AMR Similarity Metrics from Principles

Juri Opitz and Letitia Parcalabescu and Anette Frank

Department for Computational Linguistics
Heidelberg University
69120 Heidelberg
{opitz,parcalabescu,frank}@cl.uni-heidelberg.de

Abstract

Different metrics have been proposed to compare Abstract Meaning Representation (AMR) graphs. The canonical SMATCH metric (Cai and Knight, 2013) aligns the variables of two graphs and assesses triple matches. The recent SEMBLEU metric (Song and Gildea, 2019) is based on the machine-translation metric BLEU (Papineni et al., 2002) and increases computational efficiency by ablating the variable-alignment. In this paper, i) we establish criteria that enable researchers to perform a *principled assessment of metrics* comparing meaning representations like AMR; ii) we undertake a *thorough analysis* of SMATCH and SEMBLEU where we show that the latter exhibits some undesirable properties. For example, it does not conform to the *identity of indiscernibles* rule and introduces biases that are hard to control; and iii) we propose a *novel metric* S^2MATCH that is more benevolent to only very slight meaning deviations and targets the fulfilment of all established criteria. We assess its suitability and show its advantages over SMATCH and SEMBLEU.

1 Introduction

Proposed in 2013, the aim of Abstract Meaning Representation (AMR) is to represent a sentence's meaning in a machine-readable graph format (Banarescu et al., 2013). AMR graphs are rooted, acyclic, directed, and edge-labeled. Entities, events, properties, and states are represented as *variables* that are linked to corresponding *concepts* (encoded as leaf nodes) via *is-instance* relations (cf. Figure 1, left). This structure allows us to capture complex linguistic phenomena such as coreference, semantic roles, or polarity.

When measuring the similarity between two AMR graphs A and B, for instance for the purpose of AMR parse quality evaluation, the metric of choice is usually SMATCH (Cai and Knight, 2013). Its backbone is an alignment-search between the graphs' variables. Recently, the SEMBLEU metric (Song and Gildea, 2019) has been proposed that operates on the basis of a variable-free AMR (Figure 1, right),[1] converting it to a bag of k-grams. Circumventing a variable alignment search reduces computational cost and ensures full determinacy. Also, grounding the metric in BLEU (Papineni et al., 2002) has a certain appeal, since BLEU is quite popular in machine translation.

However, we find that we are lacking a principled in-depth comparison of the properties of different AMR metrics that would help informing researchers to answer questions such as: *Which metric should I use to assess the similarity of two AMR graphs, e.g., in AMR parser evaluation? What are the trade-offs when choosing one metric over the other?* Besides providing criteria for such a principled comparison, we discuss a property that none of the existing AMR metrics currently satisfies: They do not measure graded meaning differences. Such differences may emerge because of near-synonyms such as *ruin – annihilate; skinny – thin – slim; enemy – foe* (Inkpen and Hirst, 2006; Edmonds and Hirst, 2002) or paraphrases such as *be able to – can; unclear – not clear*. In a classical syntactic parsing task, metrics do not need to address this issue because input tokens are typically projected to lexical concepts by lemmatization, hence two graphs for the same sentence tend not to disagree on the concepts projected from the input. This is different in semantic parsing where the projected concepts are often more abstract.

This article is structured as follows: We first establish *seven principles* that one may expect a metric for comparing meaning representations to

[1] Most research papers on AMR display the graphs in this ''shallow'' form. This increases simplicity and readability. (Lyu and Titov, 2018; Konstas et al., 2017; Zhang et al., 2019; Damonte and Cohen, 2019; Song et al., 2016).

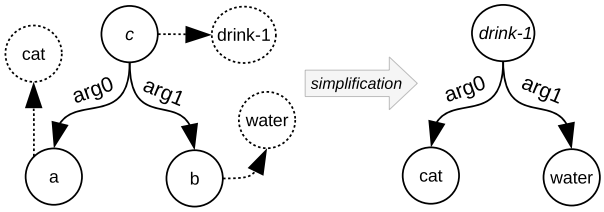


Figure 1: *A cat drinks water.* Simplified AMR graph and underlying deep form with *is-instance* relations (---▶) from variables (solid) to concepts (dashed).

satisfy, in order to obtain meaningful and appropriate scores for the given purpose (§2). Based on these principles we provide an *in-depth analysis* of the properties of the AMR metrics SMATCH and SEMBLEU (§3). We then *develop S²MATCH, an extension of SMATCH* that abstracts away from a purely symbolic level, allowing for a graded semantic comparison of atomic graph-elements (§4). By this move, we enable SMATCH to take into account fine-grained meaning differences. We show that our proposed metric retains valuable benefits of SMATCH, but at the same time is more benevolent to slight meaning deviations. Our code is available online at `https://github.com/Heidelberg-NLP/amr-metric-suite`.

2 From Principles to AMR Metrics

The problem of comparing AMR graphs $A, B \in \mathcal{D}$ with respect to the meaning they express occurs in several scenarios, for example, parser evaluation or inter-annotator agreement calculation (IAA). To measure the extent to which A and B agree with each other, we need a $metric$: $\mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ that returns a *score* reflecting *meaning distance* or *meaning similarity* (for convenience, we use similarity). Below we establish seven principles that seem desirable for this metric.

2.1 Seven Metric Principles

The first four metric principles are **mathematically motivated**:

I. continuity, non-negativity and upper-bound A similarity function should be continuous, with two natural edge cases: A, B are equivalent (maximum similarity) or unrelated (minimum similarity). By choosing 1 as upper bound, we obtain the following constraint on $metric$: $\mathcal{D} \times \mathcal{D} \rightarrow [0, 1]$.[2]

II. identity of indiscernibles This focal principle is formalized by $metric(A, B) = 1 \Leftrightarrow A = B$. It is violated if a metric assigns a value indicating equivalence to inputs that are not equivalent or if it considers equivalent inputs as different.

III. symmetry In many cases, we want a metric to be symmetric: $metric(A, B) = metric(B, A)$. A metric violates this principle if it assigns a pair of objects different scores when argument order is inverted. Together with principles I and II, it extends the scope of the metric to usages beyond parser evaluation, as it also enables sound IAA calculation, clustering, and classification of AMR graphs when we use the metric as a kernel (e.g., SVM). In parser evaluation, one may dispense with any (strong) requirements for symmetry—however, the metric must then be applied in a standardized way, with a fixed order of arguments.

In cases where there is no defined reference, the asymmetry could be handled by aggregating $metric(A, B)$ and $metric(B, A)$, for example, using the mean. However, it is open what aggregation is best suited and how to interpret results, for example, for $metric(A, B) = 0.1$ and $metric(B, A) = 0.9$.

IV. determinacy Repeated calculation over the same inputs should yield the same score. This principle is clearly desirable as it ensures reproducibility (a very small deviation may be tolerable).

The next three principles we believe to be desirable specifically when comparing meaning representation graphs such as AMR (Banarescu et al., 2013). The first two of the following principles are **motivated by computer science and linguistics**, whereas the last one is **motivated by a linguistic and an engineering perspective**.

V. no bias Meaning representations consist of nodes and edges encoding specific information types. Unless explicitly justified, a metric should not unjustifiably or in unintended ways favor correctness or penalize errors for specific substructures (e.g., leaf nodes). In case a metric favors or penalizes certain substructures more than others, in the interest of transparency, this should be made clear and explicit, and should be easily verifiable

[2] At some places in this paper, due to conventions, we project this score onto [0,100] and speak of *points*.

and consistent. For example, if we wish to give negation of the main predicate of a sentence a two-times higher weight compared with negation in an embedded sentence, we want this to be made transparent. A concrete example for a transparent bias is found in Cai and Lam (2019). They analyze the impact of their novel top–down AMR parsing strategy by integrating a root-distance bias into SMATCH to focus on structures situated at the top of a graph.

We now turn to properties that focus on the nature of the objects we aim to compare: graph-based compositional meaning representations. These graphs consist of atomic conditions that determine the circumstances under which a sentence is true. Hence, our $metric$ score should increase with increasing overlap of A and B, which we denote $f(A, B)$, the number of *matching* conditions. This overlap can be viewed from a **symbolic** or/and a **graded** perspective (cf., e.g., Schenker et al. [2005], who denote these perspectives as ''syntactic'' vs. ''semantic''). From the symbolic perspective, we compare the nodes and edges of two graphs on a symbolic level, while from the graded perspective, we take into account the degree to which nodes and edges differ. Both types of matching involve a precondition: If A and B contain variables, we need a variable-mapping in order to match conditions from A and B.[3]

VI. matching (graph-based) meaning representations – symbolic match A natural symbolic overlap-objective can be found in the Jaccard index J (Jaccard, 1912; Real and Vargas, 1996; Papadimitriou et al., 2010): Let $t(G)$ be the set of triples of graph G, $f(A, B) = |t(A) \cap t(B)|$ the size of the overlap of A, B, and $z(A, B) = |t(A) \cup t(B)|$ the size of their union. Then, we wish that A and B are considered more similar to each other than A and C iff A and B exhibit a greater relative agreement in their (symbolic) conditions: $metric(A, B) > metric(A, C) \Leftrightarrow \frac{f(A,B)}{z(A,B)} = J(A, B) > \frac{f(A,C)}{z(A,C)} = J(A, C)$. An allowed exception to this monotonic relationship can occur if we want to take into account a graded semantic match of atomic graph elements or sub-structures, which we will now elaborate on.

