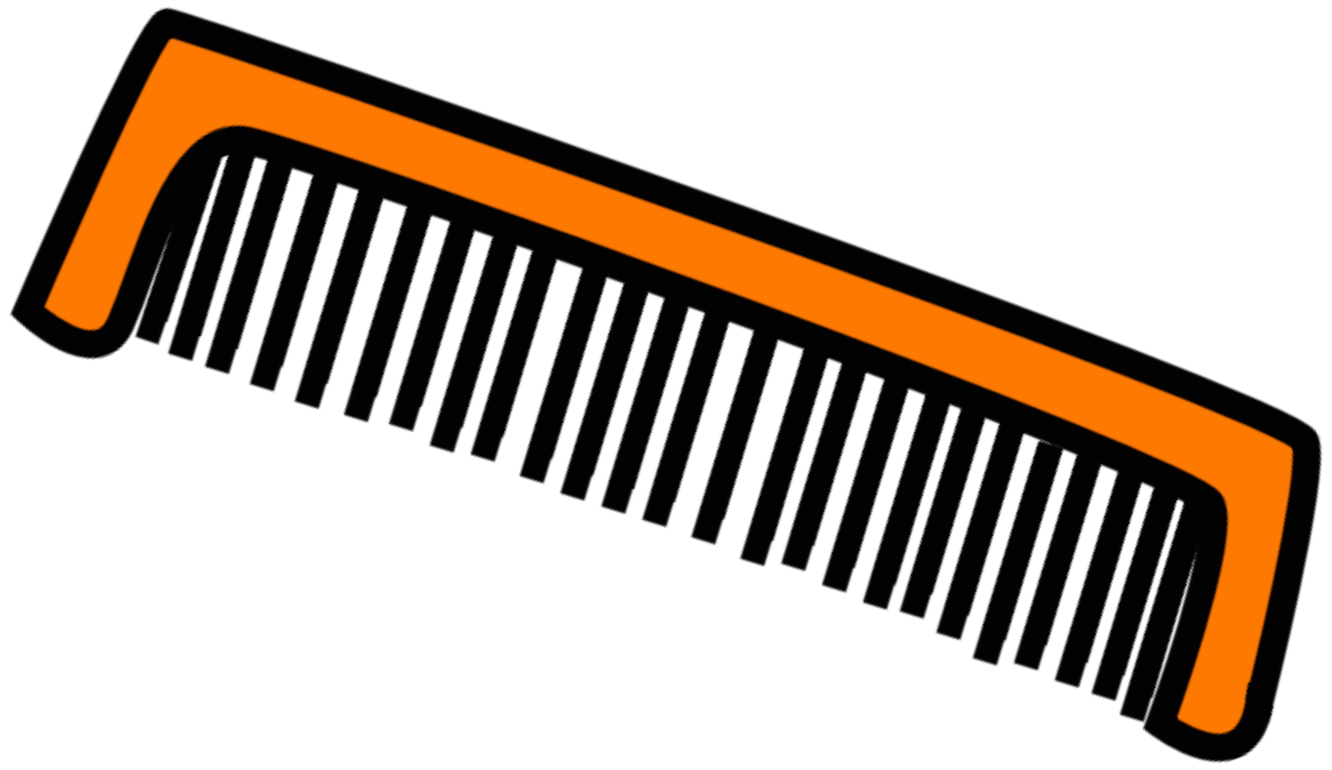
All data points normalized to a range of 0 to 1 for the Spanish-English sentence pairs, including human judgment, AMR similarity score, and mBERTscore.

Key Finding

AMR facilitates a stricter measure of fine-grained semantic equivalence in translation pairs.

(+ first attempt at AMR annotation for French!)

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



Potential Uses

- Filter out exactly semantically equivalent sentence pairs
    - Decreasing the amount of data that needs to be post-edited by human translators or annotated for human evaluation
    - Lessen the amount of annotation necessary for human evaluations of text (Saldías et al., 2022)
- Cross-lingual text reuse detection (plagiarism detection)
- Translation studies and semantic analyses could also benefit from the distinction between semantically equivalent sentence pairs and sentence pairs which have subtle or implicit differences (Bassnett, 2013)

Thanks!