Semantic Similarity as a Window into Vector- and Graph-Based Metrics

Wai Ching Leung
Georgetown University
wl607@georgetown.edu

Shira Wein
Georgetown University
sw1158@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

In this work, we use sentence similarity as a lens through which to investigate the representation of meaning in graphs vs. vectors. On semantic textual similarity data, we examine how similarity metrics based on vectors alone (SENTENCE-BERT and BERTSCORE) fare compared to metrics based on AMR graphs (SMATCH and S^{2}MATCH). Quantitative and qualitative analyses show that the AMR-based metrics can better capture meanings dependent on sentence structures, but can also be distracted by structural differences—whereas the BERT-based metrics represent finer-grained meanings of individual words, but often fail to capture the ordering effect of words within sentences and suffer from interpretability problems. These findings contribute to our understanding of each approach to semantic representation and motivate distinct use cases for graph and vector-based representations.

1 Introduction

Deriving sentence-level semantics is a nontrivial task and is fundamental to natural language understanding. Word embeddings (vectors) and graph-based formalisms are two kinds of sentence meaning representations that are widely used in NLP and NLG. One way to evaluate semantic representations is to compare their judgments on semantic similarity, often using human judgments as a reference, and there are automatic sentence similarity metrics that have been developed which make use of vector and graph-based models.

Vector-based models, such as SENTENCE-BERT (Reimers and Gurevych, 2019) and BERTSCORE (Zhang et al., 2019), rely on BERT embeddings. Though they are robust and highly efficient, they often suffer from interpretability issues and do not meet all of the expectations of a distributional semantics model (Mickus et al., 2019).

On the other hand, semantic formalisms such as Abstract Meaning Representation (AMR; Banarescu et al., 2013) are more explicit, and can be compared via graph-based similarity measures (Cai and Knight, 2013). While the explainability of these metrics is high, some studies have shown that they do not correlate with cross-lingual human-level judgments of similarity as well as embedding-based metrics (Wein and Schneider, 2022).

What is the difference between a stock and a bond?
What is the difference between a mode and a scale?

Figure 1: A sentence pair in the STS (Agirre et al., 2016) dataset which receives a human judgment of 0 (no similarity), an S^{2}MATCH similarity score of 0.75, a SENTENCE-BERT similarity score of 0.10, and a BERTSCORE score of 0.94. All three automatic metrics range from 0 to 1, where 1 indicates that the sentences are identical.

AMR-based metrics reflect the semantics of a sentence while abstracting away from syntactic features, while word-embedding based-metrics compare the tokens with contextualized embeddings. To this date, there is not a single approach that fully captures sentence meaning in a transparent fashion, so we need to investigate the strengths and weaknesses of different approaches in order to develop a better representation.

Given the importance of reflecting semantic similarity in formalisms of meaning, we hypothesize that semantic similarity is an informative way to investigate these representations. In this work, we investigate these semantic models through the lens of semantic textual similarity to investigate the differences between the various representations. These graph-based and the BERT-based approaches to automatically assessing semantic similarity have different strengths, but no work to date has thoroughly compared these metrics as a way to better understand their applicability and utility. We investigate how these metrics compare to human judgments of similarity on a semantic similarity task. Specifically, we compare and analyze the scores of two

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 106 - 115

AMR-based metrics—SMATCH and S^{2}MATCH—as well as two BERT-based metrics—BERTSCORE and SENTENCE-BERT—in relation to each other and in relation to human similarity judgments. For example, Figure 1 features a sentence pair with vastly different similarity judgments via human annotation and three of our automatic metrics.

Our primary contributions include:

- Quantitative results comparing AMR-based metrics and BERT-based metrics, with each other and with human judgments of similarity
- Analysis of points of low and high agreement between metrics
- Investigation of semantic features captured by the metrics
- Discussion of the successes and weaknesses of the performance of the metrics

Our data are publicly available.[1]

2 Background & Related Work

Abstract Meaning Representation. Graph-based semantic representations take an explicit approach to representing the meaning of the sentence by defining the roles and relations of the concepts within the sentence. AMR is a semantic representation which captures the meaning of a phrase or sentence in the form of a directed, acyclic graph (Banarescu et al., 2013). The graph's nodes and edges correspond to concepts in the sentence and the relations between the concepts, respectively. Methods for evaluating the performance of a text-to-AMR parser or computing the similarity of two AMRs include SMATCH and S^{2}MATCH, described in §3.