VII. matching (graph-based) meaning representations – graded semantic match: One motivation for this principle can be found in engineering, for example, when assessing the quality of produced parts. Here, small deviations from a reference may be tolerable within certain limits. Similarly, two AMR graphs may match almost perfectly—except for two small divergent components. The extent of divergence can be measured by the degree of similarity of the two divergent components. In our case, we need linguistic knowledge to judge what degree of divergence we are dealing with and whether it is tolerable.

For example, consider that graph A contains a triple $\langle x, \text{instance}, \textit{conceptA}\rangle$ and graph B a triple $\langle y, \text{instance}, \textit{conceptB}\rangle$, while otherwise the graphs are equivalent, and the alignment has set $x = y$. Then $f(A, B)$ should be higher when *conceptA* is similar to *conceptB* compared to the case where *conceptA* is dissimilar to *conceptB*. In AMR, concepts are often abstract, so near-synonyms may even be fully admissible (*enemy–foe*). Although such (near-)synonyms are bound to occur frequently when we compare AMR graphs of *different sentences* that may contain paraphrases, we will see, in Section §4, that this can also occur in parser evaluation, where two different graphs represent the *same sentence*. By defining $metric$ to map to a range [0,1] we already defined it to be globally graded. Here, we desire that *graded similarity* may also hold of *minimal units* of AMR graphs, such as atomic concepts or even sub-graphs, for example, to reflect that $injustice(x)$ is very similar to $justice(x) \wedge polarity(x, -)$.

2.2 AMR Metrics: SMATCH and SEMBLEU

With our seven principles for AMR similarity metrics in place, we now introduce SMATCH and SEMBLEU, two metrics that differ in their design and assumptions. We describe each of them in detail and summarize their differences, setting the stage for our in-depth metric analysis (§3).

Align and match – SMATCH The SMATCH metric operates in two steps. First, (i) we align the variables in A and B in the best possible way, by finding a mapping $map^{\star}$: $vars(A) \rightarrow vars(B)$ that yields a maximal set of matching triples between

[3] For example, consider graph A in Figure 1 and its set of triples $t(A)$: $\{\langle x_1, \text{instance}, \text{drink-1}\rangle \langle x_2, \text{instance}, \text{cat}\rangle, \langle x_1, \text{arg0}, x_2\rangle, \langle x_1, \text{arg1}, x_3\rangle, \langle x_3, \text{instance}, \text{water}\rangle\}$. When comparing A against graph B we need to judge whether a triple $\mathbf{t} \in t(A)$ is also contained in B: $\mathbf{t} \in t(B)$. For this, we need a mapping map: $vars(A) \rightarrow vars(B)$ where $vars(A) = \{x_1, .., x_n\}$, $vars(B) = \{y_1, .., y_m\}$ such that f is maximized.

A and B. For example, if $\langle x_i, \text{rel}, x_j \rangle \in t(A)$ and $\langle map^\star(x_i)\, \text{rel}, map^\star(x_j) \rangle = \langle y_k, \text{rel}, y_m \rangle \in t(B)$, we obtain one triple match. (ii) We compute Precision, Recall, and F1 score based on the set of triples returned by the alignment search. The NP-hard alignment search problem of step (i) is solved with a greedy hill-climber: Let $f_{map}(A, B)$ be the count of matching triples under any mapping function *map*. Then,

$$map^\star = \underset{map}{\text{argmax}}\, f_{map}(A, B) \quad (1)$$

Multiple restarts with different seeds increase the likelihood of finding better optima.

Simplify and match – SemBleu The SemBleu metric in Song and Gildea (2019) can also be described as a two-step procedure. But unlike Smatch it operates on a **variable-free reduction** of an AMR graph G, which we denote by G^{vf} (vf: variable-free, Figure 1, right-hand side).

In a first step, (i) SemBleu performs k-gram extraction from A^{vf} and B^{vf} in a breadth-first traversal (path extraction). It then (ii) adopts the Bleu score from MT (Papineni et al., 2002) to calculate an overlap score based on the extracted bags of k-grams:

$$\text{SemBleu} = BP \cdot \exp\left(\sum_{k=1}^{n} w_k \log p_k\right) \quad (2)$$

$$BP = e^{\min\left\{1-\frac{|B^{vf}|}{|A^{vf}|}, 0\right\}} \quad (3)$$

where p_k is Bleu's *modified k-gram precision* that measures k-gram overlap of a candidate against a reference: $p_k = \frac{|kgram(A^{vf}) \cap kgram(B^{vf})|}{|kgram(A^{vf})|}$. w_k is the (typically uniform) weight over chosen k-gram sizes. Sembleu uses NIST geometric probability smoothing (Chen and Cherry, 2014). The recall-focused ''brevity penalty'' BP returns a value smaller than 1 when the candidate length $|A^{vf}|$ is smaller than the reference length $|B^{vf}|$.

The graph traversal performed in SemBleu starts at the root node. During this traversal it simplifies the graph by replacing variables with their corresponding concepts (see Figure 1: the node c becomes Drink-01) and collects visited nodes and edges in uni-, bi- and tri grams ($k = 3$ is recommended). Here, a source node together with a relation and its target node counts as a bi-gram. For the graph in Figure 1, the extracted unigrams are $\{cat,\ water,\ drink\text{-}01\}$; the extracted bi grams are $\{drink\text{-}01\ arg1\ cat,\ drink\text{-}01 arg2\ water\}$.

Smatch vs. SemBleu in a nutshell SemBleu differs significantly from Smatch. A key difference is that SemBleu operates on reduced variable-free AMR graphs (G^{vf})—instead of full-fledged AMR graphs. By eliminating variables, SemBleu bypasses an alignment search. This makes the calculation faster and alleviates a weakness of Smatch: The hill-climbing search is slightly imprecise. However, SemBleu is not guided by aligned variables as anchors. Instead, SemBleu uses an *n*-gram statistic (Bleu) to compute an overlap score for graphs, based on *k-hop paths* extracted from G^{vf}, using the root node as the start for the extraction process. Smatch, by contrast, acts directly on variable-bound graphs matching triples based on a selected alignment. If in some application we wanted it, both metrics allow the capturing of more ''global'' graph properties: SemBleu can increase its k-parameter and Smatch may match conjunctions of (interconnected) triples. In the following analysis, however, we will adhere to their default configurations because this is how they are used in most applications.

3 Assessing AMR Metrics with Principles

This section evaluates Smatch and SemBleu against the seven principles we established above by asking: *Why does a metric satisfy or violate a given principle?* and *What does this imply?* We start with principles from mathematics.

I. Continuity, non-negativity, and upper-bound This principle is fulfilled by both metrics as they are functions of the form $metric : \mathcal{D} \times \mathcal{D} \rightarrow [0, 1]$.

II. Identity of indiscernibles This principle is fundamental: An AMR metric must return maximum score if and only if the graphs are equivalent in meaning. Yet, there are cases where SemBleu, in contrast to Smatch, does not satisfy this principle. Figure 2 shows an example.

Here, SemBleu yields a perfect score for two AMRs that differ in a single but crucial aspect: Two of its ARG$_x$ roles are filled with arguments that are meant to refer to distinct individuals that share the same concept. The graph on the left is an abstraction of, for example, *The man*$_1$ *sees the other man*$_2$ *in the other man*$_2$, while the graph on the right is an abstraction of *The man*$_1$ *sees himself*$_1$

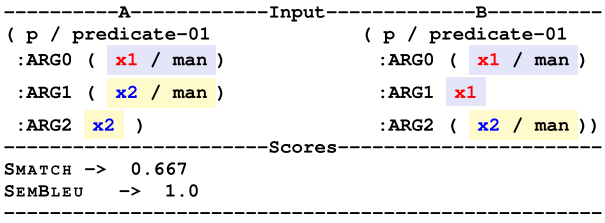


Figure 2: Two AMRs with semantic roles filled differently, SEMBLEU considers them as equivalent.

