UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

THE EVALUATION OF ABSTRACT MEANING REPRESENTATION STRUCTURES

RAFAEL TORRES ANCHIÊTA

Nº 451

RELATÓRIOS TÉCNICOS

São Carlos – SP

Dez./2024

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

The Evaluation of Abstract Meaning Representation Structures

Rafael Torres Anchiêta

São Carlos
2024

ABSTRACT

ANCHIÊTA, RAFAEL T. **The Evaluation of Abstract Meaning Representation Structures**. 2024. 18 p. Report - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Abstract Meaning Representation (AMR) is a sentence-level semantic meaning representation that got the attention of the Natural Language Processing (NLP) community because of its simpler structure, representing a sentence as a direct acyclic graph based on previous well-known NLP resources. This stimulated the development of several AMR corpora and parsers, aiming to produce better natural language understanding tools and derived applications. This research line has also fostered the creation of evaluation metrics to assess the quality of the obtained representations. SMATCH is the dominant evaluation metric, but recently SEMA, SEMBLEU, and S^2MATCH metrics were designed based on the weaknesses of SMATCH. In this study, we perform varied (but complementary) experiments to determine which metric is the most appropriate, including two tests from the literature and proposing two new ones, creating a comprehensive evaluation setup. We show that SEMA is more suitable to evaluate AMR structures than the other analyzed metrics.

Keywords: Computational Semantics, Abstract Meaning Representation, Evaluation

1 INTRODUCTION

Abstract Meaning Representation (AMR) is a semantic formalism designed to capture the meaning of a sentence (Banarescu *et al.*, 2013). AMR structures may be encoded as graphs with explicit semantic features, such as semantic roles, word sense disambiguation, negation, and other semantic phenomena. In Figure 1, we present an example of an AMR graph for the sentence "*Something was broken in my engine.*". In this figure, the nodes are concepts, and the edges are relations among them. The concept `break-01` is the root of the graph, and `:ARG1`, `:location`, and `:poss` are AMR relations.

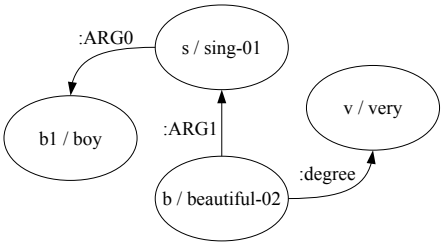

Figure 1 – An example of AMR graph for the sentence "*Something was broken in my engine.*" extracted from The Little Prince Corpus.

According to Bos (2016), AMR structures are manually easier to produce than traditional meaning representations. Because of that, AMR corpora for different languages arose. For instance, English[1], Chinese (Li *et al.*, 2016), Portuguese (Anchiêta; Pardo, 2018), Spanish (Migueles-Abraira; Agerri; Ilarraza, 2018), and others. With the availability of these corpora, the AMR parsing task, which is responsible for producing an AMR graph from a sentence got a lot of attention due to the need for better natural language understanding methods (Flanigan *et al.*, 2014; Wang; Xue; Pradhan, 2015; Artzi; Lee; Zettlemoyer, 2015; Damonte; Cohen; Satta, 2017; Noord; Bos, 2017; Anchiêta; Pardo, 2018; Lyu; Titov, 2018; Zhang *et al.*, 2019).

The growing interest in AMR representation, through the development of several corpora and AMR parsers, supported by three shared tasks (May, 2016; May; Priyadarshi, 2017; Oepen *et al.*, 2019), stimulated the development of methods/metrics to evaluate such graph structures since automatic evaluation plays an important role both in AMR parsing and annotation tasks.

Smatch (Cai; Knight, 2013) is the most famous evaluation metric. It measures the degree of overlap between two AMR structures, computing precision, recall, and f-score over AMR annotation triples. Anchiêta, Cabezudo e Pardo (2019) and Song and Gildea (Song; Gildea, 2019) pointed out that Smatch has some shortcomings. For comparing related nodes in the graph, Smatch obtains the maximum f-score via one-to-one matching of variables/nodes and comparing the edge labels, possibly producing high scores for some completely different AMR graphs. This one-to-one node mapping produces search errors, weakening the robustness of the metric. Moreover, the Smatch metric overvalues the `TOP` relation used to indicate which node is the root node in the graph. As the `TOP` relation produces a self-relation at the root node, this results in either a double penalty (if the root nodes of the hypothesis graph and the reference graph are different) or a double score (otherwise). To deal with these problems, the authors proposed two new metrics: Sema (Anchiêta; Cabezudo; Pardo, 2019) and SemBleu (Song; Gildea, 2019). Also, recently, Opitz, Parcalabescu e Frank (2020) proposed seven quality criteria to compare metrics and propose a new one named S^{2}match.

It is important to realize that evaluating AMR structures is not trivial since the metrics may generate different results for AMR graph pairs. For example, the sentences "*The girl asked the boy to leave*" and "*The boy wants to go*" may be encoded as graphs according to Figures 2 and 3, respectively.

[1] https://amr.isi.edu/download.html

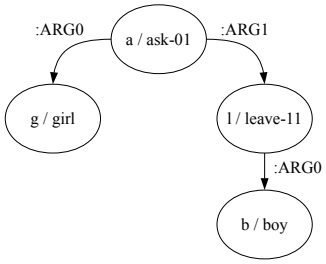

Figure 2 – AMR graph for the sentence "*The girl asked the boy to leave*".

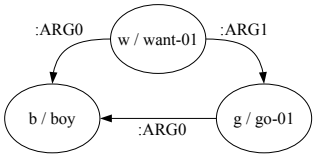

Figure 3 – AMR graph for the sentence "*The boy wants to go*".