Wein and Schneider (2022) introduce three methods for comparing cross-lingual pairs of AMRs and evaluate the AMR-based metrics against human judgments of sentence similarity and BERTSCORE. Cross-lingually, Wein and Schneider (2022) found that BERTSCORE was more correlated with human judgments than the AMR-based metrics.

Semantic Textual Similarity. Semantic textual similarity (STS) is the task of judging the semantic equivalence of two sentences (Agirre et al., 2016).

In the most recent SemEval Semantic Textual Similarity for monolingual data in 2016 (Agirre et al., 2016), the top performing team at that time incorporates WordNet information into word embeddings in their model (Rychalska et al., 2016). Wang et al. (2022) combine FrameNet information with SENTENCE-BERT (Reimers and Gurevych, 2019) to compute sentence similarity. WordNet and FrameNet focus on lexical information and relations between words or frames, which is distinctively different from AMR which represents lexical concepts and their relations within the *same* sentence. The state-of-the-art systems most correlated with human judgments consistently make use of Transformers, such as SMART-Roberta Large (Jiang et al., 2020), which achieves state-of-the-art 92.9 and 92.5 on Spearman's and Pearson's correlations.

Opitz and Frank (2022) introduced a similarity metric, Semantically Structured SENTENCE-BERT (S^{3}BERT), which combines the explainability of AMR metrics with the high performance of BERT-based approaches. For the STS task, S^{3}BERT obtains a correlation score of 83.7 on Spearman's rank correlation coefficient. S^{3}BERT separates SENTENCE-BERT embeddings into partitions and trains the sub-embeddings on individual aspects of AMR metrics. Opitz and Frank (2022) developed S^{3}BERT with the motivation of uncovering the semantic features that contribute to similarity ratings, and in doing so, develops hypotheses and conjectures about the reasons to combine these two methods based on their potential strengths and weaknesses. For example, AMR metrics are said to be able to capture specific aspects related to semantic similarity, such as semantic roles, but are less correlated with human judgments; the BERT-based metrics are more similar to human judgments but might lack sensitivity to word order. Relatedly, Mohebbi et al. (2022) combine semantic roles labels and dependency grammar on top of the BERT Transformer model (Devlin et al., 2019) with the aim of computing semantic textual similarity.

In order to provide a more fine-grained evaluation of existing AMR parsers, Damonte et al. (2016) compare their parser with JAMR (Flanigan et al., 2014) and CAMR (Wang et al., 2015) on various sub-tasks, such as named entity identification and semantic role labeling, and conclude that there is not a single parser that outperforms others in all sub-tasks.

While prior research efforts have focused on either combining explicit information from graph-based resources with vectors, to the best of our knowledge, there has not been a direct, fine-grained comparison between these two formalisms. In our

[1] https://github.com/chingachleung/Vector-and-Graph-Based-Metrics

work, we perform a comparative analysis testing the hypotheses proposed in Opitz and Frank (2022) and analyzing the distinct strengths of graph versus vector-based representations on the task of semantic similarity.

3 Sentence Similarity Metrics

In this work, we investigate the performance of and differences between four metrics of sentence-level semantics: SMATCH (Cai and Knight, 2013), S^{2}MATCH (Opitz et al., 2020), SENTENCE-BERT (Reimers and Gurevych, 2019), and BERTSCORE (Zhang et al., 2019). We select these four metrics, two AMR-based and two BERT-based, because they are popularly used to compute sentence similarity (Wang et al., 2022), and are often used for downstream NLP tasks which depend on sentence semantics, such as paraphrase detection (Issa et al., 2018) and coreference predictions (Anikina et al., 2020). Moreover, we wish to contrast the weaknesses and strengths of graph and BERT based metrics.