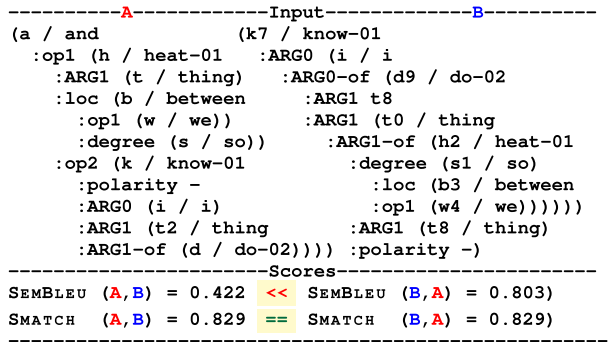


Figure 3: Symmetry violation for two parses of *Things are so heated between us, I don't know what to do.*

in the other man$_2$. SEMBLEU does not recognize the difference in meaning between a reflexive and a non-reflexive relation, assigning maximum similarity score, whereas SMATCH reflects such differences appropriately because it accounts for variables.

In sum, SEMBLEU does not satisfy principle II because it operates on a variable-free reduction of AMRs (G^{vf}). One could address this problem by reverting to canonical AMR graphs and adopting variable alignment in SEMBLEU. But this would adversely affect the advertised efficiency advantages over SMATCH. Re-integrating the alignment step would make SEMBLEU *less* efficient than SMATCH because it would add the complexity of breadth-first traversal, yielding a total complexity of $\mathcal{O}(\text{SMATCH})$ plus $\mathcal{O}(|V| + |E|)$.

III. Symmetry This principle is fulfilled if $\forall A, B \in \mathcal{D} : metric(A, B) = metric(B, A)$. Figure 3 shows an example where SEMBLEU does not comply with this principle, to a significant extent: When comparing AMR graph A against B, it yields a score greater than 0.8, yet, when comparing B to A the score is smaller than 0.5. We perform an experiment that quantifies this effect on a larger scale by assessing the frequency and the extent of such divergences. To this end, we parse 1,368 development sentences from an AMR corpus (LDC2017T10) with an AMR parser (obtaining graph bank $\mathcal{A}$) and evaluate it against another graph bank $\mathcal{B}$ (gold graphs or another parser-output). We quantify the symmetry violation by the *symmetry violation ratio* (Eq. 4) and the *mean symmetry violation* (Eq. 5) given some metric m:

$$svr = \frac{\sum_{i=1}^{|\mathcal{A}|} \mathbb{I}[m(\mathcal{A}_i, \mathcal{B}_i) \neq m(\mathcal{B}_i, \mathcal{A}_i)]}{|\mathcal{A}|} \quad (4)$$

$$msv = \frac{\sum_{i=1}^{|\mathcal{A}|} |m(\mathcal{A}_i, \mathcal{B}_i) - m(\mathcal{B}_i, \mathcal{A}_i)|}{|\mathcal{A}|} \quad (5)$$

	symmetry violation			
	svr (%, Δ>0.0001)		msv (in points)	
Graph banks	SMATCH	SEMBLEU	SMATCH	SEMBLEU
Gold ↔ GPLA	2.7	81.8	0.1	3.2
Gold ↔ CAMR	7.8	92.8	0.2	3.1
Gold ↔ JAMR	5.0	87.0	0.1	3.2
JAMR ↔ GPLA	4.2	86.0	0.1	3.0
CAMR ↔ GPLA	7.4	93.4	0.1	3.4
CAMR ↔ JAMR	7.9	91.6	0.2	3.3
avg.	5.8	88.8	0.1	3.2

Table 1: svr (Eq. 4), msv (Eq. 5) of AMR metrics.

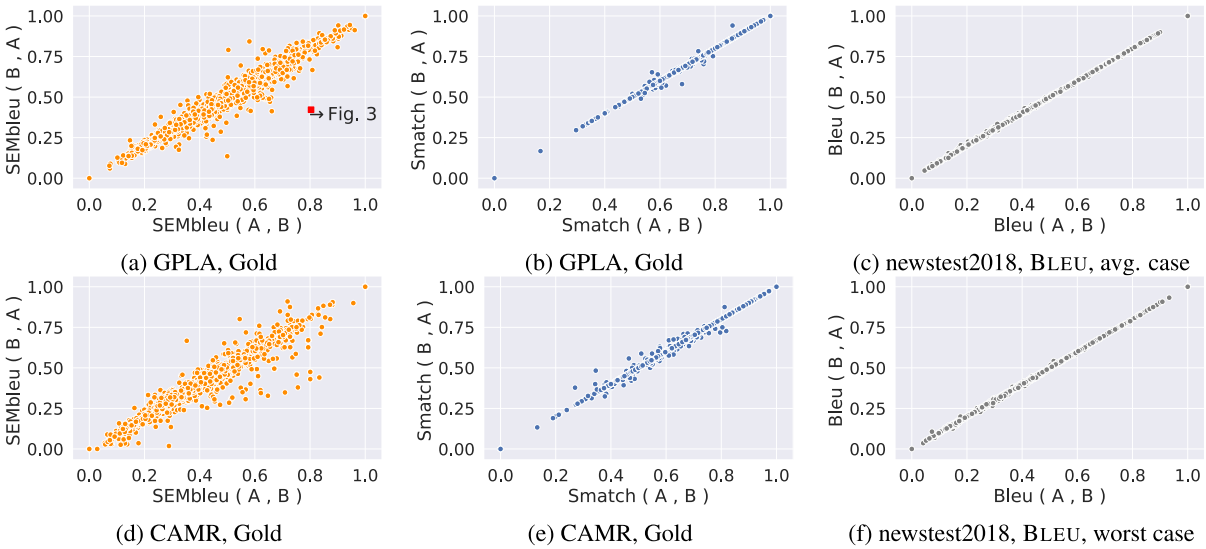
	BLEU symmetry violation, MT	
data: newstest2018 ↔ (·)	svr (%, Δ > 0.0001)	msv (in points)
worst-case	81.3	0.2
avg-case	72.7	0.2

Table 2: svr (Eq. 4), msv (Eq. 5) of BLEU, MT setting.

We conduct the experiment with three AMR systems, CAMR (Wang et al., 2016), GPLA (Lyu and Titov, 2018), and JAMR (Flanigan et al., 2014), and the gold graphs. Moreover, to provide a baseline that allows us to better put the results into perspective, we also estimate the symmetry violation of BLEU (SEMBLEU's MT ancestor) in an MT setting. Specifically, we fetch 16 system outputs of the WMT 2018 EN-DE metrics task (Ma et al., 2018) and calculate BLEU(A,B) and BLEU(B,A) of each sentence-pair (A,B) from the MT system's output and the reference (using the same smoothing method as SEMBLEU). As *worst-case*/*avg.-case*, we use the outputs from the team where BLEU exhibits maximum/median msv.[4]

Table 1 shows that more than 80% of the evaluated AMR graph pairs lead to a symmetry violation with SEMBLEU (as opposed to less than

[4]worst: LMU uns.; avg.: LMU NMT (Huck et al., 2017).



(a) GPLA, Gold (b) GPLA, Gold (c) newstest2018, BLEU, avg. case

(d) CAMR, Gold (e) CAMR, Gold (f) newstest2018, BLEU, worst case

Figure 4: Symmetry evaluations of metrics. SEMBLEU (left column) and SMATCH (middle column) and BLEU as a 'baseline' in an MT task setting on newstest2018. SEMBLEU: large divergence, strong outliers. SMATCH: few divergences, few outliers; BLEU: many small divergences, zero outliers. (a) marks the case in Figure 3.

10% for SMATCH). The msv of SMATCH is considerably smaller compared to SEMBLEU: 0.1 vs. 3.2 points F1 score. Even though the BLEU metric is inherently asymmetric, most of the symmetry violations are negligible when applied in MT (high svr, low msv, Table 2). However, when applied to AMR graphs ''via'' SEMBLEU the asymmetry is amplified by a factor of approximately 16 (0.2 vs. 3.2 points). Figure 4 visualizes the symmetry violations of SEMBLEU (left), SMATCH (middle), and BLEU (right). The SEMBLEU-plots show that the effect is widespread, some cases are extreme, many others are less extreme but still considerable. This stands in contrast to SMATCH but also to BLEU, which itself appears well calibrated and does not suffer from any major asymmetry.