We may see that the two graphs have a somewhat similar structure. Both graphs have outgoing edges from the root node with the same edge label (`:ARG0` and `:ARG1`) and an incoming edge at the `boy` node with the same edge label (`:ARG0`). Despite this similarity, the meanings of the sentences are different. Comparing the left graph against the right graph, SMATCH returns an f-score of 0.53, SEMA gives an f-score of 0.00, SEMBLEU yields a value of 0.21, and S^{2}MATCH shows an f-score of 0.45. Choosing an evaluation metric that gives over- or under-value results may overlook several issues, producing unreal values that may mislead the development and improvement of AMR parsing methods and AMR-based applications.

In this context, we carry out a study aiming to investigate which metric produces results most related to human judgment and shows to be more appropriate for evaluating AMR structures. We initially performed three analyses: at the corpus, the sentence, and representation graph levels. In the first one, we replicate the experiment carried out by Song and Gildea (Song; Gildea, 2019). They compared SEMBLEU and SMATCH metrics on the outputs of 4 systems concerning manual evaluation, over 100 sentences from the test set of the LDC2015E86 corpus. In the second one, to validate the first results, we create noisy sentences from the original ones and ask 5 annotators to rank these sentences from the best (most similar to the original) to the worst sentence (less similar to the original). Next, we use the metrics to create automatic ranks to be compared to those the annotators indicate. In the third one, we analyze two AMR properties (inverse relation and reification) to explore their impact on evaluation metrics. The first property allows swapping two AMR relations, inverting a relation, while the second property turns an AMR relation into a concept. These properties are interesting because they enable altering the original graph without losing meaning. Therefore, the metrics should (ideally) return an f-score of 1.0 even though the graphs differ. Finally, we also check if the metrics meet the seven criteria established by Opitz et al. (Opitz; Parcalabescu; Frank, 2020). Overall, we show that SEMA is the most suitable measure to evaluate AMR structures.

2 AMR BACKGROUND

AMR is a sentence-level semantic representation that incorporates several semantic features into its structure, such as semantic role, coreference, named entity, and word sense. Words that supposedly do not contribute to the meaning of the analyzed sentence are not annotated, as 'to' infinitive particle and articles, since they are referred to as "syntactic sugar" in the AMR original paper (Banarescu *et al.*, 2013).

As mentioned before, AMR may be encoded as a graph, but it is also usual to find it in PENMAN notation (Matthiessen; Bateman, 1991) or conjunction of logical triples. Figure 4 shows the canonical form in PENMAN and logical triples for the sentence "*That was by a Turkish astronomer, in 1909.*".

The AMR formalism has two types of nodes: concrete and abstract (or keywords). For instance, in the cited figure, the concrete nodes are `see-01` and `astronomer`, since they are present in the text, whereas the abstract nodes are: `country`, `name`, and `date-entity`, as they are not in the text. These notations

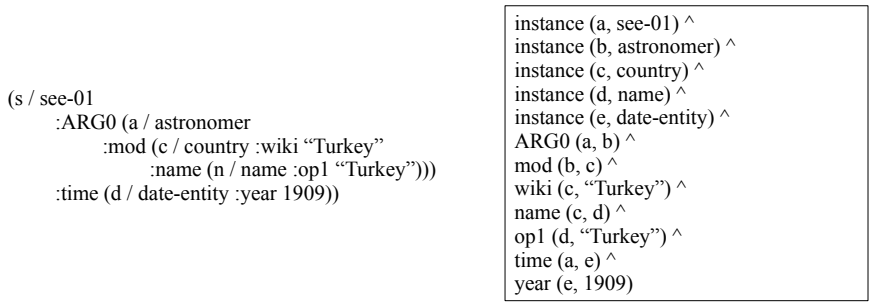

Figure 4 – In the left, the PENMAN notation and, in the right, the logical triples.

also have two constants: `Turkey` and `1909`, as they get no variables. AMR semantic relations are `:ARG0`, `:mod`, `:wiki`, `:name`, `:op1`, `:time`, and `:year`. The first AMR relation is a core role, while the others are non-core. `:mod` indicates a modifier, `:wiki` refers to wikification, `:name` introduces a named entity, `:op1` is used for conjunctions, and `:time` and `:year` are date-entity relations.

The existing metrics adopt two approaches to evaluate AMR structures: conjunction of logical triples or string-match. SMATCH, SEMA, and S^{2}MATCH metrics explore the first strategy, while SEMBLEU uses the second one.

SMATCH calculates the degree of overlap between two AMR structures, trying to find the one-to-one node mapping between two AMR structures. To compute precision and recall, it follows Equations 2.1 and 2.2, respectively:

$$P = \frac{M}{C} \qquad (2.1) \qquad\qquad R = \frac{M}{T} \qquad (2.2)$$

where M is the correct (according to a reference) number of triples, C is the produced number of predicted triples (e.g., by a parser or a different human annotator), and T is the total number of triples in some AMR reference. In practice, it considers a `TOP` attribute[1] that indicates which node is the root node of the graph. For example, when computing the AMR graph of Figure 2 (hypothesis) against the AMR graph of Figure 3 (reference), the Smatch metric produces the results in Figure 5. It considers as correct the triples `TOP(a, “top”)`, `ARG0(a, b)`, `ARG1(a, c)`, and `ARG0(c, d)`.