SMATCH: SMATCH is a widely used metric that measures whole sentence semantic structure similarity by computing the overlaps of structures between sentences that are represented in AMR graphs (Cai and Knight, 2013). Since AMR abstracts away from syntax, syntactic paraphrases are expected to be represented with the same graph. A SMATCH score of 1 indicates semantic equivalence between two sentences, and a SMATCH score of 0 indicates that two sentences are completely dissimilar. Figure 2 shows the AMR graph of two semantically identical sentences. In order to compute sentence similarity, sentences are first parsed into AMR graphs. SMATCH then aligns the graphs by finding the maximum number of triple matches (there are two types of triple matches: `<var, instance, concept>` and `<var, relationship, var>`), which is achieved by using a hill-climbing method with a smart initialization and 4 random starts to increase the probability in finding the highest number of matches (Cai and Knight, 2013).

```
(g / give-01
      :ARG0 (h / he)
      :ARG1 (m / money)
      :ARG2 (s / school)
```

Figure 2: The AMR graph for two syntactic paraphrases: *He gives the school money*, and *He gives money to the school*

S^{2}MATCH: This metric is an extension of SMATCH which incorporates word-embeddings to match synonyms or near-synonyms (Opitz et al., 2020). During graph alignment, if the cosine similarity between the word-embeddings of two nodes meets a threshold, these two nodes, even if they have a different surface form, are considered a match. As a result, the S^{2}MATCH score will go up according to the cosine similarity score. This addresses the disadvantage of SMATCH (Cai and Knight, 2013) that graded meanings are not measured. For example, `<var, instance, skinny>` and `<var, instance, thin>` are considered a match in S^{2}MATCH since "skinny" and "thin" are synonyms, but not in SMATCH since "skinny" and "thin" are two different words.

Following Reimers and Gurevych (2019), we set the S^{2}MATCH alignment threshold value to 0.5. Therefore, we only consider the similarity between nodes if their cosine similarity reaches 0.5 or higher.

In order to use SMATCH and S^{2}MATCH to compare sentence similarity, we first use an automated AMR parser (Bai et al., 2022) to convert sentence pairs into AMR graphs. The parser is a BART-based model (Lewis et al., 2020) that is trained on a self-supervised graph-to-graph method. The accuracy of the parser is 83.6% on the **AMR2.0** (LDC2017T10) dataset.

SENTENCE-BERT: SENTENCE-BERT (Reimers and Gurevych, 2019) is a BERT-based model that is pre-trained on the SNLI (Bowman et al., 2015) and the Multi-Genre NLI (Williams et al., 2017) datasets. It generates sentence embeddings using the Siamese BERT model architecture (Devlin et al., 2019). To measure sentence similarity using SENTENCE-BERT, we pass sentence pairs into this model to obtain sentence embeddings, and compute their cosine similarity.

BERTSCORE: This metric was designed with the intention to evaluate text generation quality (Zhang et al., 2019). To obtain BERTSCORES between reference and candidate sentences, this metric first matches the tokens in the sentences using a greedy method: each token in a sentence is matched to the most similar token in the other sentence. Therefore, tokens are potentially matched more than once. After, the normalized pairwise cosine similarity between their word embeddings are computed with an optional idf-importance weight-

Metric	Pearson	Spearman
SMATCH	0.54	0.52
S^2MATCH	0.55	0.53
SENTENCE-BERT	0.80	0.81
BERTSCORE	0.67	0.66

Table 1: Pearson and Spearman's Rho correlations with gold labels for each of the four metrics.

Metric	Mean
SMATCH	0.54
S^2MATCH	0.56
SENTENCE-BERT	0.63
BERTSCORE	0.93
Gold	0.53

Table 2: Mean scores of the metrics and the gold labels

ing which can put more weight on more indicative words of sentences during computation.

4 Data & Evaluation Protocol

We use the test data from the SemEval-2016 Semantic Textual Similarity English Subtask (Agirre et al., 2016) to evaluate the metrics on their degree of alignment with human judgments. The data contains 1,189 pairs of English snippets from five sources: newswire headlines, short-answer plagiarism, machine translation post-editing, Q&A forum answers, and Q&A forum questions. All the pairs are labeled on an ordinal scale from 0 to 5, with 0 indicating the texts are completely dissimilar, and 5 indicating they are semantically equivalent. For example, the sentences *the bird is bathing in the sink* and *birdie is washing itself in the water basin* are labeled as 5, while *John went horse riding at dawn with a whole group of friends* and *Sunrise at dawn is a magnificent view to take in if you wake up early enough for it* are labeled as 0.