In sum, symmetry violations with SMATCH are much fewer and less pronounced than those observed with SEMBLEU. In theory, SMATCH is fully symmetric, however, violations can occur due to alignment errors from the greedy variable-alignment search (we discuss this issue in the next paragraph). By contrast, the symmetry violation of SEMBLEU is intrinsic to the method because the underlying overlap measure BLEU is inherently asymmetric, however, this asymmetry is amplified in SEMBLEU compared to BLEU.[5]

IV. Determinacy This principle states that repeated calculations of a metric should yield identical results. Because there is no randomness in SEMBLEU, it fully complies with this principle. The reference implementation of SMATCH does not fully guarantee deterministic variable alignment results, because it aligns the variables by means of greedy hill-climbing. However, multiple random initializations together with the small set of AMR variables imply that the deviation will be $\leq \epsilon$ (a small number close to 0).[6] In Table 3 we measure the expected ϵ: it displays the SMATCH F1 standard deviation with respect to 10 independent runs, on a corpus level and on a graph-pair level (arithmetic mean).[7] We see that ϵ is small, even when only one random start is performed (corpus level: $\epsilon = 0.0003$, graph level: $\epsilon = 0.0013$). We conclude that the hill-climbing in SMATCH is unlikely to have any significant effects on the final score.

	# restarts				
	1	2	3	5	7
corpus vs. corpus	$2.6e^{-4}$	$1.7e^{-4}$	$8.1e^{-5}$	$5.7e^{-5}$	$5.6e^{-5}$
graph vs. graph	$1.3e^{-3}$	$1.0e^{-3}$	$8.5e^{-4}$	$5.3e^{-4}$	$4.0e^{-4}$

Table 3: Expected determinacy error ϵ in SMATCH F1.

V. No bias A similarity metric of (A)MRs should not unjustifiably or unintentionally favor

[5] As we show below (principle V), this is due to the way in which k-grams are extracted from variable-free AMR graphs.

[6] Additionally, $\epsilon = 0$ is guaranteed when resorting to a (costly) ILP calculation (Cai and Knight, 2013).

[7] Data: dev set of LDC2017T10, parses by GPLA.

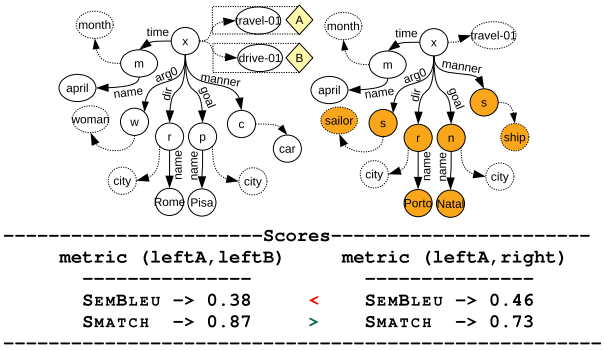


Figure 5: Left: *In April, a woman rides a car from Rome to Pisa.* root nodes A: *travel-01* vs. B: *drive-01*. Right: *In Apr., a sailor travels with a ship from P. to N.*

the correctness or penalize errors pertaining to any (sub-)structures of the graphs. However, we find that SemBleu is affected by a bias that affects (some) leaf nodes attached to high-degree nodes. The bias arises from two related factors: (i) when transforming G to G^{vf}, SemBleu replaces variable nodes with concept nodes. Thus, nodes that were leaf nodes in G can be raised to highly connected nodes in G^{vf}. (ii) breadth-first k-gram extraction starts from the root node. During graph traversal, concept leaves—now occupying the position of (former) variable nodes with a high number of outgoing (and incoming) edges—will be visited and extracted more frequently than others.

The two factors in combination make SemBleu penalize a wrong concept node harshly when it is attached to a high-degree variable node (the leaf is raised to high-degree when transforming G to G^{vf}). Conversely, correct or wrongly assigned concepts attached to nodes with low degree are only weakly considered.[8] For example, consider Figure 5. SemBleu considers two graphs that express quite distinct meanings (left and right) as more similar than graphs that are almost equivalent in meaning (left, variant A vs. B). This is because the leaf that is attached to the root is raised to a highly connected node in G^{vf} and thus is over-frequently contained in the extracted k-grams, whereas the other leaves will remain leaves in G^{vf}.

Analyzing and quantifying SemBleu's bias To better understand the bias, we study three limiting

[8]This may have severe consequences, e.g., for *negation*, since negation *always* occurs as a leaf in G and G^{vf}. Therefore, SemBleu, by-design, is benevolent to polarity errors.

	🌿	√	🌴
SemBleu	$\mathcal{O}(3d)$	$\mathcal{O}(d^2+d)$	$\mathcal{O}(d^2+2d)$
Smatch	$\mathcal{O}(d)$	$\mathcal{O}(d)$	$\mathcal{O}(d)$

Table 4: Error impact depending on error location in a tree with node degree d.

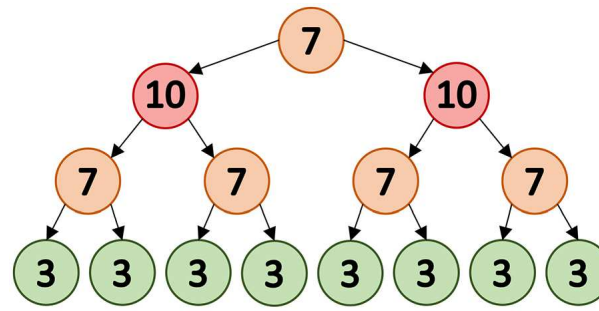


Figure 6: # of k-grams entered by a node in SemBleu.

cases: (i) the root is wrong (√) (ii) d leaf nodes are wrong (🌿) and (iii) one branching node is wrong (🌴). Depending on a specific node and its position in the graph, we would like to know onto how many k-grams (SemBleu) or triples (Smatch) the errors are projected. For the sake of simplicity, we assume that the graph always comes in its simplified form G^{vf}, that it is a tree, and that every non-leaf node has the same out-degree d.

The result of our analysis is given in Table 4[9] and exemplified in Figure 6. Both show that the number of times k-gram extraction visits a node heavily depends on its position and that with growing d, the bias gets amplified (Table 4).[10] For example, when $d = 3$, 3 wrong leaves yield 9 wrong k-grams, and 1 wrong branching node can already yield 18 wrong k-grams. By contrast, in Smatch the weight of d leaves always approximates the weight of 1 branching node of degree d.

In sum, in Smatch the impact of a wrong node is constant for all node types and rises linearly with d. In SemBleu the impact of a node rises approximately quadratically with d and it also depends on the node type, because it raises some (but not all) leaves in G to connected nodes in G^{vf}.

[9]Proof sketch, Smatch, d leaves: d triples, a root: d triples, a branching node: d+1 triples. $\text{SemBleu}_{k=3}^{w_k=1/3}$, d leaves: $3d$ k-grams (d tri, d bi, d uni). A root: d^2 tri, d bi, 1 uni. A branching node: d^2+d+1 tri, d+1 bi, 1 uni. □

[10]Consider that in AMR, d can be quite high, e.g., a predicate with multiple arguments and additional modifiers.

Eliminating biases A possible approach to reduce SEMBLEU's biases could be to weigh the extracted k-gram matches according to the degree of the contained nodes. However, this would imply that we assume some k-grams (and thus also some nodes and edges) to be of greater importance than others—in other words, we would eliminate one bias by introducing another. Because the breadth-first traversal is the metric's backbone, this issue may be hard to address well. When BLEU is used for MT evaluation, there is no such bias because the k-grams in a sentence appear linearly.

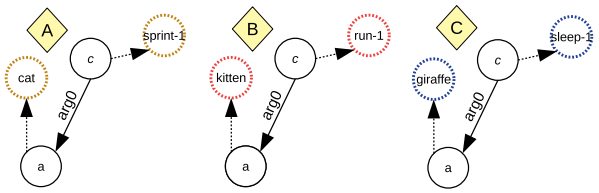
VI. Graph matching: Symbolic perspective This principle requires that a metric's score grows with increasing overlap of the conditions that are simultaneously contained in A and B. SMATCH fulfills this principle since it matches two AMR graphs inexactly (Yan et al., 2016; Riesen et al., 2010) by aligning variables such that the triple matches are maximized. Hence, SMATCH can be seen as a graph-matching algorithm that works on any pair of graphs that contain (some) nodes that are variables. It fulfills the Jaccard-based overlap objective, which symmetrically measures the amount of triples on which two graphs agree, normalized by their respective sizes (since SMATCH F1 = $2J/(1+J)$ is a monotonic relation).

Because SEMBLEU does not satisfy principles II and III (id. of indescernibles and symmetry), it is a corollary that it cannot fulfill the overlap objective.[11] Generally, SEMBLEU does not compare and match two AMR graphs per se, instead it matches the results of a graph-to-bag-of-paths projection function (§2.2) and the input may not be recoverable from the output (surjective-only). Thus, matching the outputs of this function cannot be equated to matching the inputs on a graph-level.