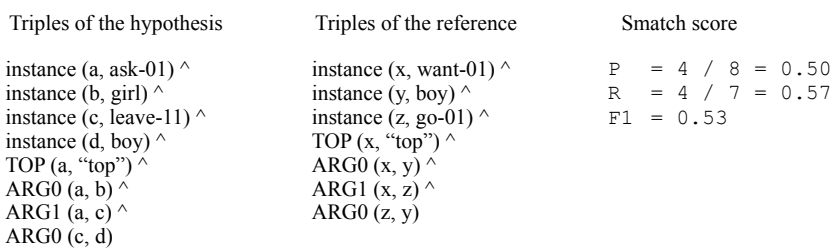

Figure 5 – Triples and results produced by the SMATCH metric.

S^{2}MATCH uses the same equations as defined by SMATCH to compute precision and recall. However, instead of maximizing the number of triple matches between two graphs as SMATCH does, S^{2}MATCH maximizes the triple matches by considering the degree of semantic similarity among concepts, allowing a soft match, even though they may not be identical. For example, in Figure 6, we present three AMR graphs (A, B, C) representing the following sentences “*The cat sprints*”, “*The kitten runs*”, and “*The giraffe sleeps*”, respectively.

[1] The last released version swapped the `TOP` relation to the `TOP` attribute.

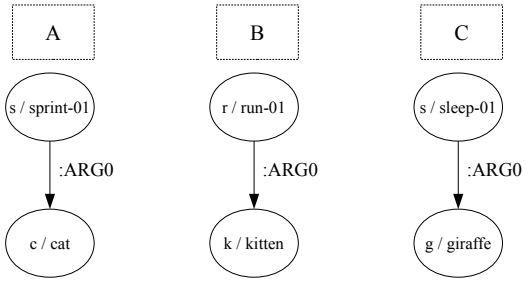

Figure 6 – Example of soft-match by the S^{2}MATCH metric.

From this figure, S^{2}MATCH produces 0.39, 0.25, and 0.25 when comparing (A, B), (B, C), and (A, C), respectively. Although the node labels are different, the authors claim that **A** is more similar to **B**, so the result of (A, B) is higher than (B, C) and (A, C). To capture these similarities and produce these results, S^{2}MATCH uses the 100-dimensional GloVe vectors (Pennington; Socher; Manning, 2014). In this way, when comparing the hypothesis graph in Figure 2 against the reference graph in Figure 3, S^{2}MATCH returns 0.45 f-score. Overall, S^{2}MATCH metric is more benevolent than SMATCH because of its soft-match computation.

SEMBLEU, in a different approach, extends the BLEU metric (Papineni *et al.*, 2002), defining the size of an AMR graph as the number of nodes and edges. This value is used to calculate the brevity penalty (BP) in BLEU, according to Equation 2.3:

$$BLEU = BP \cdot exp\left(\sum_{k=1}^{n} w_k \log p_k\right) \tag{2.3}$$

where w_k is the weight for matching k-grams and p_k is the precision. Moreover, SEMBLEU considers unigrams, bigrams, and trigrams as k-grams for matching and 1/3 as weight for each n-gram. Applying this equation to the previous example, SEMBLEU yields a value of 0.21 when comparing the hypothesis graph (Figure 2) against the reference graph (Figure 3).

Compared to SMATCH, SEMBLEU performs a higher-order comparison of AMR structures, as it considers longer sequences (k-grams) of AMR elements than SMATCH. It also does not suffer from the eventual errors introduced by the one-to-one node mapping that SMATCH performs.

SEMA adopts the same equations of SMATCH to compute precision and recall. However, comparing a hypothesis graph with a reference graph has two steps. In the first one, SEMA tries to match only the root node of the hypothesis with the root node of the reference graph. For example, when comparing the graph in Figure 2 (hypothesis) against the graph in Figure 3 (reference), SEMA scores only if the root nodes (`ask-01` and `want-01`) are the same. In the second one, for a triple to be counted as correct, the target node and at least a node with outgoing edges leading to the target node (in the hypothesis graph) must be in the reference graph. For example, comparing the graph **A** ("*They boy refused to go*") with the graph **B** ("*The boy wants to go*"), in Figure 7, SEMA computes as correct the triples `instance (b, boy)`, `instance (g, go-1)`, and `ARG0 (g, b)`, resulting in a 0.50 f-score. Considering the target node `boy`, there is a node (`go-1`) with an outgoing edge (`:ARG0`) leading to the target node in both graphs. That is, in both graphs, there is the subgraph `boy` $\xleftarrow{\text{:ARG0}}$ `go-01`.

Looking at the example of Figures 2 and 3, SEMA returns the result as in Figure 8. We can see that SEMA does not consider the `TOP` attribute in the triples and does not compute any triple as correct. Although there are `:ARG0` and `:ARG1` relations outgoing from the root nodes, there is neither the `girl` ← `ask-01` nor `leave-11` ← `ask-01` triples in the reference graph. Similarly, the reference graph has no `boy` ← `leave-11` triple.

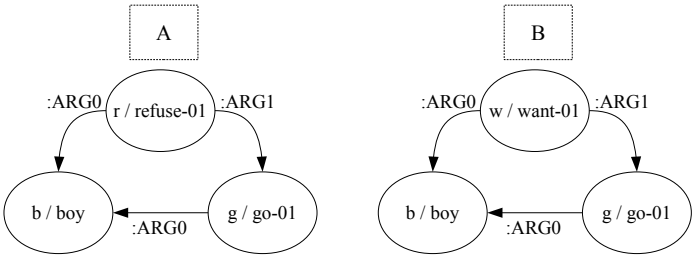

Figure 7 – Example of SEMA computation.

Triples of the hypothesis	Triples of the reference	SEMA score
instance (a, ask-01) ^	instance (x, want-01) ^	`P = 0 / 7 = 0.00`
instance (b, girl) ^	instance (y, boy) ^	`R = 0 / 6 = 0.00`
instance (c, leave-11) ^	instance (z, go-01) ^	`F1 = 0.00`
instance (d, boy) ^	ARG0 (x, y) ^	
ARG0 (a, b) ^	ARG1 (x, z) ^	
ARG1 (a, c) ^	ARG0 (z, y)	
ARG0 (c, d)		

Figure 8 – Triples and result produced by the SEMA metric.

SEMA is an intermediate metric in relation to SMATCH and SEMBLEU. It does not perform higher-order comparisons as SEMBLEU, but it is more rigorous than SMATCH in relation to the possible AMR graph matchings.

Overall, one may see that the metrics show different behaviors. It is important, therefore, to assess to what extent they are fair in their results. We do this by checking their correlation with human judgment to determine the best metric. In what follows, we detail our analyses and results.

3 ANALYSIS AND RESULTS

3.0.1 Corpus level evaluation

In this level, we replicated the experiment carried out by Song and Gildea (Song; Gildea, 2019). They compared SEMBLEU against SMATCH on the output of 4 systems: `JAMR` (Flanigan *et al.*, 2014), `CAMR` (Wang; Xue; Pradhan, 2015), `Gros` (Groschwitz *et al.*, 2018), and `Lyu` (Lyu; Titov, 2018) over 100 sentences from the test set of the LDC2015E86 corpus. In summary, the authors divided these four systems into two groups and asked 3 annotators to indicate the preferred group; the systems' final score corresponded to the number of times a group was preferred over the other by majority vote. For this study, we added the SEMA and S^{2}MATCH metrics to compare them with the same sentences and systems. We present the results in Table 1.

Table 1 – Results of the corpus level experiment.

Metric	**JAMR**	**CAMR**	**Gros**	**Lyu**
HUMAN	30	33	63	74
SEMBLEU	31	27	38	47
SMATCH	56	56	64	67
SEMA	39	41	49	54
S^{2}MATCH	59	58	65	69

The four systems may be viewed as two weak and two strong parsers, where, in this experiment, `JAMR` is the weakest and `Lyu` is the stronger. The human scores reflect this classification. Looking at these values, we may perform four analyses: (i) checking which metric discriminates the output of all systems,

(ii) verifying which metric discriminates the difference between the `CAMR` and `JAMR` systems, (iii) checking which metric discriminates the difference between the `Gros` and `CAMR` parsers, and (iv) verifying which metric discriminates the difference between the `Lyu` and `Gros` systems.

Analyzing the output of systems, SEMA is the only metric that discriminates all the systems concerning the other metrics, also discriminating against the small magnitude between `CAMR` and `JAMR` systems. On the other hand, SEMBLEU discriminates better between `Gros` and `CAMR` systems and between `Lyu` and `Gros`.

One can see that the SEMA and SEMBLEU metrics are more consistent with human judgments than SMATCH and S^{2}MATCH, as SEMA was able to distinguish between all the systems' output (as a human did) and between `CAMR` and `JAMR`, whereas SEMBLEU was able to distinguish between `Gros` and `CAMR` and between `Lyu` and `Gros`. SMATCH and S^{2}MATCH, on the other hand, were not able to distinguish the output of the parsers.

Overall, although SEMA was the only metric to distinguish all the output of systems, there is a match between SEMA and SEMBLEU, being important to go on in the evaluation, as follows.

We also replicated the bootstrap tests performed by Song and Gildea (Song; Gildea, 2019). They used bootstrap resampling (Koehn, 2004) to obtain 1,000 new datasets, each having 100 instances. Every dataset contains the references, four system outputs, and the corresponding human scores. Using the new datasets, the authors checked how frequently SEMBLEU and SMATCH are consistent with human judgments. As in the previous experiment, we added the SEMA and S^{2}MATCH metrics in this investigation and presented the achieved results in Table 2.

Table 2 – Bootstrap accuracies (%) for each system pair.

Metric	**CAMR** vs **JAMR**	**CAMR** vs **Gros**	**CAMR** vs **Lyu**	**JAMR** vs **Gros**	**JAMR** vs **Lyu**	**Gros** vs **Lyu**
SMATCH	67.9	99.9	100.0	100.0	100.0	90.3
SEMBLEU	68.4	99.9	100.0	100.0	100.0	90.9
SEMA	71.0	99.9	100.0	100.0	100.0	91.2
S^{2}MATCH	67.9	99.9	100.0	100.0	100.0	90.3

Overall, SEMA is equal to or slightly better than SEMBLEU, SMATCH, and S^{2}MATCH across all system pairs. However, the advantages are not significant at $p < 0.05$. As pointed out by Song and Gildea (Song; Gildea, 2019), this may be because of the small data size. Despite this, based on this experiment, SEMA is slightly better than other metrics, showing it may be more consistent with human evaluation.

Another experiment carried out by Song and Gildea (Song; Gildea, 2019) was at sentence-level. In this investigation, they calculated how many times metrics and humans chose the same output system. We did not replicate this experiment because we do not know which output the humans chose. On the other hand, we perform another sentence-level experiment. Our experiment investigates how metrics capture or behave with noisy sentences. In what follows, we detail our sentence-level experiment.

3.0.2 Sentence level evaluation

We performed a sentence-level experiment complementary to the corpus-level evaluation, including actual sentences in the setting. Although AMR metrics are predominantly used in automated parser evaluation by comparing two AMR graphs, these graphs reflect the meaning of sentences. Thus, when the meaning of a sentence changes, its respective AMR graph should also change. Therefore, a metric should be able to capture these changes (or noises) when comparing two AMR graphs.

To evaluate SMATCH, SEMBLEU, SEMA, and S^{2}MATCH metrics in the sentence level, we used the $1,527$ sentences from the AMR-annotated "The Little Prince" corpus in Portuguese (Anchiêta; Pardo, 2018). We also computed some statistics about the sentences and their respective graphs and presented them in Table 3. On average, the sentences have 8.31 tokens, producing 4 concepts and 4 relations.

Table 3 – Some statistics regarding the sentences and their graphs.

Statistics	Value
Avg. sentence length	8.31
Avg. number of concepts	4
Avg. number of relations	4

For each sentence, we created 3 noisy sentences from the original one and asked 5 annotators to rank these 3 sentences from the best (most similar to the original) to the worst sentence (less similar to the original)[1]. We chose sentences in Portuguese because the annotators were native speakers of this language.

To create the noisy sentences, we followed a systematic approach, as depicted in Figure 9. For the first sentence, we altered the predicate of the original sentence[2]; for the second, we changed the arguments of the predicate; and, at last, we removed the adjuncts of the sentence.

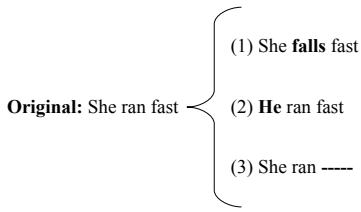

Figure 9 – Process to create noisy sentences.

The annotators produced ranking agreement greater than 80% for 1,070 sentences, i.e., at least 4 annotators agreed on the same rank. From these 1,070 sentences, 82.31% agreed that the worst sentence is the one for which we changed the predicate, 86.00% agreed that the best sentence is the one for which we removed the adjuncts, and 83.00% agreed that the medium sentence is the one for which we changed the arguments. Table 4 presents these agreement results. These 1,070 sentences were the ones we selected for testing the metrics.

Table 4 – Annotation agreement.

Noise	Agreement (%)		
	Worst	Medium	Best
Predicate	**82.31**	10.15	7.54
Arguments	10.11	**86.00**	3.89
Adjuncts	7.00	10.00	**83.00**

We manually developed AMR graphs for the noisy sentences to make our evaluation possible. Next, we compared the AMR graphs from the original sentences and the AMR graphs from the noisy sentences, using the Smatch, Sema, SemBleu, and S^{2}match metrics, producing an automatic ranking for each metric from their output scores. In Figure 10, we illustrate this approach.

We computed their output scores for each metric, comparing the gold AMR graph with the AMR graphs of the noisy sentences. As an illustration, for the above example, the annotators ranked the first (**1**), second (**2**), and third (**3**) sentences as the worst, medium, and best cases, respectively. The Smatch metric produced the f-scores 0.83, 0.67, and 0.80 for the (1), (2), and (3) sentences, respectively (i.e., this metric changed the best, medium, and worst sentences). SemBleu yielded 0.40, 0.57, and 0.51 (changing

[1] Notice that, in this evaluation, the annotators did not have access to the AMR annotation. Moreover, we asked annotators to evaluate the sentence similarity in meaning.

[2] We changed the predicate by a possible antonym, as this completely alters the meaning of the sentence.

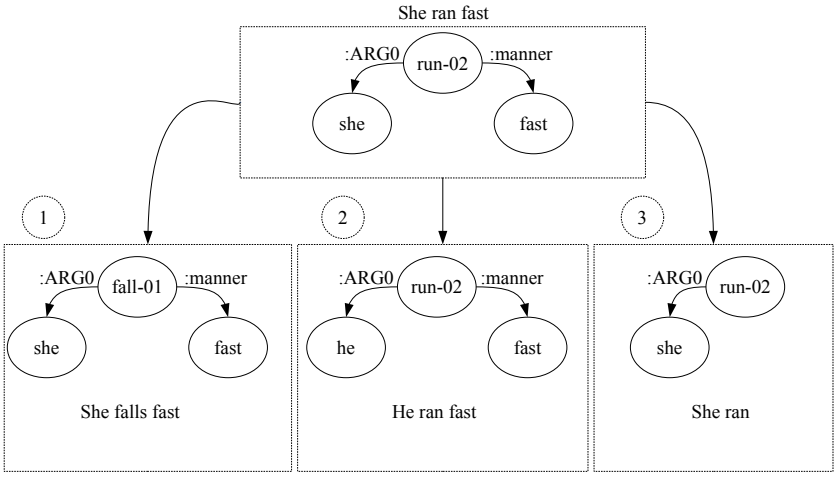

Figure 10 – Comparison between the gold AMR graph at the top and the AMR graphs at the bottom.

the best and medium sentence). SEMA produced 0.00, 0.60, and 0.75, respectively, agreeing with the manual ranking. S^{2}MATCH returned 0.70, 0.97, and 0.80, making the same confusion as SEMBLEU.

To compare the manual and automatic rankings, we used Kendall's tau coefficient (Kendall, 1938), which measures rank correlation. The correlation between the two ranks will be 1 when the ranks are identical, and -1 when the ranks are fully different, as defined by Equation 3.1:

$$\tau = \frac{C - D}{\binom{n}{2}} \tag{3.1}$$

where C is the number of concordant pairs, D is the number of discordant pairs, and $\binom{n}{2} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items. In Figure 11, we show the result of the Kendall's tau correlation for the $1,070$ sentences.

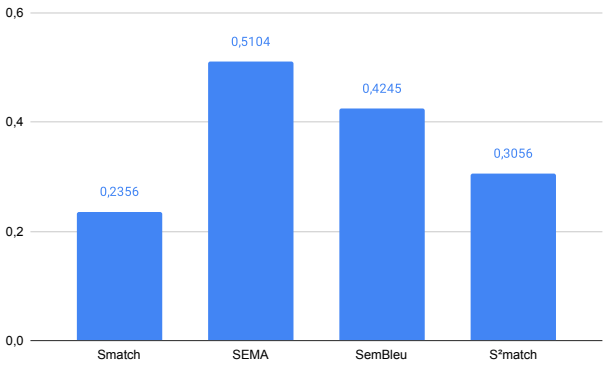

Figure 11 – Result of the Kendall's tau correlation.

One can see that the ranking produced by the SEMA and SEMBLEU metrics are the most similar to the manual ranking. In the same manner as the corpus level experiment, SEMA and SEMBLEU metrics are more consistent with manual judgments than the SMATCH metric, with SEMA being more consistent than SEMBLEU (now the difference among them is more evident than in the first experiment). SMATCH and S^{2}MATCH performed very poorly.

We also checked how often the metrics agreed with the manual ranking, as shown in Table 5. Analyzing the two best metrics, we see that SEMA agreed more with the sentences manually ranked as medium and best cases. The SEMBLEU metric agreed more with sentences considered the worst cases.

Table 5 – Agreement result concerning the manual ranking.

Metric	Sentence (#)		
	Worst	**Medium**	**Best**
SMATCH	1,020	1,017	1,018
SEMA	1,032	**1,035**	**1,046**
SEMBLEU	**1,042**	1,029	1,030
S^2MATCH	1,041	1,026	1,020

3.0.3 Graph level evaluation

At this level, we investigated AMR properties' inverse relation and reification. In the first, it is possible to invert two relations without losing the meaning of the sentence, while, in the second, it is possible to turn a relation into a concept. Figure 12 shows an example of inverse relation, where the `:ARG1` relation is converted into `:ARG1-of`, and Figure 13 presents an example of reification, where the `:location` relation is turned into `be-located-at-91` concept.

Such phenomena are sophisticated, and evaluating the metrics' discriminative power may reveal their true potential. Therefore, this may be a hard test for the metrics.

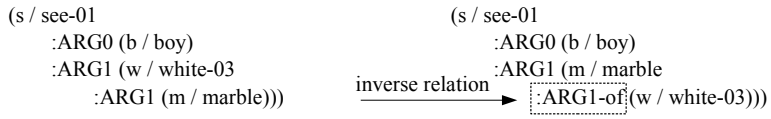

Figure 12 – An example of inverse relation for the sentence "*The boy sees that the marble is white*".

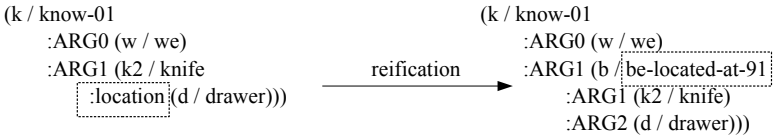

Figure 13 – An example of reification for the sentence "*We know the knife is in the drawer*".

To perform our analysis, we chose AMR graphs from The Little Prince corpus[3] and manually generated graph versions with inverse relations and reifications. Then, we used the metrics to evaluate the original graphs against those with inverse relation and reification. Table 6 presents the average of the obtained results.

Table 6 – Results for inverse relation and reification properties

Metric	Inverse relation	Reification
SMATCH	0.89	0.44
Sema	**1.00**	**0.58**
SEMBLEU	0.74	0.42
S^{2}MATCH	0.81	0.43

We may see that the SEMA metric achieves the best inverse relation and reification results. For that, we believe that the metric produces fairer results. The low values for reification may undervalue AMR parsing and annotation tasks since an AMR parser may generate a correct AMR graph and be penalized by the metrics (that would evaluate it as a bad quality one). To mitigate the low results in reification, it may be useful for metrics to normalize the AMR structures before evaluating them. Goodman (2019) (Goodman,

[3] https://amr.isi.edu/download/amr-bank-struct-v3.0.txt

2019) developed an AMR normalizer to transform a reified AMR into a non-reified AMR and vice-versa. We applied this normalizer to our reified version and reassessed the metrics. With this normalized version, SMATCH, SEMA, SEMBLEU and S^{2}MATCH achieved scores of 0.96, 0.94, 0.89, and 0, 92, respectively, improving their performances and making them fairer.

3.0.4 Other criteria to compare the metrics

Recently, Opitz et al. (Opitz; Parcalabescu; Frank, 2020) proposed seven criteria to evaluate and compare AMR metrics, as listed below. We suggest consulting the original paper (Opitz; Parcalabescu; Frank, 2020) for a complete description of the criteria.

1. **Continuity, non-negativity, and upper-bound.** This criterion verifies if a metric satisfies the following constraints: $metric : D \times D \rightarrow [0, 1]$; if graphs A and B are equivalent, $metric(A, B) \rightarrow 1$; if they are unrelated, $metric(A, B) \rightarrow 0$.
2. **Identity of indiscernibles**. This criterion states that $metric(A, B) = 1 \leftrightarrow A = B$.
3. **Symmetry**. This criterion guarantees $metric(A, B) = metric(B, A)$.
4. **Determinacy**. This criterion guarantees that repeated calculation over the same inputs yields the same score.
5. **No (low) bias**. This criterion checks if a metric favors correctness or penalizes errors for substructures of certain types in an unjustifiable or unintended manner.
6. **Symbolic graph matching**. This criterion verifies if the score of a metric increases when the overlap between two graphs increases.
7. **Graded graph matching**. This criterion checks if the metrics can properly distinguish AMR graphs with local variations.

The authors used these seven criteria to compare the SMATCH and SEMBLEU metrics, and, based on them, they developed S^{2}MATCH to meet these criteria. Here, besides summarizing the criteria, we replicated the comparison carried out by Opitz et al. (Opitz; Parcalabescu; Frank, 2020), added the SEMA metric in the analysis, and highlighted the found differences concerning the performed investigation by Opitz et al. (Opitz; Parcalabescu; Frank, 2020).

Table 7 shows our results by analyzing the seven criteria. The symbol ✓$_\epsilon$ means that the criterion was fulfilled with a small error, and ✓^{LEX} means that the criterion was fulfilled with the support of a lexical resource.

In the evaluation of Opitz et al. (Opitz; Parcalabescu; Frank, 2020), they found out that SMATCH, SEMBLEU, and S^{2}MATCH fulfill the first criterion. However, this does not occur with the SMATCH metric. For instance, Figure 14 presents two unrelated graphs, where graph **A** means "*Chapter 1*" and graph **B** means "*He described the mission as a failure*". When computing $Smatch(A, B)$, it returns an f-score of 0.20, violating that criterion. This issue occurs because SMATCH considers the `TOP` attribute in its computation. Removing this attribute solves this issue. For this example, the other metrics return 0.0 as a result.

In the second criterion, **identity of indiscernibles**, different from Opitz et al. (Opitz; Parcalabescu; Frank, 2020), we found out that SMATCH and S^{2}MATCH violate it. In Figure 15, we can see two graphs (**A** and **B**) that may mean "*The marble is white*". When evaluating $Smatch(A, B)$ and $S^2match(A, B)$ both produce 0.75, violating the criterion. To apply an AMR normalizer (Goodman, 2019) on these graphs may solve this problem.

Table 7 – Results of the criteria evaluation. $\checkmark_\epsilon$ fulfilled with small ϵ-error and $\checkmark^{LEX}$ fulfilled with a lexical resource.

Criterion	Smatch	SemBleu	S^{2}match	SEMA
Cont., non-neg, & upper-bound	✗	✓	✓	✓
Identity of indiscernibles	✗	✗	✗	✓
Symmetry	$\checkmark_\epsilon$	✗	$\checkmark_\epsilon$	✓
Determinacy	$\checkmark_\epsilon$	✓	$\checkmark_\epsilon$	✓
No/Low bias	✓	✗	✓	✓
Symbolic graph matching	✗	✗	✗	✗
Graded graph matching	✗	✗	$\checkmark^{LEX}$	✗

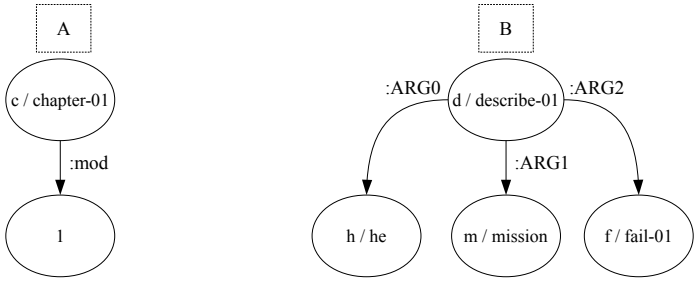

Figure 14 – Example of unrelated graphs.

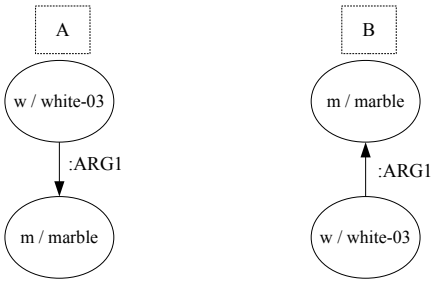

Figure 15 – Example of graphs for which SMATCH and S^{2}MATCH fail in the second criterion.

To analyze the other criteria, we follow Opitz et al. (Opitz; Parcalabescu; Frank, 2020) and found the same results for the **symmetry**, **determinacy**, and **no/low bias** criteria. Moreover, the SEMA metric fulfills these criteria without errors. On the other hand, the SEMA metric violates the symbolic graph matching and the graded graph matching criteria.

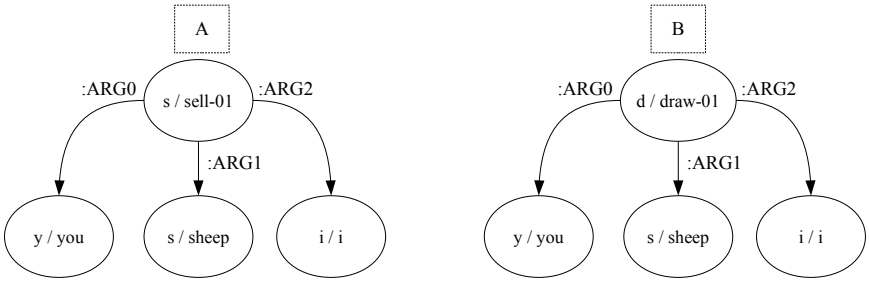

Figure 16 – Example of graphs for which SEMA fails in the sixth criterion.

Looking at the graphs **A** "*Sell me a sheep*" and **B** "*Draw me a sheep*", they share three nodes and three edges. Despite this, SEMA returns an f-score of 0.0. This occurs due to its computation strategy that

takes into account the nodes, which have outgoing edges leading to the node under analysis. SMATCH and S^{2}MATCH metrics also violate this criterion because they violate the identity of indiscernibles criterion.