To measure the correlation between each metric with human judgments, we use both the Spearman's and Pearson's rank correlation statistics. SENTENCE-BERT (Reimers and Gurevych, 2019) also use the same task to evaluate their model and report 74% Spearman's Rank correlation.

Since the AMR parser (Bai et al., 2022) used in this experiment may output multiple alternative AMR graphs due to linguistic ambiguity, we only select sentences that are only parsed into a single graph to avoid impact caused by ambiguity on the correlation results. As a result, a total of 1138 sentence pairs are selected in our test set. This pre-processing step is very likely the reason why there is a discrepancy between our reported score as shown in Table 1 and the reported score from (Reimers and Gurevych, 2019).

5 Results and Analyses

In this section, we present qualitative and quantitative analyses of the performance of each of the four metrics: SMATCH, S^2MATCH, SENTENCE-BERT, and BERTSCORE. Specifically, we aim to:

- Evaluate metric quality for measuring semantic similarity, using human judgments as a reference
- Identify the challenges of incorporating embeddings into graph-based metrics
- Identify challenging and easy scenarios, by looking at what types of sentences the metrics correlate best and worst with each other
- Uncover the strengths and weaknesses of each metric, by analyzing what semantic aspects these metrics are able to capture

5.1 Semantic Metric Quality

In our experiment, we use both Pearson's and Spearman's correlation coefficients to compute correlations between the metrics and human judgments to avoid bias towards certain metrics due to our choice of correlation tests. As shown in Table 1, SENTENCE-BERT has the strongest correlation with the gold labels on both the Pearson's and Spearman's Rho. SMATCH and S^2MATCH have the lowest correlations with the the gold labels: 0.54 and 0.55 on Pearson's, and 0.52 vs. 0.53 on Spearman's respectively.

Although SENTENCE-BERT has the highest similarity with human judgments, it suffers from low interpretability. In particular, we find it hard to account for seemingly inconsistent predictions. For example, the sentence pair *What is the best way to store fresh berries?* vs. *What is the best way to cite an anonymous writer?* receives a similarity score of 0.06 from SENTENCE-BERT, but this pair *What is the best way to treat a feline ringworm?* vs. *What is the best way to clean a grater?* receives a similarity of 0.4 from the same model. Intuitively, the differences in these two sentence pairs are very similar, but they have very different similarity scores.

Besides correlation with human judgments, we also look at the mean scores of the metrics versus that of the human judgments, in order to understand the likelihood of each metric considering sentences similar or dissimilar. This will be particularly useful if the metrics are used as an off-the-shelf tool to

compute similarity in downstream NLP tasks. The mean score of the test data in our experiment is 3.2 on an ordinal scale from 0 to 5, which translates to 0.53 on a 0–1 scale. We find that the graph-based metrics' scores are closest to the mean score of human judgments, whereas BERTSCORE's mean is significantly higher (see details in Table 2). The high scores produced by BERTSCORE could be misleading, especially in cases where sentences are completely dissimilar. Therefore, we also investigate how this metric scores dissimilar sentences. Out of 198 sentence pairs that are annotated as completely dissimilar by human judgments, BERTSCORE gives an average score of 0.89, remarkably higher than the average scores of S^{2}MATCH and SENTENCE-BERT for the same pairs, which are 0.36 and 0.28. For example, *Step towards* and *Not in sight* is a sentence pair rated as 0 by human judgments, but BERTSCORE gives a similarity score of 0.86. A potential cause is its greedy approach for token matching: tokens are matched with the most similar tokens from the other sentence, even if they have already been matched with other tokens. The way the sentences are constructed in the STS data also exacerbates this behavior: on average, 57% of the tokens in reference sentences also occur in their candidate sentences, which means that there are over 50% of the tokens which are considered exact matches by BERTSCORE, even if the tokens are used differently.

Based on the correlations with human judgments and the mean scores of the metrics, we conclude that SENTENCE-BERT's scores are the most indicative of semantic similarity between sentences.

5.2 Challenges of Incorporating Embeddings into AMR Metrics

Theoretically, S^{2}MATCH combines the strengths of graph-based and vector-based metrics, but its low correlation with human judgments calls into question how embeddings have been incorporated into graphs. In order to distinguish the performance of SMATCH and S^{2}MATCH, we first compare the similarity between SMATCH and S^{2}MATCH by running the Spearman's and Pearson's correlation tests on their similarity scores. We obtain 0.98 on both of the tests, which serves as a strong indicator that these metrics have extremely similar behavior. As posited in Opitz et al. (2020), S^{2}MATCH scores are the same as or higher than SMATCH scores due to the additional consideration of graded meaning. Their extremely similar correlation scores with human judgments also implies that the use of vectors in S^{2}MATCH does not improve its representations of meanings. Digging into the STS data, we observe several challenges that help explain why this metric fails to achieve the 'best of both worlds'.

Cosine similarity may not reflect semantic similarity: For the sentence pair *What type of faucet is this?* vs. *What kind of socket is this?*, the words 'kind' and 'type' have a cosine similarity score of 0.6, which is above the threshold we set. As a result, S^{2}MATCH considers these two tokens a match, and incorporates the cosine similarity score into the final similarity score. This makes the S^{2}MATCH score of this pair higher than the SMATCH score. However, we also see that the cosine similarity score of 'this' and 'kind' is 0.778, which is higher than the cosine similarity score between 'kind' and 'type'. Although 'this' and 'kind' are not matched, hence their cosine similarity score is not computed into the final similarity score, it shows that embedding similarity might not be always reliable for comparing semantic similarity, which directly impacts the performance of S^{2}MATCH.

Conversion of words into frames potentially hinders embedding comparison: During AMR parsing, words can be represented with a frame that looks different. This poses a challenge for comparison via embeddings since the embeddings of words and their frames can be different. For example, there was a pair in the test data where the synonyms 'therefore' and 'thus' were used in the same way, but given different frames, `cause-01` vs. `infer-01`. The cosine similarity between similarity between 'therefore' and 'thus' is 0.91, whereas the cosine similarity between 'infer' and 'cause' is 0.23, which is lower than the threshold we set. Since S^{2}MATCH computes similarity between frames instead of words, the synonyms 'therefore' and 'thus' could not be matched during S^{2}MATCH score computation. As a result, the S^{2}MATCH and SMATCH scores are the same for this pair.

Another scenario is when AMR 'unpacks' the lexical semantics depending on derivational morphology which may differ between synonyms, obscuring their semantic similarity. For example, the relational meaning of 'employer' in *How do I maintain a good relationship with an employer after resigning?* is expressed with the AMR frame `employ-01`. This not only causes subsequent changes in its graph structures, but also makes S^{2}MATCH

Metric	Mean
S^{2}MATCH	0.59 (0.23_z)
SENTENCE-BERT	0.29 (-1.42_z)
BERTSCORE	0.93 (0.11_z)

Table 3: Mean absolute scores and z-scores of the metrics on 30 most dissimilar pairs (label 0 or 1). The z-score refers to the number of standard deviations from the mean value from each metric.

less likely to match it with the word 'boss' in the sentence *How do I maintain a good relationship with my old boss after being promoted?*. (The AMR graph does not unpack the relational meaning of 'boss' in the same way because it is not signaled with a derivational suffix.) Even if `boss` and `employ-01` were matched, their cosine similarity would be artificially low because they have different parts of speech.

Given the similar behavior of SMATCH and S^{2}MATCH, the following subsections will focus on S^{2}MATCH in relation to the vector-based metrics SENTENCE-BERT and BERTSCORE. We look at examples where the metrics exhibit low (§5.3) and high (§5.4) agreement with each other, to identify challenging and easy cases, and discuss the impact of three semantic features: negation (§5.5), semantic roles (§5.7) and paraphrases (§5.6).

5.3 Low Agreement Between Metrics

In order to compare how the metrics rank semantic similarity differently, we first convert the raw scores from each metric into z-scores using standard scaling. Next, for each sentence pair, we compute the variance of the 3 metrics' z-scores. We examine the sentence pairs with high cross-metric variance and observe that most are judged by humans as dissimilar in meaning (i.e., disagreement among metrics predicts low meaning similarity).

As a means of comparing the metrics, we then investigate the reverse: how pairs with the lowest meaning similarity as judged by humans tend to fare on different metrics (see Table 3). We find that SENTENCE-BERT's judgment is similar to the gold labels, whereas S^{2}MATCH and BERTSCORE tend to consider them more similar than they actually are. For example, the pair in Figure 1, which receives a human judgment score of 0, is found to have the highest variance between the metrics. S^{2}MATCH and BERTSCORE give a similarity score of 0.75 (0.97_z) and 0.94 (0.4_z) respectively, whereas SENTENCE-BERT gives a significantly lower score, 0.1 (-2.2_z).

We believe it is challenging for BERTSCORE and S^{2}MATCH to overcome surface level similarity when computing semantic similarity. In contrast, because of how SENTENCE-BERT is pretrained on data, surface features might not necessarily obstruct their semantic similarity judgment.

5.4 High Agreement Between Metrics

Based on our observation of the top 30 pairs that have the lowest cross-metric variance, we find that the metrics agree strongly with human judgments as well as each other on sentences that exhibit either of these two patterns:

1. rated with high similarity by all the metrics as well as human judgments; exist great overlap of words and argument structures
2. rated with low similarity by all the metrics as well as human judgments; have little or no overlap of words or argument structures

For example, this sentence pair falls into the first pattern type, and is ranked as having the most similar judgments from all the metrics, with a gold label of 4: *Sudanese soldiers had done this Sunday six of the kidnappers in the border area between Sudan, Chad and Egypt, and had arrested two of them.* and *Sudanese soldiers had killed six of the kidnappers this Sunday in the border area between Sudan, Chad and Egypt, and had arrested two of them.*

Meanwhile, this sentence pair exhibits the second pattern, has the fourth highest cross-metric agreement score and receives a gold label of 0: *The other method is the top down approach which is a method that combines memorization and recursion* vs. *The easiest way to look at inheritance is as an '... is a kind of' relationship.*

Since all of the metrics show a similar behavior with each other as well as with human judgments on both highly similar and highly dissimilar sentences, we can conclude that sentences with semantic similarity strongly correlated with their number of mutual surface features are "easy cases", i.e represented well by all the metrics.

5.5 Negation

Ettinger (2019) finds that pre-trained BERT is unable to capture the effect on negation on meaning. By contrast, AMR explicitly encodes negation through the inclusion of a polarity role, and the remainder of the graph is structured as if the negated statement did not appear. Currently, scope of nega-

Metric	Mean
S^{2}MATCH	0.92
SENTENCE-BERT	0.88
BERTSCORE	0.99

Table 4: Mean scores of the metrics on negated pairs.

Metric	Mean
S^{2}MATCH	0.74
SENTENCE-BERT	0.90
BERTSCORE	0.96

Table 5: Mean scores of the metrics on paraphrases.

tion is not captured in the AMR annotation schema (Stein and Donatelli, 2021).

We find that there is a significant discrepancy between human judgments and the metrics on the evaluation of negation. For example, the pair *You should do it* vs. *You should never do it* is considered very dissimilar (with label 1 on a scale from 0 to 5) by annotators , but is rated relatively more similar by the metrics: 0.86 by S^{2}MATCH, 0.97 by BERTSCORE and 0.45 by SENTENCE-BERT. Since there are only two negated pairs in the test data, we also randomly select 10 negated sentences from the NegDDI-DrugBank corpus (Bokharaeian et al., 2014) and the BioScope corpus (Szarvas et al., 2018), and manually construct their positive equivalents to compute their semantic similarity, in order to form a more robust analysis. For example, the positive equivalent of *These differences in gene expression have not been molecularly defined.* is *These differences in gene expression have been molecularly defined.*. As shown in Table 4, all the metrics rate the 20 pairs with high similarity. Since negation often reverses meanings of sentences, we believe it is essential to address the degree of impact of negation on meaning which both the graph-based and vector-based metrics fail to capture. We hope this result encourages more robust research in the future on determining the role of negation in semantic metrics.

5.6 Paraphrases

Compared with the BERT-based metrics, S^{2}MATCH is found to struggle more with paraphrases, especially when there are syntactic differences. We compare their scores on 164 paraphrase pairs (with gold label 5) in the test data, and find that S^{2}MATCH on average gives considerably lower similarity scores in comparison with the BERT-based metrics (see Table 5 for details). For example, the two sentences in Figure 3 are semantically identical, but S^{2}MATCH scores it with 0.65, which is considerably lower than the SENTENCE-BERT and BERTSCORE scores, 0.92 and 0.97 respectively. There are multiple potential explanations for this outcome: first, S^{2}MATCH cannot capture that 'use' and 'cash out' have the same contextual meaning because it uses static GloVe embeddings to compute similarity. Second, the arguments of 'to' are parsed into `:ARG2` and `:purpose` respectively, even they serve the same function in the sentences. In other words, the lexical and structural differences of these sentences lead to differences in their AMR graphs, resulting in a lower S^{2}MATCH score.

Therefore, although AMR is intended to abstract away from syntax and sentences with similar meanings should have similar graph structures, we have observed that this is not always the case: sentence structure affects AMR graphs and thus affects semantic similarity for AMR-based metrics.

Sentence 1: *Should I use IRA money to pay down my student loans?*

```
(r / recommend-01
      :ARG1 (u / use-01
            :ARG0 (ii / i)
            :ARG1 (m / money
                  :source (o / organization
                        :name (n / name
                              :op1 "IRA")))
            :ARG2 (p / pay-down-05
                  :ARG0 ii
                  :ARG1 (l / loan-01
                        :ARG2 ii
                        :mod (p2 / person
                              :ARG0-of (s / study-01)))))
      :polarity (a / amr-unknown))
```

Sentence 2: *Should I cash out my IRA to pay my student loans?*

```
(r / recommend-01
      :ARG1 (c / cash-out-03
            :ARG0 (ii / i)
            :ARG1 (p / product
                  :name (n / name
                        :op1 "IRA")
                  :poss ii)
            :purpose (p2 / pay-01
                  :ARG0 ii
                  :ARG3 (l / loan-01
                        :ARG2 ii
                        :ARG3 (p3 / person
                              :ARG0-of (s / study-01)))))
      :polarity (a / amr-unknown))
```

Figure 3: AMR graphs for two paraphrases.

5.7 Semantic Roles

We observe that the AMR-based metrics are able capture semantic roles and argument structures,

which might not be captured by BERT-based metrics, or even human judgment.

For example, the pair in Figure 4 has the 9th highest cross-metric variance, which has a z-score of -0.655_z from S^{2}MATCH, -3.31_z from SENTENCE-BERT and -2.67_z from BERTSCORE, and is annotated as completely dissimilar by human annotators. In other words, S^{2}MATCH considers this pair much more similar than the BERT-based metrics as well as human judgments.

Sentence 1: *Spanish bulls gore seven to death.*

```
(g / gore-01
      :ARG0 (b / bull
            :mod (c / country
                  :name (n / name
                        :op1 "Spain")))
      :ARG1 (p / person
            :quant 7)
      :ARG2 (d / die-01
            :ARG1 p))
```

Sentence 2: *Obama queries Turnbull over China port deal.*

```
(q / query-01
      :ARG0 (p / person
            :name (n / name
                  :op1 "Obama"))
      :ARG1 (p2 / person
            :name (n2 / name
                  :op1 "Turnbull"))
      :ARG2 (d / deal-01
            :ARG2 (p3 / port
                  :location (c / country
                        :name (n3 / name
                              :op1 "China")))))
```

Figure 4: AMR graphs for two sentences that have similar argument structures.

One reason that accounts for such a distinct judgment from S^{2}MATCH is that the two sentences share certain similarity in their argument structures: their main predicates have these three arguments `ARG0`, `ARG1`, `ARG2`. Both the `ARG0`s refer to an agent, and the `ARG1`s refer to a patient or theme.

Although we use human judgments as a reference to evaluate the performance of the metrics, this example pairs shed some lights on the dimensions of meaning: argument structures represent semantic relationship between arguments, providing a high level of meaning representation of a sentence. It is then worth asking if we should take argument structure into consideration when comparing semantic similarity, and in what use cases we should or should not.

One might argue that since this pair is completely irrelevant, it makes sense not to consider them similar at all. However, we observe that the sensitivity to argument structures of the AMR-based metrics can address drastic change of meaning due to change of semantic roles. For example, the sentence pair *(A|B) is the conditional probability of A, given B* vs. *P(B|A) is the conditional probability of B given A* has a gold label of 3, S^{2}MATCH score of 0.39, SENTENCE-BERT score of 0.99 and BERTSCORE score of 0.98. In this case, S^{2}MATCH's judgment is much more similar with human judgments than the BERT-based metrics which regard this pair as almost equivalent.

The cause for such a difference between the graph-based and the vector-based metrics is that the former identifies the argument of `give-01` changes from `B` to `A`, whereas since word order is not explicitly encoded in the computations of the vector metrics, such a change is likely to have no impact in their similarity judgments.

5.8 Summary of Findings

Among the metrics we compared, we found that SENTENCE-BERT is most similar to human judgments, but its judgment lacks interpretability because it is a pretrained model with its performance dependent on the training data.

S^{2}MATCH's design takes advantage of both graph-based and vector-based metrics, but fails to take full advantage of vectors to compare word similarity due to changes caused by AMR parsing. Therefore, we suggest concepts that S^{2}MATCH aligns could have their embeddings represented by their original words in the sentence, not the concept labels themselves, so embeddings of words instead of embeddings of their concepts are compared for cosine similarity. For example, taking advantage of a system that maps AMR concepts to tokens, 'employer' would be aligned with 'boss', not `employ-01` with `boss`.

We have also identified challenging cases where S^{2}MATCH and BERTSCORE fail to account for the inverse relationship between surface level similarity and semantic similarity, and easy scenarios when semantic similarity positively correlates with surface level similarity.

Finally, we looked at three semantic features, negation, semantic role arguments and paraphrases. We found that all the metrics do not account for the impact of negation on meaning, only the graph-based metrics are sensitive to role arguments but fail to capture semantic similarity of paraphrases.

6 Conclusion

We compared four graph- and vector-based semantic metrics via a semantic similarity task. In the task, we used human judgments as a reference and explored various scenarios to investigate the strengths and weaknesses of these metrics, both qualitatively with examples and quantitatively via correlation with human judgments. We found that graph-based metrics are highly accurate in capturing meaning variations driven by change in sentence structures, whereas vector-based metrics allow more fine-grained meanings of individual words due to contextual embeddings.

As we used automatic parsers in our experiments, the results were certainly affected by some amount of parser error. In future work, it would be interesting to see how gold AMR graphs perform in the same experiment. We also hope that our analyses can motivate more robust research on utilizing the strengths of both vector- and graph-based meaning representations, allowing more effective semantic representation at both the word and sentence levels.

Acknowledgements

Thank you to anonymous reviewers for their feedback. This work is partially 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.

Tatiana Anikina, Alexander Koller, and Michael Roth. 2020. Predicting coreference in Abstract Meaning Representations. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–38, Barcelona, Spain (online). Association for Computational Linguistics.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. *ArXiv*, abs/2203.07836.

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.

Behrouz Bokharaeian, Alberto Diaz, Mariana Neves, and Virginia Francisco. 2014. Exploring negation annotations in the drugddi corpus. *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BIOTxtM 2014)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

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.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2016. An incremental parser for abstract meaning representation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2019. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *CoRR*, abs/1907.13528.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

Fuad Issa, Marco Damonte, Shay B. Cohen, Xiaohui Yan, and Yi Chang. 2018. Abstract Meaning Representation for paraphrase detection. 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 442–452, New Orleans, Louisiana. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 2177–2190, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees Van Deemter. 2019. What do you mean, BERT? Assessing BERT as a distributional semantics model. *arXiv preprint arXiv:1911.05758*.

Majid Mohebbi, Seyed Naser Razavi, and Mohammad Ali Balafar. 2022. Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information. *Scientific Reports*, 12(1):1–11.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: decomposing sentence embeddings into explainable AMR meaning features. arXiv:2206.07023 [cs].

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP team at SemEval-2016 task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 602–608, San Diego, California. Association for Computational Linguistics.

Katharina Stein and Lucia Donatelli. 2021. Representing implicit positive meaning of negated statements in AMR. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 23–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gyorgy Szarvas, Veronika Vincze, Richard Farkas, and Janos Csirik. 2018. The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. A transition-based algorithm for amr parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

Tiexin Wang, Hui Shi, Wenjing Liu, and Xinhua Yan. 2022. A joint FrameNet and element focusing Sentence-BERT method of sentence similarity computation. *Expert Systems with Applications*, 200:117084.

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.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.