4 Towards a More *Semantic* Metric for *Semantic* Graphs: S²MATCH

This section focuses on principle VII, semantically graded graph matching, a principle that none of the AMR metrics considered so far satisfies. A

[11] Proof by symmetry violation:
w.l.o.g. $\exists A, B$: $metric(A,B) > metric(B,A) \Rightarrow f(A,B) > f(B,A) \rightarrow$ ϟ , since $f(A,B) = |t(A) \cap t(B)| = |t(B) \cap t(A)| = f(B,A)$ □ /// Proof by identity of indiscernibles:
w.l.o.g. $\exists\ A, B, C$: $metric(A,B) = metric(A,C) = 1 \wedge f(A,B)/z(A,B) = 1 > f(A,C)/z(A,C)$ϟ □



Scores		
metric (A,B)	metric (B,C)	metric (A,C)
SEMBLEU -> 0.00	SEMBLEU -> 0.00	SEMBLEU -> 0.00
SMATCH -> 0.25	SMATCH -> 0.25	SMATCH -> 0.25
S²MATCH -> 0.39	S²MATCH -> 0.25	S²MATCH -> 0.25

Figure 7: Three different AMR graphs representing *The cat sprints; The kitten runs; The giraffe sleeps* and pairwise similarity scores from SEMBLEU, SMATCH, and S²MATCH (see (§4) for S²Match).

fulfilment of this principle also increases the capacity of a metric to assess the semantic similarity of two AMR graphs from *different sentences*. For example, when clustering AMR graphs or detecting paraphrases in AMR-parsed texts, the ability to abstract away from concrete lexicalizations is clearly desirable. Consider Figure 7, with three different graphs. Two of them (A, B) are similar in meaning and differ significantly from C. However, both SMATCH and SEMBLEU yield the same result in the sense that $metric(A,B) = metric(A,C)$. Put differently, neither metric takes into account that *giraffe* and *kitten* are two quite different concepts, while *cat* and *kitten* are more similar. However, we would like this to be reflected by our metric and obtain $metric\ (A, B) > metric(A,C)$ in such a case.

S²MATCH We propose the S²MATCH metric (*Soft Semantic match*, pronounced: [estuːmætʃ]) that builds on SMATCH but differs from it in one important aspect: Instead of maximizing the number of (hard) triple matches between two graphs during alignment search, we maximize the (soft) triple matches by taking into account the semantic similarity of concepts. Recall that an AMR graph contains two types of triples: instance and relation triples (e.g., Figure 7, left: $\langle a, \text{instance}, \text{cat}\rangle$ and $\langle c, \arg 0, a\rangle$). In SMATCH, two triples can only be matched if they are identical. In S²MATCH, we relax this constraint, which has also the potential to yield a different, and possibly, a better variable alignment. More precisely, in SMATCH we match two instance triples

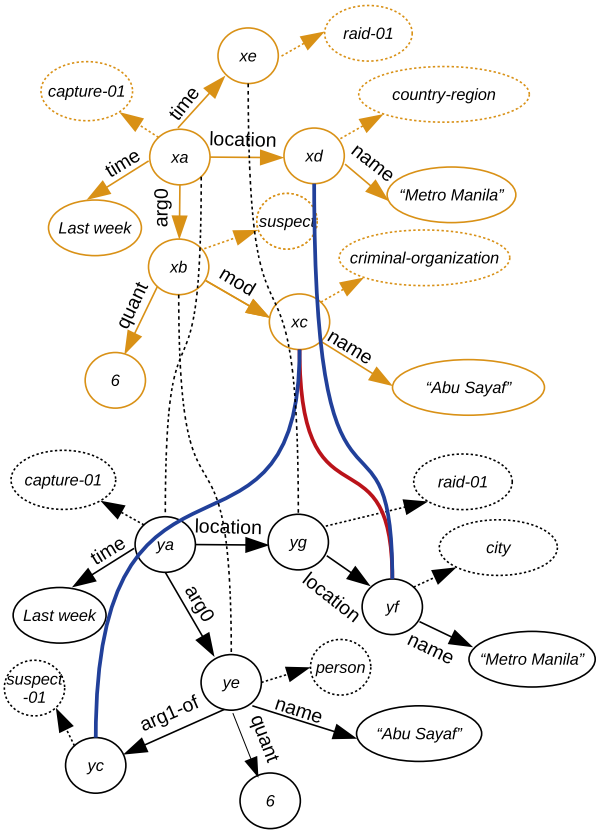


Figure 8: *'6 Abu Sayyaf suspects were captured last week in a raid in Metro Manila.'* gold (top) vs. parsed AMR (bottom). SMATCH aligns *criminal-organization* to *city* (red); S^{2}MATCH aligns *criminal-organization* to *suspect-01*, *city* to *country-region* (blue).

$\langle a, \text{instance}, \text{x} \rangle \in A$ and $\langle map(a), \text{instance}, y \rangle \in B$ as follows:

$$hardMatch = \mathbb{I}[\text{x} = \text{y}] \quad (6)$$

where $\mathbb{I}(c)$ equals 1 if c is true and 0 otherwise. S^{2}MATCH relaxes this condition:

$$softMatch = 1 - d(\text{x}, \text{y}), \quad (7)$$

where d is an arbitrary distance function $d : X \times X \rightarrow [0, 1]$. For example, in practice, if we represent the concepts as vectors $x, y \in \mathbb{R}^n$, we can use

$$d(x, y) = \min \left\{ 1, 1 - \frac{y^T x}{\|x\|_2 \, \|y\|_2} \right\}. \quad (8)$$

When plugged into Eq. 7, this results in the *cosine similarity* $\in [0, 1]$. It may be suitable to set a threshold τ (e.g., $\tau = 0.5$), to only consider the similarity between two concepts if it is above τ ($softMatch = 0$ if $1 - d(x, y) < \tau$). In the following pilot experiments, we use cosine (Eq. 8) and $\tau = 0.5$ over 100-dimensional GloVe vectors (Pennington et al., 2014).

		determinacy error		
	avg. msv (Eq. 5)	1 restart	2 restarts	4 restarts
SMATCH	0.0011	1.3e^{-3}	1.0e^{-3}	5.3e^{-4}
S^{2}MATCH	0.0005	9.0e^{-4}	6.1e^{-4}	2.1e^{-4}
relative change	−54.6%	−30.7%	−39.0%	−60.3%

Table 5: S^{2}MATCH improves upon SMATCH by reducing the extent of its non-determinacy.

To summarize, S^{2}MATCH is designed to either yield the same score as SMATCH—or a slightly increased score when it aligns concepts that are symbolically distinct but semantically similar. An example, from parser evaluation, is shown in Figure 8. Here, S^{2}MATCH increases the score to 63 F1 (+10 points) by detecting a more adequate alignment that accounts for the graded similarity of two related AMR concepts pairs. We believe that this is justified: The two graphs are very similar and an F1 of 53 is too low, doing the parser injustice.

On a technical note, the changes in alignments also have the outcome that S^{2}MATCH mends some of SMATCH's flaws: It better addresses principles III and IV, reducing the symmetry violation and determinacy error (Table 5).

Qualitative study: Probing S^{2}MATCH's choices This study randomly samples 100 graph pairs from those where S^{2}MATCH assigned higher scores than SMATCH.[12] Two annotators were asked to judge the similarity of all aligned concepts with similarity score <1.0: Are the concepts dissimilar, similar, or extremely similar? For concepts judged dissimilar, we conclude that S^{2}MATCH erroneously increased the score; if judged as (extremely) similar, we conclude that the decision was justified. We calculate three agreement statistics that all show large consensus among our annotators (Cohen's kappa: 0.79, squared kappa: 0.87, Pearson's ρ: 0.91) According to the annotations, the decision to increase the score is mostly justified: in 56% and 12% of cases both annotators voted that the newly aligned concepts are *extremely similar* and *similar*, respectively, while the agreed *dissimilar* label makes up 25% of cases.

[12] Automatic graphs by GPLA, on LDC2017T10, dev set.

input span region (excerpt)	amr region gold (excerpt)	amr region parser (excerpt)	cos	points F1↑	annotation
40 km southwest of	:quant 40:unit (k2 / kilometer)	(k22 / km :unit-of (d23 / distance-quantity	0.72	1.2	ex. similar
improving agricultural prod.	(i2 / improve-01 . . . :mod (f2 / farming)	(i31 / improve-01:mod (a23 / agriculture)	0.73	3.0	ex. similar
other deadly bacteria	op3 (b / bacterium . . . :mod (o / other)))	op3 (b13 / bacteria :ARG0-of:mod (o12 / other)))	0.80	5.1	ex. similar
drug and law enforcement aid	(a / and:op2 (a3 / aid-01	:ARG1 (a9 / and:op1 (d8 / drug) :op2 (l10 / law)))	0.67	1.8	similar
Get a lawyer and get a divorce.	:op1 (g / get-01:mode imp. :ARG0 (y / you)	:op1 (g0 / get-01 :ARG1 (l2 / lawyer):mode imp.)	0.80	4.8	dissimilar
The unusual development.	ARG0 (d / develop-01:mod (u / usual :polarity -))	:ARG0 (d1 / develop-02:mod (u0 / unusual))	0.60	14.0	dissimilar

Table 6: Examples where S^{2}MATCH assigns a higher score, accounting for the similarity of aligned concepts.