Only the S^{2}MATCH metric fulfills this criterion through a lexical resource in the graded graph matching criterion. Although this criterion is linguistically motivated, we believe that the other metrics do not fulfill this criterion for two reasons: (i) AMR parsers will hardly produce a synonym of a concept because they have access to the lexicon of the sentence, and (ii) the AMR formalism does not use synonyms into its structure. Therefore, we believe that the last criterion is not a shortcoming of the metrics.

Based on all these experiments, we may conclude that SEMA is the most appropriate metric for evaluating AMR structures. It is the only one that is considered the best in all the tests or is close to a tie with some other metric (which varies depending on the experiment).

Besides showing each metric's limitations and potentialities, this paper also contributes to establishing a systematic and varied test set for new metrics that eventually arise in the area. We have used tests from the literature and also proposed two new ones. They may be useful strategies for assessing new proposals.

4 FINAL REMARKS

In this study, we conducted an in-depth investigation to know which AMR evaluation metric produces results most related to human judgment. Based on these studies, we showed that the SMATCH metric, the most popular AMR metric, shows severe limitations. We also found out that the SEMA metric may be considered the best one, followed by SEMBLEU in the first three experiments and by S^{2}MATCH in the fourth experiment.

These findings directly impact AMR-related investigations, as the evaluation metrics are used to assess annotation agreement and parsing results. Consequently, AMR-based applications may also be affected. Adopting a metric that produces over- or under-values may overlook several AMR structuring issues.

To the more interested reader, more information about this work may be found on the OPINANDO project webpage (https://sites.google.com/icmc.usp.br/opinando/).

REFERENCES

ANCHIÊTA, R.; PARDO, T. Towards AMR-BR: A SemBank for Brazilian Portuguese language. *In*: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. p. 974–979.

ANCHIêTA, R. T.; CABEZUDO, M. A. S.; PARDO, T. A. S. SEMA: an extended semantic evaluation metric for amr. *In*: **(To appear) Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing**. [*S.l.: s.n.*], 2019.

ANCHIêTA, R. T.; PARDO, T. A. S. A rule-based amr parser for portuguese. *In*: SIMARI, G. R. *et al.* (ed.). **Advances in Artificial Intelligence - IBERAMIA 2018**. [*S.l.: s.n.*], 2018. p. 341–353.

ARTZI, Y.; LEE, K.; ZETTLEMOYER, L. Broad-coverage CCG semantic parsing with AMR. *In*: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 1699–1710.

BANARESCU, L. *et al.* Abstract Meaning Representation for sembanking. *In*: **Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse**. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 178–186.

BOS, J. Squib: Expressive power of Abstract Meaning Representations. **Computational Linguistics**, v. 42, n. 3, p. 527–535, set. 2016.

CAI, S.; KNIGHT, K. 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)**. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 748–752.

DAMONTE, M.; COHEN, S. B.; SATTA, G. An incremental parser for abstract meaning representation. *In*: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**. [*S.l.: s.n.*], 2017. p. 536–546.

FLANIGAN, J. *et al.* 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)**. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 1426–1436.

GOODMAN, M. W. AMR Normalization for Fairer Evaluation. *In*: **Proceedings of the 33rd Pacific Asia Conference on Language, Information, and Computation**. Hakodate, Japan: [*S.l.: s.n.*], 2019. p. 37–46.

GROSCHWITZ, J. *et al.* AMR dependency parsing with a typed semantic algebra. *In*: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 1831–1841.

KENDALL, M. G. A New Measuere of Rank Correlation. **Biometrika**, v. 30, n. 1-2, p. 81–93, 1938.

KOEHN, P. Statistical significance tests for machine translation evaluation. *In*: **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 388–395.

LI, B. *et al.* Annotating the little prince with Chinese AMRs. *In*: **Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 7–15.

LYU, C.; TITOV, I. AMR parsing as graph prediction with latent alignment. *In*: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 397–407.

MATTHIESSEN, C.; BATEMAN, J. A. **Text generation and systemic-functional linguistics: experiences from English and Japanese**. [*S.l.: s.n.*]: Pinter Publishers, 1991.

MAY, J. SemEval-2016 task 8: Meaning representation parsing. *In*: **Proceedings of the 10th International Workshop on Semantic Evaluation**. San Diego, California: Association for Computational Linguistics, 2016. p. 1063–1073.

MAY, J.; PRIYADARSHI, J. SemEval-2017 task 9: Abstract meaning representation parsing and generation. *In*: **Proceedings of the 11th International Workshop on Semantic Evaluation**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 536–545.

MIGUELES-ABRAIRA, N.; AGERRI, R.; ILARRAZA, A. Diaz de. Annotating Abstract Meaning Representations for Spanish. *In*: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. Miyazaki, Japan: European Language Resources Association, 2018. p. 3074–3078.

NOORD, R. van; BOS, J. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. **Computational Linguistics in the Netherlands Journal**, v. 7, p. 93–108, 2017.

OEPEN, S. *et al.* (ed.). **Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning**. Hong Kong: Association for Computational Linguistics, 2019.

OPITZ, J.; PARCALABESCU, L.; FRANK, A. AMR similarity metrics from principles. **Transactions of the Association for Computational Linguistics**, v. 8, p. 522–538, 2020.

PAPINENI, K. *et al.* Bleu: a method for automatic evaluation of machine translation. *In*: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. *In*: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543.

SONG, L.; GILDEA, D. SemBleu: A robust metric for AMR parsing evaluation. *In*: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [*S.l.: s.n.*], 2019. p. 4547–4552.

WANG, C.; XUE, N.; PRADHAN, S. 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**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 366–375.

ZHANG, S. *et al.* AMR parsing as sequence-to-graph transduction. *In*: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [*S.l.: s.n.*], 2019. p. 80–94.