Table 6 lists examples of good or ill-founded score increases. We observe, for example, that S^{2}MATCH accounts for the similarity of two concepts of different number: *bacterium* (gold) vs. *bacteria* (parser) (line 3). It also captures abbreviations (*km* – *kilometer*) and closely related concepts (*farming* – *agriculture*). SEMBLEU and SMATCH would penalize the corresponding triples in exactly the same way as predicting a truly dissimilar concept.

An interesting case is seen in line 7. Here, *usual* and *unusual* are correctly annotated as dissimilar, since they are opposite concepts. S^{2}MATCH, equipped with GloVe embeddings, measures a cosine of 0.6, above the chosen threshold, which results in an increase of the score by 14 points (the increase is large as these two graphs are tiny). It is well known that synonyms and antonyms are difficult to distinguish with distributional word representations, because they often share similar contexts. However, the case at hand is orthogonal to this problem: *usual* in the gold graph is modified with the polarity '−', whereas the predicted graph assigned the (non-negated) opposite concept *unusual*. Hence, given the context in the gold graph, the prediction is semantically almost equivalent. This points to an aspect of principle VII that is not yet covered by S^{2}MATCH: It assesses graded similarity at the lexical, but not at the phrasal level, and hence cannot account for compositional phenomena. In future work, we aim to alleviate this issue by developing extensions that measure semantic similarity for larger graph contexts, in order to fully satisfy all seven principles.[13]

Quantitative study: Metrics vs. human raters This study investigates to what extent the judgments of the three metrics under discussion resemble human judgements, based on the following **two expectations**. First, the more a human rates two sentences to be semantically *similar* in their *meaning*, the higher the metric should rate the corresponding AMR graphs (**meaning similarity**). Second, the more a human rates two sentences to be *related* in their *meaning* (maximum: equivalence), the higher the score of our metric of the corresponding AMR graphs should tend to be (**meaning relatedness**). Albeit not the exact same (Budanitsky and Hirst, 2006), the tasks are closely related and both in conjunction should allow us to better assess the performance of our AMR metrics.

As ground truth for the **meaning similarity** rating task we use test data of the Semantic Textual Similarity **(STS)** shared task (Cer et al., 2017), with 1,379 sentence pairs annotated for meaning similarity. For the **meaning-relatedness** task we use **SICK** (Marelli et al., 2014) with 9,840 sentence pairs that have been additionally annotated for semantic relatedness.[14] We proceed as follows: We normalize the human ratings to [0,1]. Then we apply GPLA to parse the sentence tuples (s_i, s_i'), obtaining tuples $(parse(s_i), parse(s_i'))$ and score the graph pairs with the metrics: SMATCH(i), S^{2}MATCH(i), SEMBLEU(i), and H(i), where H(i) is the human score. For both tasks SMATCH and S^{2}MATCH yield better or equal correlations with human raters than SEMBLEU (Table 7). When considering the RMS error $\sqrt{n^{-1}\sum_{i=1}^{n}(H(i) - metric(i))^2}$. the difference is even more pronounced.

This deviation in the absolute scores is also reflected by the score density distributions plotted in Figure 9: SEMBLEU underrates a good proportion of graph pairs whose input sentences were rated as highly semantically similar or related by humans. This may well relate to the biases of different node types (cf. §3). Overall, S^{2}MATCH appears to provide

[13] As we have seen, this requires much care. We therefore consider this next step to be out of scope of the present paper.

[14] An example from SICK. Max. score: *A man is cooking pancakes–The man is cooking pancakes*. Min. score: *Two girls are playing outdoors near a woman.–The elephant is being ridden by the man.* To further enhance the soundness of the SICK experiment we discard pairs with a *contradiction* relation and retain 8,416 pairs with *neutral* or *entailment*.

	RMSE			RMSE (quant)			Pearson's ρ			Spearman's ρ		
task	SB	SM	S^2M	SB	SM	S^2M	SB	SM	S^2M	SB	SM	S^2M
STS	0.34	0.25	0.25	0.25	0.11	0.10	0.52	0.55	0.55	0.51	0.53	0.53
SICK	0.38	0.25	0.24	0.32	0.14	0.13	0.62	0.64	0.64	0.66	0.66	0.66

Table 7: RMSE (lower is better) and correlation results of our metrics in our **STS** and **SICK** investigations. RMSE (quant): RMSE on empirical quantile distribution with quantiles 0.1,0.2,. . . ,0.9.

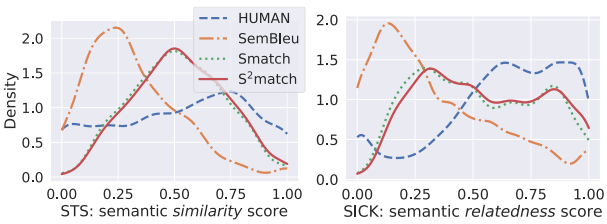


Figure 9: Sentence meaning similarity distributions.

a better fit with the score-distribution of the human rater when measuring **semantic similarity** and **relatedness**, the latter being notably closer to the human reference in some regions than the otherwise similar SMATCH. A concrete example from the STS data is given in Figure 10. Here, S^2MATCH detects the similarity between the abstract anaphors *it* and *this* and assigns a score that better reflects the human score compared to SMATCH and SEMBLEU, the latter being far too low. However, in total, we conclude that S^2MATCH's improvements seem rather small and no metric is perfectly aligned with human scores, possibly because gradedness of semantic similarity that arises in combination with constructional variation is not yet captured—more research is needed to extend S^2MATCH's scope to such cases.

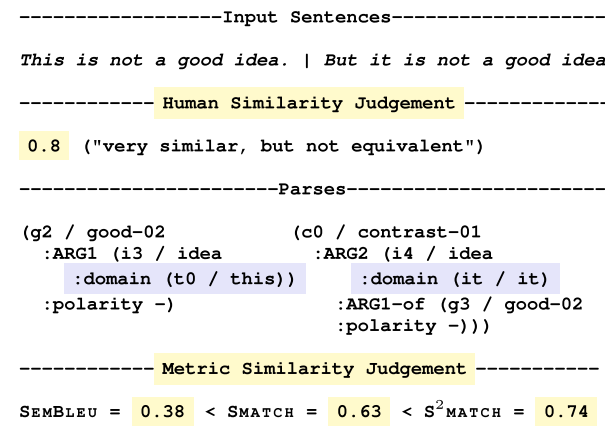


Figure 10: An example from STS, where S^2MATCH yields a score that better reflects the human judgement, due to detecting a similarity between the abstract anaphora *it* and *this*.

5 Metrics' effects on parser evaluation

We have seen that different metrics can assign different scores to the same pair of graphs. We now want to assess to what extent this affects rankings: Does one metric rank a graph higher or lower than the other? And can this affect the ranking of parsers on benchmark datasets?

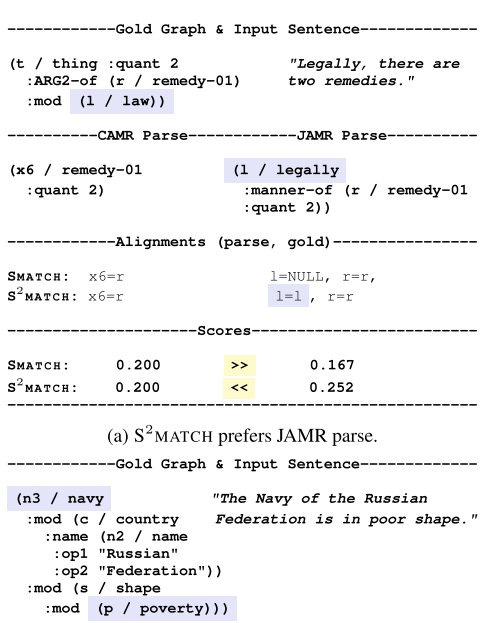
Quantitative study: Graph rankings In this experiment, we assess whether our metrics rank graphs differently. We use LDC2017T10 (dev) parses by CAMR $[c_1 \ldots c_n]$, JAMR $[j_1 \ldots j_n]$ and gold graphs $[y_1 \ldots y_n]$. Given metrics $\mathcal{F}$ and $\mathcal{G}$ we obtain results $\mathcal{F}^C := [\mathcal{F}(c_1, y_1) \ldots \mathcal{F}(c_n, y_n)]$ and analogously $\mathcal{F}^J$, $\mathcal{G}^C$ and $\mathcal{G}^J$. We calculate two statistics: (i) the ratio of cases i where the metrics differ in their preference for one parse over the other $(\mathcal{F}_i^J - \mathcal{F}_i^C) \cdot (\mathcal{G}_i^J - \mathcal{G}_i^C) < 0$, and, to assess significance, (ii) a t-test for paired samples on the differences assigned by the metrics between the parsers: $\mathcal{F}^J - \mathcal{F}^C$ and $\mathcal{G}^J - \mathcal{G}^C$. Table 8 shows that SMATCH and S^2MATCH both differ (significantly) from SEMBLEU in 15% – 20% of cases. SMATCH and S^2MATCH differ on individual rankings in appr. 4% of cases. Furthermore, we note a considerable amount of cases (8.1%) where SEMBLEU disagrees with itself in the preference for one parse over the other.[15]

The differing preferences of S^2MATCH for either candidate parse can be the outcome of small divergences due to the alignment search or because S^2MATCH accounts for the lexical similarity of concepts, perhaps supported by a new variable alignment. Figure 11 shows two examples where S^2MATCH prefers a different candidate parse compared to SMATCH. In the first example (Figure 11a), S^2MATCH prefers the parse

[15] That is, SB(A,G)>SB(B,G) albeit SB(G,A)<SB(G,B).

	$\text{SM}^{G}_{A,B}$	$\text{SM}^{A,B}_{G}$	$\text{SB}^{G}_{A,B}$	$\text{SB}^{A,B}_{G}$	$\text{S2M}^{G}_{A,B}$	$\text{S2M}^{A,B}_{G}$
$\text{SM}^{G}_{A,B}$	0.0	1.5	17.6†	19.0†	4.0	4.1
$\text{SM}^{A,B}_{G}$	–	0.0	17.9†	19.5†	3.9	4.0
$\text{SB}^{G}_{A,B}$	–	–	0.0	8.1†	18.4†	19.2†
$\text{SB}^{A,B}_{G}$	–	–	–	0.0	19.1†	19.3†
$\text{S2M}^{G}_{A,B}$	–	–	–	–	0.0	1.2
$\text{S2M}^{A,B}_{G}$	–	–	–	–		0.0

Table 8: Cross-metric comparison on individual graph rankings. Percentage of cases where metrics differ in their preference for one parse over the other. metric^{Y}_{X}: short for $\text{metric}(X, Y)$. † indicates significance in score differences assigned to parse pairs at $p<0.005$.



(a) S^2MATCH prefers JAMR parse.

(b) S^2MATCH prefers CAMR parse.

Figure 11: Two examples, where S^2MATCH disagrees with SMATCH in its preference of a candidate parse (for clarity, wiki-links are omitted in this display).

produced by JAMR and changes the alignment *legally-NULL* (SMATCH) to *legally-law* (S^2MATCH). In the second example 11b, S^2MATCH prefers the parse produced by CAMR, because it detects the similarity between *military* and *navy* and *poor* and *poverty*. Therefore, S^2MATCH can assess that the CAMR parse and the gold graph substantially agree on the root concept of the graph, which is not the case in the JAMR parse.

Quantitative study: Parser rankings Having seen that our metrics disagree on the ranking of individual graphs, we now quantify the effects on the ranking of parsers. We collect outputs of three state-of-art parsers on the test set of LDC2017T10: GPLA, a sequence-to-graph transducer (STOG), and a neural top-down parser (TOP-DOWN).

Table 9 shows that SMATCH and S^2MATCH agree on the ranking of all three parsers, but both disagree with SEMBLEU on the ranks of the 42^{nd} and 3^{rd} parser: unlike SEMBLEU, the SMATCH variants rate GPLA higher than TOP-DOWN. A factor that may have contributed to the different rankings perhaps lies in SEMBLEU's biases towards connected nodes: Compared with TOP-DOWN, GPLA delivers more complex parses, with more edges (avg. $|E|$: 32.8 vs. 32.1) and higher graph density (avg. density: 0.065 vs. 0.059). This is a nice property, because it indicates that the graphs of GPLA better resemble the rich gold graph structures (avg. density: 0.063, avg. $|E|$: 34.2). When inspecting this more closely, and looking at single (parse, gold) pairs, we observe further evidence for this: the structural error, in degree and density, is lower for GPLA than for TOP-DOWN (Table 9, right columns), with an error reduction of -27% (degree, 0.08 vs. 0.11) and -14% (density, 0.0067 vs. 0.0078).

	metric scores				structure error	
	SM	SB^G_A	SB^A_G	S^2M	node degree	graph density
STOG	$76.3_{\|1}$	$59.6_{\|1}$	$58.9_{\|1}$	$77.9_{\|1}$	0.08	0.0069
GPLA	$74.5_{\|2}$	$54.2_{\|3}$	$52.9_{\|3}$	$76.2_{\|2}$	0.08	0.0068
TOP-DOWN	$73.2_{\|3}$	$54.5_{\|2}$	$53.1_{\|2}$	$75.0_{\|3}$	0.11	0.0078

Table 9: Ranking parsers: STOG (Zhang et al.); GPLA (Lyu and Titov); TOP-DOWN (Cai and Lam, 2019). The structure error is defined as $\frac{1}{1371}\sum_{i=1}^{1371}|f(gold_i) - f(pred_i)|$, where f either is node degree or graph density. All four metrics differ significantly in their scores (paired t-test, p<0.05).

principle	SMATCH	SEMBLEU	S^2MATCH	Sec.
I. Cont., non-neg. & upper-bound	✓	✓	✓	-
II. id. of indescernibles	$✓_\epsilon$	✗	$✓_{\delta<\epsilon}$	§3
III. symmetry	$✓_\epsilon$	✗	$✓_{\delta<\epsilon}$	§3
IV. determinacy	$✓_\epsilon$	✓	$✓_{\delta<\epsilon}$	§3
V. low bias	✓	✗	✓	§3
VI. symbolic graph matching	✓	✗	✓	§3
VII. graded graph matching	✗	✗	$✓^{LEX}$	§4

Table 10: Evaluation of three AMR metrics using our seven principles. $✓_\epsilon$: fulfilled with a very small ϵ-error.

In sum, by building graphs of higher complexity, GPLA takes a greater risk when attaching wrong concepts to connected nodes where errors are penalized more strongly by SEMBLEU than SMATCH, according to the biases we have studied in §3 (Table 4). In that sense, STOG also takes more risks, but it may get more of such concepts right and so the bias transitions from penalty to reward, potentially explaining the large performance Δ (+6) of STOG to the other parsers, as measured by SEMBLEU, in contrast to S(2)MATCH (Δ: +2).

6 Summary of Our Metric Analyses

Table 10 summarizes our analyses' integral results. Principle I is fulfilled by all metrics as they exhibit *continuity, non-negativity and an upper bound*. Principle II, however, is not satisfied by SEMBLEU because it can mistake two graphs of different meaning as equivalent. This is because it ablates a variable-alignment and therefore cannot capture facets of coreference. Yet, a positive outcome of this is that it is *fast to compute*. This could make it first choice in some recent AMR parsing approaches that use reinforcement learning (Naseem et al., 2019), where rapid feedback is needed. It also marks a point by fully satisfying Principle IV, yielding fully deterministic results. SMATCH, by contrast, either needs to resort to a costly ILP solution or (in practice) uses hill-climbing with multiple restarts to reduce divergence to a negligible amount.

A central insight brought out by our analysis is that SEMBLEU exhibits *biases* that are hard to control. This is caused by two (interacting) factors: (i) The extraction of k-grams is applied on the graph top to bottom and visits some nodes more frequently than others. (ii) It raises some (but not all) leaf nodes to connected nodes, and these nodes will be overly frequently contained in extracted k-grams. We have shown that these two factors in combination lead to large biases that researchers should be aware of when using SEMBLEU (§3). Its ancestor BLEU does not suffer from such biases since it extracts k-grams linearly from a sentence.

Given that SEMBLEU is built on BLEU, it is inherently *asymmetric*. However, we have shown that the asymmetry (Principle III) measured for BLEU in MT is amplified by SEMBLEU in AMR, mainly due to the biases it incurs (Principle V). Although asymmetry can be tolerated in parser evaluation if outputs are compared against gold graphs in a standardized manner, it is difficult to

apply an asymmetric metric to measure IAA or to compare parses for detecting paraphrases, or in tri-parsing, where no reference is available. If the asymmetry is amplified by a bias, it becomes harder to judge the scores. Finally, considering that SEMBLEU does not match AMR graphs on the graph-level but matches extracted bags-of-k-grams, it turns out that it cannot be categorized as a graph matching algorithm as defined in Principle VI.

Principle VI is fulfilled by SMATCH without any transformation on AMR graphs. It searches for an optimal variable alignment and counts matching triples. As a corollary, it fulfills principles I, II, III and V. The fact that SMATCH fulfills all but one principle backs up many prior works that use it as sole criterion for IAA and parse evaluation.

Our principles also helped us detect a weakness of all present AMR metrics: They operate on a discrete level and cannot assess graded meaning differences. As a first step, we propose S^{2}MATCH: It preserves beneficial properties of SMATCH but is benevolent to slight lexical meaning deviations. Besides parser evaluation, this property makes the metric also more suitable for other tasks, for example, it can be used as a kernel in an SVM that classifies AMRs to determine whether two sentences are equivalent in meaning. In such a case, S^{2}MATCH is bound to detect meaning-similarities that cannot be captured by SMATCH or SEMBLEU, for example, due to paraphrases being projected into the parses.

7 Related Work

Developing similarity metrics for meaning representations (MRs) is important, as it, inter alia, affects semantic parser evaluation and computation of IAA statistics for sembanking. MRs are designed to represent the meaning of text in a well-defined, interpretable form that is able to identify meaning differences and support inference. Bos (2016, 2019) has shown how AMR can be translated to FOL, a well-established MR formalism. Discourse Representation Theory (DRT; Kamp, 1981; Kamp and Reyle, 1993) is based on and extends FOL to discourse representation. A recent shared task on DRS parsing used the COUNTER metric (Abzianidze et al., 2019; Evang, 2019), an adaption of SMATCH, underlining SMATCH's general applicability. Its extension S^{2}MATCH may also prove beneficial for DRS.

Other research into AMR metrics aims at making the comparison fairer by normalizing graphs (Goodman, 2019). Anchiêta et al. (2019) argue that one should not, for example, insert an extra *is-root* node when comparing AMR graphs (as done in SEMBLEU and SMATCH). Damonte et al. (2017) extend SMATCH to analyze individual AMR facets (co-reference, WSD, etc.). Cai and Lam (2019) adapt SMATCH to analyze their parser's performance in predicting triples that are in close proximity to the root. Our metric S^{2}MATCH allows for straightforward integration of these approaches.

Computational AMR tasks Since the introduction of AMR, many AMR-related tasks have emerged. Most prominent is AMR parsing (Wang et al., 2015, 2016; Damonte et al., 2017; Konstas et al., 2017; Lyu and Titov, 2018; Zhang et al., 2019). The inverse task generates text from AMR graphs (Song et al., 2017, 2018; Damonte and Cohen, 2019). Opitz and Frank (2019) rate the quality of automatic AMR parses without costly gold data.

8 Conclusion

We motivated seven principles for metrics measuring the similarity of graph-based (Abstract) Meaning Representations, from mathematical, linguistic and engineering perspectives. A metric that fulfills all principles is applicable to a wide spectrum of cases, ranging from parser evaluation to sound IAA calculation. Hence **(i) our principles can inform (A)MR researchers who desire to compare and select among metrics**, and **(ii) they ease and guide the development of new metrics**.

We provided examples for both scenarios. We showcased (i) by utilizing our principles as guidelines for an in-depth analysis of two AMR metrics: SMATCH and the recent SEMBLEU metrics, two quite distinct approaches. Our analysis uncovered that the latter does not satisfy some principles, which might reduce its safety and applicability. In line with (ii), we target the fulfilment of all seven principles and propose S^{2}MATCH, a metric that accounts for graded similarity of concepts as atomic graph components. In future work, we aim for a metric that accounts for graded compositional similarity of subgraphs.

Acknowledgments

We are grateful to the anonymous reviewers and the action editors for their valuable time and comments. This work has been partially funded by DFG within the project *ExpLAIN. Between the Lines – Knowledge-based Analysis of Argumentation in a Formal Argumentation Inference System*; FR 1707/-4-1, as part of the RATIO Priority Program; and by the the *Leibniz ScienceCampus Empirical Linguistics & Computational Language Modeling*, supported by Leibniz Association grant no. SAS2015-IDS-LWC and by the Ministry of Science, Research, and Art of Baden-Wurttemberg.

References

Lasha Abzianidze, Rik van Noord, Hessel Haagsma, and Johan Bos. 2019. The first shared task on discourse representation structure parsing. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden.

Rafael Torres Anchiêta, Marco Antonio Sobrevilla Cabezudo, and Thiago Alexandre Salgueiro Pardo. 2019. SEMA: An extended semantic evaluation for AMR. In *(To appear) Proceedings of the 20th Computational Linguistics and Intelligent Text Processing*. Springer International Publishing.

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.

Johan Bos. 2016. Expressive power of abstract meaning representations. *Computational Linguistics*, 42(3):527–535.

Johan Bos. 2019. Separating argument structure from logical structure in AMR. *arXiv preprint arXiv:1908.01355*.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Deng Cai and Wai Lam. 2019. Core semantic first: A top-down approach for AMR parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3797–3807, Hong Kong, China.

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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA.

Marco Damonte and Shay B. Cohen. 2019. Structural neural encoders for AMR-to-text Generation. 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 3649–3658, Minneapolis, Minnesota.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. 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*, pages 536–546, Valencia, Spain.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Kilian Evang. 2019. Transition-based DRS parsing using stack-LSTMs. In *Proceedings of*

the IWCS Shared Task on Semantic Parsing, Gothenburg, Sweden.

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.

Michael Wayne Goodman. 2019. AMR normalization for fairer evaluation. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information, and Computation*. Hakodate.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017. LMU Munich's neural machine translation systems for news articles and health information texts. In *Proceedings of the Second Conference on Machine Translation*, pages 315–322, Copenhagen, Denmark.

Diana Inkpen and Graeme Hirst. 2006. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, 32(2):223–262.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Hans Kamp. 1981. A theory of truth and semantic representation. *Formal Semantics: The Essential Readings*, pages 189–222.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic. Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, Kluwer, Dordrecht.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 146–157.

Chunchuan Lyu and Ivan Titov. 2018. 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)*, pages 397–407, Melbourne, Australia.

Qingsong Ma, Ondrej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 671–688.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland.

Tahira Naseem, Abhishek Shah, Hui Wan, Radu Florian, Salim Roukos, and Miguel Ballesteros. 2019. Rewarding smatch: Transition-based AMR parsing with reinforcement learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4586–4592, Florence, Italy.

Juri Opitz and Anette Frank. 2019. Automatic Accuracy Prediction for AMR Parsing. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 212–223, Minneapolis, Minnesota.

Panagiotis Papadimitriou, Ali Dasdan, and Hector Garcia-Molina. 2010. Web graph similarity for anomaly detection. *Journal of Internet Services and Applications*, 1(1):19–30.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Raimundo Real and Juan M Vargas. 1996. The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, 45(3):380–385.

Kaspar Riesen, Xiaoyi Jiang, and Horst Bunke. 2010. Exact and inexact graph matching: Methodology and applications, *Managing and Mining Graph Data*, Springer, pages 217–247.

Adam Schenker, Horst Bunke, Mark Last, and Abraham Kandel. 2005. *Graph-Theoretic Techniques for Web Content Mining*, World Scientific Publishing Co., Inc., USA.

Linfeng Song and Daniel Gildea. 2019. SemBleu: A robust metric for AMR parsing evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4552, Florence, Italy.

Linfeng Song, Xiaochang Peng, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2017. AMR-to-text generation with synchronous node replacement grammar. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 7–13, Vancouver, Canada.

Linfeng Song, Yue Zhang, Xiaochang Peng, Zhiguo Wang, and Daniel Gildea. 2016. AMR-to-text generation as a traveling salesman problem. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2084–2089, Austin, Texas.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia.

Chuan Wang, Sameer Pradhan, Xiaoman Pan, Heng Ji, and Nianwen Xue. 2016. CAMR at SemEval-2016 task 8: An extended transition-based AMR parser. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1173–1178, San Diego, California.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. Boosting transition-based AMR parsing with refined actions and auxiliary analyzers. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 857–862, Beijing, China.

Junchi Yan, Xu-Cheng Yin, Weiyao Lin, Cheng Deng, Hongyuan Zha, and Xiaokang Yang. 2016. A short survey of recent advances in graph matching. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 167–174. ACM.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy.