Better Smatch = Better Parser? AMR evaluation is not so simple anymore

Juri Opitz
Dept. of Computational Linguistics
Heidelberg University
69120 Heidelberg
opitz.sci@gmail.com

Anette Frank
Dept. of Computational Linguistics
Heidelberg University
69120 Heidelberg
frank@cl.uni-heidelberg.de

Abstract

Recently, astonishing advances have been observed in AMR parsing, as measured by the structural SMATCH metric. In fact, today's systems achieve performance levels that seem to surpass estimates of human inter annotator agreement (IAA). Therefore, it is unclear how well SMATCH (still) relates to human estimates of parse quality, as in this situation potentially fine-grained errors of similar weight may impact the AMR's meaning to different degrees.

We conduct an analysis of two popular and strong AMR parsers that – according to SMATCH– reach quality levels on par with human IAA, and assess how human quality ratings relate to SMATCH and other AMR metrics. Our main findings are: i) While high SMATCH scores indicate otherwise, we find that **AMR parsing is far from being solved**: we frequently find structurally small, but semantically unacceptable errors that substantially distort sentence meaning. ii) Considering high-performance parsers, **better SMATCH scores may not necessarily indicate consistently better parsing quality**. To obtain a meaningful and comprehensive assessment of quality differences of parse(r)s, we recommend augmenting evaluations with macro statistics, use of additional metrics, and more human analysis.

1 Introduction

Abstract Meaning Representation (AMR), proposed by Banarescu et al. (2013), aims at capturing the meaning of texts in an explicit graph format. Nodes describe *entities, events*, and *states*, while edges express key semantic relations, such as ARG_x (indicating semantic roles as in PropBank (Palmer et al., 2005)), or *instrument* and *cause*.

Albeit the development of parsers can be driven by multiple desiderata, better performance on benchmarks often serves as main criterion. For AMR, this goal is typically measured using SMATCH (Cai and Knight, 2013) against a refer-
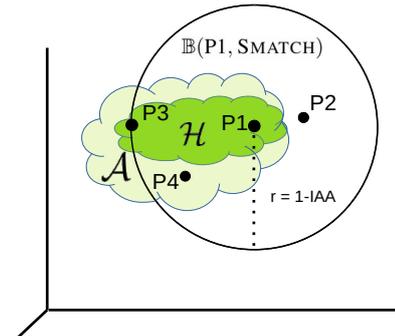

Figure 1: Sketch of AMR IAA ball. The center (P1) is a reference AMR, while P2, P3, P4 are candidates. Any AMR x from the ball has high structural SMATCH agreement with P1, i.e., $\text{SMATCH}(x, P1) \geq$ estimated human IAA. However, they may fall in different categories: $\mathcal{H}$ (green cloud) contains correct AMR alternatives. Its superset $\mathcal{A}$ (light cloud) contains acceptable AMRs that may misrepresent the sentence meaning up to a minor degree. Other parses from the ball, e.g., P2, mis-represent the sentence's meaning – despite possibly having higher SMATCH agreement with the reference than all other candidates.

ence corpus. The metric measures to what extent the reference has been reconstructed by the parser.

However, thanks to astonishing recent advances in AMR parsing, mainly powered by the *language modeling and fine-tuning paradigm* (Bevilacqua et al., 2021), parsers now achieve benchmark scores that surpass IAA estimates.[1] Therefore, it is difficult to assess whether (fine) differences in SMATCH scores i) can be attributed to minor but valid divergences in interpretation or AMR structure, as they may also occur in human assessments, or ii) if they constitute significant meaning distorting errors.

This fundamental issue is outlined in Figure 1. Four parses are located in the ball $\mathbb{B}(P1, \text{SMATCH})$

---

[1] Banarescu et al. (2013) find that an (optimistic) average annotator vs. consensus IAA (SMATCH) was 0.83 for newswire and 0.79 for web text. When newly trained annotators doubly annotated web text sentences, their annotator vs. annotator IAA was 0.71. Recent BART and T5 based models range between 0.82 and 0.84 SMATCH F1 scores.

of estimated IAA, (gold) parse P1 being the center. However, the true set of possible human candidates $\mathcal{H}$ is very likely much smaller than the ball and its shape is unknown.[2] Besides, a superset of $\mathcal{H}$ is a set of *acceptable* parses $\mathcal{A}$, i.e., parses that may have a small flaw which does not significantly distort the sentence meaning. Now, it can indeed happen that parse P2, as opposed to P3, has a lower distance to reference P1, i.e., to the center of $\mathbb{B}(\textsc{Smatch})$ – but is not found in $\mathcal{A} \supseteq \mathcal{H}$, which marks it as an inaccurate candidate. On the other hand, P4 is contained in $\mathcal{A}$, but not in $\mathcal{H}$, which would make it acceptable, but less preferable than P3.

Research questions   Triggered by these considerations, this paper tackles the key questions: *Do high-performance AMR parsers indeed deliver accurate semantic graphs, as suggested by high benchmark scores that surpass human IAA estimates*? *Does a higher* SMATCH *against a single reference necessarily indicate better overall parse quality*? And *what steps can we take to mitigate potential issues when assessing the true performance of high-performance parsers?*

Paper structure   After discussing background and related work (Section 2), we describe our data setup and give a survey of AMR metrics (Section 3). We then evaluate the metrics with regard to scoring i) corpora (Section 5), ii) AMR pairs (Section 6) and iii) cross-metric differences in their ranking behavior (Section 7). We conclude by discussing limitations of our study (Section 8), give recommendations and outline future work (Section 9).[3]

2   Background and related work

AMR parsing and applications   Over the years, we have observed a great diversity in approaches to AMR parsing, ranging from graph prediction with a pipeline (Flanigan et al., 2014), or a neural network (Lyu and Titov, 2018; Cai and Lam, 2020) to transition-based parsing (Wang et al., 2015) and sequence-to-sequence parsing, e.g., by exploiting large parallel corpora (Xu et al., 2020). A recent trend is to exploit the knowledge in large pre-trained sequence-to-sequence language models such as T5 (Raffel et al., 2019) or BART (Lewis et al., 2020), by fine-tuning them on AMR corpora,

as show-cased, e.g., by Bevilacqua et al. (2021). Such models are on par or tend to surpass estimates for human AMR agreement (Banarescu et al., 2013), when measured in SMATCH points.

AMR, by virtue of its properties as a graph-based abstract meaning representation, is attractive for many key NLP tasks, such as machine translation (Song et al., 2019), summarization (Dohare et al., 2017; Liao et al., 2018), NLG evaluation (Opitz and Frank, 2021; Manning and Schneider, 2021; Ribeiro et al., 2022) and measuring semantic sentence similarity (Opitz and Frank, 2022).

Metric evaluation for MT evaluation   Metric evaluation for machine translation (MT) has received much attention over the recent years (Ma et al., 2019; Mathur et al., 2020; Freitag et al., 2021). When evaluating metrics for MT evaluation, it seems generally agreed upon that the main goal of a MT metric is high correlation to human ratings, mainly with respect to rating adequacy of a candidate against one (or a set of) gold reference(s).

A recent shared task (Freitag et al., 2021) meta-evaluates popular metrics such as BLEU (Papineni et al., 2002) or BLEURT (Sellam et al., 2020), by comparing the metrics' scores to human scores for systems and individual segments. They find that the performance of each metric varies depending on the underlying domain (e.g., TED talks or news), and that most metrics struggle to penalize translations with errors in reversing negation or sentiment polarity, and show lower correlations for semantic phenomena including subordination, named entities and terminology. This indicates that there is potential for cross-pollination: clearly, AMR metric evaluation may profit from the vast amount of experience of metric evaluation for other tasks. On the other hand, MT evaluation may profit from relating semantic representations, to better differentiate semantic errors with respect to their type and severity. A first step in this direction may have been made by Zeidler et al. (2022), who assess the behaviour of MT metrics, AMR metrics, and hybrid metrics when analyzing sentence pairs that differ in only one linguistic phenomenon.

3   Study Setup: Data creation and AMR metric overview

In this Section, first we select data and two popular high-performance parsers for creating candidate AMRs. Then we describe the human quality annotation, and give an overview of automatic AMR

---

[2] Under the unrealistic assumptions of an omniscient annotator and AMR being the ideal way of meaning representation, one might require that $\mathcal{H}$ always has exactly one element.

[3] Code and data for our study are available at https://github.com/Heidelberg-nlp/AMRParseEval.

```
----------------Reference AMR and Sentence-----------------
(l / look-over-06                  ``Looking over to the flag''
    :ARG1 (f / flag))
--------------------Candidate parses----------------------
(l / look-01                    (z0 / look-01
    :direction (o / over)           :ARG2 (z1 / flag)
    :destination (f / flag))        :direction (z2 / over))
-------------------------Eval-----------------------------
 Smatch (ref, cand): both score 0.2 (indicates low quality)
 Human (sent, cand): both are acceptable
 Human (cand, cand): no preference
-----------------------------------------------------------
```

Figure 2: Data example: acceptable, low SMATCH. That is, $P \in \mathcal{H}$ but $P \notin \mathbb{B}(\text{SMATCH}, ref)$.

metrics that we consider in our subsequent studies.

Parsers and corpora We choose the AMR3 benchmark[4] and the literary texts from the freely available Little Prince corpus.[5] As parsers we choose T5- and BART-based systems, both on par with human IAA estimates, where BART achieves higher scores on AMR3.[6] We proceed as follows: we 1. parse the corpora with T5 and BART parsers and use SMATCH to select diverging parse candidate pairs, and 2. sample 200 of those pairs, both for AMR3, and for Little Prince (i.e., 800 AMR candidates in total).

3.1 Annotation dimensions

Annotation dimension I: pairwise ranking The annotator is presented the sentence and two candidate graphs, assigning one of three labels and a free-text rationale. The labels are either +1 (prefer first graph), −1 (prefer second graph), or 0 (both are of same or very similar quality).

Annotation dimension II: parse acceptability In addition, each graph is independently assigned a single label, considering only the sentence that it is supposed to represent. Here, the annotator makes a binary decision: +1, if the parse is acceptable, or 0, if the graph is not acceptable. A graph that is acceptable is fully valid, or may allow a very minor meaning deviation from the sentence, or a slightly weird but allowed interpretation that may differ from a normative interpretation. All other graphs are deemed not acceptable (0).

Example: Acceptable candidates, low SMATCH Figure 2 shows an example of two graphs that have very low structural overlap with the reference (SMATCH = 0.2), but are acceptable. Here, the candidate graphs both differ from the reference

---

[4]LDC corpus `LDC2020T02`

[5]From `https://amr.isi.edu/download.html`

[6]See `https://github.com/bjascob/amrlib-models` for more benchmarking statistics.

```
---------------Reference AMR (excerpt)--------------------
(i2 / imagine-01
    :ARG0 (y / you)
    :ARG1 (a / amaze-01
        :ARG1 (i / i)))
        :time-of (w / wake-01
            :ARG0 (v / voice
                :mod (o / odd)
                :mod (l / little))
            :ARG1 i))))))
---------------Candidate parse (excerpt)------------------
(ii / imagine-01
    :ARG0 (y / you)
    :ARG1 (a / amaze-01
        :ARG0 (v / voice
            :mod (l / little)
            :mod (o / odd))
        :ARG1 (ii2 / i)))
--------------------------------------------------------
Means: (..) imagine my amazement (..) by an odd little voice
Should mean: (..) imagine my amazement (..) when I was
         awakened by an odd little voice
-------------------------Eval----------------------------
 Smatch (ref, cand): scores 0.88 (indicates high quality)
 Human (sent, cand): not acceptable
--------------------------------------------------------
```

Figure 3: Data example excerpt that shows an unacceptable parse with high SMATCH. That is, $P \notin \mathcal{A} \supseteq \mathcal{H}$ but $P \in \mathbb{B}(\text{SMATCH}, ref)$

because they tend to a more conservative interpretation, using the more general *look-01* predicate instead of the *look-over-06* predicate in the human reference. In fact, the meaning of the reference can be considered, albeit valid, slightly weird, since *look-over-06* is defined in PropBank as *examining something idly*, which is a more 'specific' interpretation of the sentence in question. On the other hand, the candidate graphs differ from each other in the semantic role assigned to *flag*. In the first, *flag* is the destination of the *looking* action (which can be accepted), while in the second, we find a more questionable but still acceptable interpretation that *flag* is an *attribute of the thing that is looked at*.

Example: Candidate not acceptable, high SMATCH An inverse example (high SMATCH, unacceptable) is shown in Figure 3, where the parse omits *awaken*. Albeit the factuality of the sentence is not (much) changed, and the structural deviation may legitimately imply that the odd voice is the cause of amazement, it misses a relevant piece of meaning and is therefore rated unacceptable.

Label statistics will be discussed in Section 5, where the human annotations are also contrasted against parser rankings of automatic metrics.

3.2 Metric overview

We distinguish metrics targeting *monolingual AMR parsing evaluation* from *multi-purpose AMR metrics*. AMR metrics that are designed for evaluation of monolingual parsers typically have two features in common. First, they compare a candidate against

34

a reference parse that both (try to) represent the *same sentence*. Second, they measure the amount of successfully reconstructed reference structure.[7]

We also consider multi-purpose AMR metrics that aim to extend to use cases where AMRs represent *different sentences*, such as evaluation of cross-lingual AMR parsing, natural language generation (NLG) or rating semantic sentence similarity.

3.2.1 Monolingual AMR parsing metrics

Triple matching strategies SMATCH (Cai and Knight, 2013) and SEMA (Anchiêta et al., 2019) consider graph triples as the elementary constituents of an AMR graph. Both compute a triple overlap score between candidate and reference parses. SMATCH computes an alignment between the variable nodes of two AMRs, which is accurate but slow. The SEMA metric achieves a large speed-up by removing AMR variables from the graphs, replacing them with concept labels.

Inspired by BLEU: SEMBLEU BLEU (Papineni et al., 2002) is a popular (but debated) metric for machine translation evaluation. It matches bags-of-k-grams from candidate and reference, with a geometric mean of the precision scores over the k different bags. Inspired by BLEU, and, similar to SEMA, driven by the goal to make AMR evaluation more fast and efficient, Song and Gildea (2019) propose the SEMBLEU metric for AMR graphs. It extracts bags-of-k-grams from graphs, collected via breadth-first traversal. A point of motivation, similarly to SEMA, is that the metric skips the costly graph alignment. Per default, SEMBLEU uses k=3. In this work we additionally use k=2, following Opitz et al. (2021) who find that k=2 better relates to human notions of sentence similarity.

3.2.2 Multi-purpose metrics

S²MATCH and WLK/WWLK Targeting AMR metric application cases beyond monolingual parsing evaluation, such as measuring AMR similarity of different sentences, or cross-lingual AMR parsing evaluation, Opitz et al. (2020, 2021); Uhrig et al. (2021) propose three metrics: i) **S²MATCH** is an adaption of SMATCH that computes graded concept similarity (reflecting that, e.g., *cat* is more similar to *kitten* than to *plant*). ii) **WLK** applies the Weisfeiler-Leman kernel (Shervashidze et al., 2011) to compute a similarity score

over feature vectors that describe graph statistics in different iterations of node contextualization. iii) **WWLK** (Wasserstein WLK, Togninalli et al. (2019)) projects the nodes of the graphs to a latent space partitioned into different degrees of node contextualization. Wasserstein distance is then used to match the graphs, based on a pair-wise node distance matrix.

Setup of multi-purpose metrics For S²MATCH, WLK and WWLK we use the default setup, which consists of GloVe (Pennington et al., 2014) embeddings and k=2 in WLK and WWLK, where k indicates the depth of node contextualizations.

Default WWLK initializes parameters randomly, if tokens are out of vocabulary (a random embedding for each OOV token type). To achieve deterministic results, without fixing a random seed, we could initialize the OOV parameters to 0. However, with this we'd lose valuable discriminative information on graph similarity. We therefore adopt a slight adaptation for WWLK and calculate the *expected* distance matrix before Wasserstein metric calculation, making results more reproducible while keeping discriminative power.

We also introduce WWLK-**k3e2n**, a WWLK variant with *edge2node (e2n)* transforms, more tailored to monolingual AMR parsing evaluation, which is the focus of this paper. It increases the score impact of edge labels, motivated by the insight that edge labels are of particular importance in AMR parsing evaluation. It transforms an edge-labeled graph into an *equivalent* graph without edge-labels.[8] This is also known as 'Levi transform' (Levi, 1942), and has been previously advocated for AMR representation by Beck et al. (2018) and Ribeiro et al. (2019). Since due to the transform the distances in the graph will grow, we increase k by one (**k=3**). With this, we can set all edge weights to 1.

3.2.3 Simple baseline

To put the results into perspective, we introduce a very SIMPLE baseline: SIMPLE extracts bag-of-words (relation and concept labels) from two AMR graphs and computes the size of their intersection vs. the size of their union (aka *Jaccard Coefficient*).

4 Preliminaries

We denote an AMR metric m over AMRs as:

$$m : \mathcal{A} \times \mathcal{A} \to \mathbb{R}, \qquad (1)$$

---

[7]The notion of *success* is mostly focused on structural matches, and can vary among metrics, usually depending on theoretical arguments of the developers of the metric.

[8]E.g., $(x, arg0, z) \to (x, y) \land (y, z) \land (y, arg0)$.

and a human metric h as

$$h : \mathcal{A} \times \mathcal{S} \to \mathbb{R}, \qquad (2)$$

where $\mathcal{S}$ contains sentences.

5 Study I: System-level scoring

Research questions We focus on two questions:

1. How are the two parsers rated by humans?

2. How do metrics score our two parsers?

With 1. we aim to assess whether there is still room for AMR parser improvement, even though their SMATCH scores pass estimated human IAA. And for 2. we aim to know whether the metric rankings (still) appropriately reflect parser quality.

5.1 System scoring

Aggregation strategies: Micro vs. Macro We have defined a metric between two AMRs. For ranking systems, we need to aggregate the individual pair-wise assessments into a single score. At this point, it is important to note that most papers use (only) micro SMATCH for ranking parsers, i.e., counting triple matches of aligned AMR pairs over all AMR pairs (before a final F1 score calculation).

Naturally, such micro corpus statistics are *unbiased* w.r.t. to whatever is defined as a single evaluation instance (in SMATCH: triples), but the trade-off is that they are biased towards instance type frequency and sentence length, since longer sentences tend to yield substantially more triples. Hence, the influence of a longer sentence may marginalize the influence of a shorter sentence. This issue may be further aggravated by the fact that longer sentences tend to contain more named entity phrases, and entity phrases typically trigger large simple structures, that are mostly easy to project.[9] Therefore, micro corpus statistics alone *could* potentially yield an incomplete assessment of parser performance. To shed more light on this issue, we provide additional evaluation via macro aggregation.

---

[9]As a small example, consider *The bird sings* vs. *Jon Bon Jovi sings*. The first sentence yields 3 triples, while the second sentence yields 8 triples, where the *John Bon Jovi* named entitiy structure has added 6 triples, outweighing the key semantic event *x sings*. Micro score would assign 2.6 times more importance to the second sentence/AMR.

Statistics for micro and macro system scoring We calculate two statistics. The first statistic shows the (micro/macro)-aggregated corpus score for a metric m, parsed corpus X and gold corpus G:

$$\mathbb{S}(m, X, G) = AGGR(\{m(X_1, G_1), ..., m(X_n, G_n)\}),$$

For macro metrics, $AGGR$ is the mean of pairwise scores over all instances in a corpus X. In case of the human metric, this is the ratio of acceptable parses in X. For micro metrics, $AGGR$ computes overall matching triple F1 (SMATCH, SEMA) or overall k-gram BLEU (SEMBLEU). For WLK and WWLK, a micro variant is not implemented, hence we only show their macro scores.

The second statistic shows how often m prefers the parses in a parse corpus X over the these in Y:

$$\mathbb{P}(m, X, Y, G) = \sum_{i=1}^{n} \mathbb{I}[m(X_i, G_i) > m(Y_i, G_i)].$$

Here, $\mathbb{I}[c]$ denotes a function that returns 1 if the condition c is true, and zero in all other cases. For better comparability of numbers, we distribute cases where $m(X_i, G_i) = m(Y_i, G_i)$, which are frequent for the human metric, evenly over $\mathbb{P}(m, X, Y, G)$ and $\mathbb{P}(m, Y, X, G)$.

5.2 Results

Results are shown in Table 1. In view of our research questions, we make interesting observations.

AMR parsing is far from solved Considering the ratio of parses that were rated acceptable by the human (HUM, $\mathbb{S}$), they are surprisingly low, at only 0.58 (BART, Little Prince, Table 1); 0.69 (T5, Little Prince). Other parses have errors that substantially distort sentence meaning, even though major parts of the AMRs may structurally overlap.

Better SMATCH on AMR benchmark may not (always) imply a better parser On AMR3, when inspecting corpus-SMATCH (micro SMATCH, Table 1), BART is considered the better parser, in comparison to T5 (+2 points). However, when consulting macro statistics, a different picture emerges. Here, BART and T5 obtain the same scores: AMR3, 0.62 vs. 0.62, Table 1. On the literary texts (Little Prince), where the domain is different and sentences tend to be shorter, T5 significantly (binomial test, $p < 0.05$) outperforms BART, both in the ratio

36

| | | Little Prince | | | | | | AMR3 | | | | | |
| | | $\mathbb{P}$ | | | $\mathbb{S}$ | | | $\mathbb{P}$ | | | $\mathbb{S}$ | | |
		BART	T5	Δ	BART	T5	Δ	BART	T5	Δ	BART	T5	Δ
	HUM	87	113	-26	0.58	0.69	-0.11	100	100	0.0	0.62	0.62	0.00
	SIMPLE	87	113	-26	0.69	0.7	-0.01	82	118	-36	0.75	0.75	0.00
Macro	SEMA	84	116	-32	0.6	0.63	-0.03	89	111	-22	0.68	0.68	0.00
	SEMBLEU-k2	90	110	-20	0.61	0.63	-0.02	98	102	-4	0.70	0.69	0.01
	SEMBLEU-k3	90	110	-20	0.51	0.53	-0.02	103	97	6	0.58	0.58	0.00
	SMATCH	94	106	-12	0.73	0.74	-0.01	95	105	-10	0.77	0.77	0.00
	S^{2}MATCH	93	107	-14	0.75	0.76	-0.01	95	105	-10	0.79	0.79	0.00
	WLK-k2	92	108	-16	0.63	0.65	-0.02	96	104	-8	0.69	0.69	0.00
	WWLK-k2	91	109	-18	0.79	0.8	-0.01	102	98	4	0.84	0.84	0.00
	WWLK-k3e2n	97	103	-6	0.72	0.73	-0.01	94	106	-12	0.78	0.78	0.00
Micro	SEMA	-	-	-	0.62	0.64	-0.02	-	-	-	0.69	0.68	0.01
	SEMBLEU	-	-	-	0.53	0.54	-0.01	-	-	-	0.60	0.57	0.03
	SMATCH	-	-	-	0.74	0.74	-0.01	-	-	-	0.77	0.75	0.02
	S^{2}MATCH	-	-	-	0.76	0.76	0.00	-	-	-	0.80	0.77	0.03

Table 1: Corpus level scoring results. Negative Δ shows preference for T5, positive Δ shows preference for BART.

of acceptable sentences (BART: 0.58, T5: 0.69), and in number of preferred candidates (BART: 87, T5: 113). Note that this insight is independent from our human annotations.

All in all, this may suggest that BART tends to provide better performance for longer sentences, while T5 tends to provide better performance especially for shorter and medium-length sentences. Further analysis provides more evidence for this, cf. Appendix A.1: Figure 6 and Figure 7).

Metrics for system ranking Regarding our tested metrics, especially the macro metrics, a clear pattern is that they mostly agree with the human ranking. However, our current results for the different metrics do not tell much, yet, about their suitability for AMR assessment and ranking. Even if a metric ranks a parser more similarly to the human, this may be for the wrong reasons, since this statistic filters out pair-wise correspondences to the human. This is also indicated by results of the simplistic bag-of-structure metric SIMPLE, which achieves the same results as human (HUM) on Little Prince, with respect to the number of preferred parses ($\mathbb{P}$, Little Prince, Table 1, HUM vs. SIMPLE). In that respect, it is more important to assess the pair-wise metric accuracy and metric specificity, which we will visit next in Sections 6 and 7.

6 Study II: Metric accuracy on parse level

Research questions Now, we are interested in the metric accuracy, that is, agreement of AMR metrics with the human ratings. In particular, we would like to know, regarding:

- Pair-wise parse accuracy: How do metrics agree with human preferences when ranking two candidates?

- Individual parse accuracy: Can metrics tell apart acceptable from unacceptable parses?

Note that these are hard tasks for metrics, since both T5 and BART show performance levels on par or above estimated measurements for human IAA. Therefore, smaller structural divergences from the reference can potentially have a bigger impact on parse acceptability (or preference) than larger structural deviations, that could express different (but valid) interpretations or (near-)paraphrases.

6.1 Evaluation metrics

Pairwise accuracy Recall that the human assigned one of three ratings: 1, if AMR x is better, -1, if AMR y is better, and 0 if there is no considerable quality difference between two candidate graphs x and y. A metric assigns two real values, $m(x,g)$ and $m(y,g)$, where g is the reference graph. Mapping the score to -1 or 1 is simple and intuitive, prompting us to introduce pair-wise accuracy. Consider a data set SD that contains all graph triplets (x,y,g) with a human preference sign (label -1 or $+1$). Further, let $\delta^m(x,y,g) = m(x,g) - m(y,g)$ the (signed) quality difference between x and y when using m. Anal-

| | Little Prince | | AMR3 | |
	PA	$\mathbb{A}\Delta$	PA	$\mathbb{A}\Delta$
HUM	1.0	233	1.0	234
RAND	0.5	0.0	0.5	0.0
SIMPLE	$0.66^\dagger$	11.0	$0.68^\dagger$	$39.5^\dagger$
SEMA	$0.66^\dagger$	24.3	$0.7^\dagger$	$35.3^\dagger$
SEMBLEU-k2	$0.67^\dagger$	25.0	$0.74^\dagger$	28.0
SEMBLEU-k3	$0.63^\dagger$	32.0	$0.68^\dagger$	29.0
SMATCH	$\mathbf{0.72}^\dagger$	42.0†	$0.7^\dagger$	$35.0^\dagger$
S^{2}MATCH	$\mathbf{0.72}^\dagger$	35.3	$0.7^\dagger$	$42.3^\dagger$
WLK	$0.66^\dagger$	28.0	$0.68^\dagger$	$41.5^\dagger$
WWLK-k2	$0.63^\dagger$	20.5	$0.73^\dagger$	$51.0^\dagger$
WWLK-k3e2n	$0.66^\dagger$	$\mathbf{48.0}^\dagger$	$\mathbf{0.76}^\dagger$	$\mathbf{57.0}^\dagger$

Table 2: Metric agreement with human. †: random baseline (RAND) not contained in 95% confidence interval.

ogously $\delta^h(x,y)$ is the human preference. Then, the pairwise accuracy is

$$PA = \frac{1}{|D|} \sum_{(x,y,g) \in D} \mathbb{I}[\delta^m(x,y,g) \cdot \delta^h(x,y) > 0] \quad (3)$$

This measures the ratio of candidate pairs where the metric has made the same signed decision as the human, in preferring one over the other parse.

Acceptability score When rating acceptability, the human rates a single parse (given its sentence), assigning 1 (acceptable) or 0 (no acceptable). The metrics make use of the reference graph to compute a score. Aiming at an evaluation metric that makes as few assumptions as possible, we formulate the following expectation for an AMR graph metric to fulfill: the average rank of the scores for parses that have been labeled acceptable by the human should surpass the average rank of the scores for parses labeled as being not acceptable. Let $\mathcal{I}^+$ ($\mathcal{I}^-$) be the set of indices for which the human has assigned a label that indicates (un-)acceptability. Let $S = \{m(X_1, G_1)...m(X_n, G_n)\}$ be the metric m's scores over all (x,g) parse/reference pairs, and R be the ranks of D. Let R^+ (and R^-) be the set of ranks indexed by $\mathcal{I}^+$ (and $\mathcal{I}^-$). Then

$$\mathbb{A}\Delta = avg(R^+) - avg(R^-) \quad (4)$$

To increase robustness, we use $avg := \text{median}$.

6.2 Results

The results are shown in Table 2. We conclude:

All metrics are suitable for pairwise-ranking of parses from high-performance parsers All met-

rics significantly outperform the random baseline with regard to the pair-wise ranking accuracy (PA). For Little Prince, SMATCH and S^{2}MATCH yield the best performance, while for AMR3, WWLK-k3e2n has the best performance (closely followed by SEMBLEU-k2). Among different metrics, however, the differences are not large enough to confidently recommend one metric over the other.

Parse acceptability rating is hard When tasked to rate parse acceptability ($\mathbb{A}\Delta$), all metrics show issues. For Little Prince, only SMATCH and WWLK-k3e2n significantly outperform the chance baseline, while for AMR3 all metrics are significantly above chance level, except SEMBLEU. Overall, however, the differences are not large enough to confidently recommend one metric over the other. On both corpora, best results are achieved with WWLK-k3e2n (Little Prince: 48.0, AMR3: 57.0).

Control experiment of metrics We additionally parse a subset of 50 sentences with an older parser (Flanigan et al., 2014) that scores more than 20 points lower SMATCH, when compared with IAA as estimated in Banarescu et al. (2013). All metrics (with the exception of SIMPLE for one pair) correctly figure out all rankings and acceptability (according to the human, BART and T5 are preferred in all cases, except two cases where all three systems deliver equally valid graphs). This indicates that metrics indeed can accurately tell apart quality differences, *if* they are large enough and do not lie beyond human IAA.

7 Metric specificity

We found little evidence that could help us giving recommendations on which metrics to prefer over others for monolingual parser evaluation in the high-performance regime. On the contrary, we found evidence that no metric can sufficiently assess parse acceptability. Therefore, it is interesting to see whether the metrics can provide *different* views on parse quality.

7.1 Correlation analysis

Statistics We compute Spearman's ρ over metric pairs. Spearman's ρ calculates Pearson's ρ on the ranked predictions, which increases robustness.

Results Results are plotted in Figures 4 and 5. For both datasets, we see that the Wasserstein metrics provide rankings that differ more from the rankings assigned by other metrics, suggesting that they
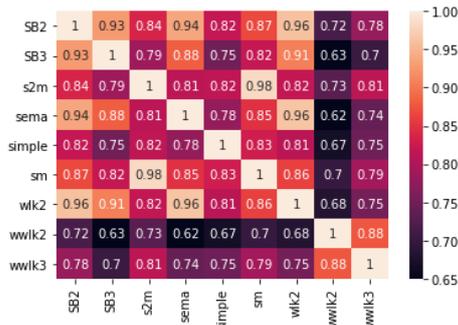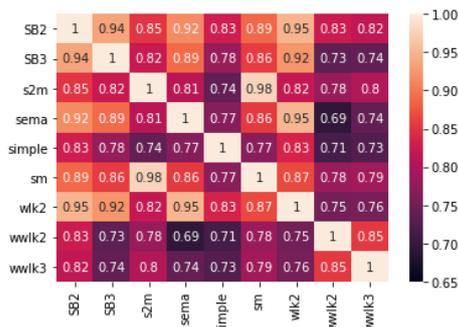
Figure 4: Inter-metric correlation on Little Prince.

Figure 5: Inter-metric correlation on AMR3.

have unique features. On the other hand, the SEM-BLEU metrics tend to agree the most with the rankings of the other metrics, suggesting that they share more features with other metrics. On a pair-wise level, the most similar metrics are SMATCH and S^{2}MATCH, which is intuitive, since S^{2}MATCH is an adaption of SMATCH that also targets the comparison of AMRs from different sentences. Indeed, synonyms and similar concepts are unlikely to often occur in monolingual parsing, where parses contain exactly matching concepts. Further, WLK very much agrees with SEMBLEU, which seems intuitive, since both aim at comparing larger AMR subgraphs. Lowest agreement is exhibited between SEMA and WWLK, perhaps because these metrics are of different complexity and share different goals: simple and fast match of structures vs. graded assessment for general AMR similarity.

8 Discussion of study limitations

There are limitations of our study:

Limitation I: *Single vs. Double annotation*
While our quality annotations stem from an experienced human annotator, we would have liked to obtain annotations from a second annotator to measure IAA for AMR quality rating. This was partly

precluded by the high costs of AMR annotation, which requires much time and experience. This is also reflected in the AMR benchmark corpora: the majority of graphs were created by a single annotator. Note, however, that some findings are independent of annotation (e.g., macro vs. micro metric corpus scoring, metric specificity).

Limitation II: *Assessing individual suitability of metrics for rating high-performance parsers*
Our study reports relevant findings on (monolingual) AMR parsing evaluation in high-performance regimes, and on upper bounds of AMR parsing. But an important question we had to leave open is the individual suitability of the metrics for comparing high-performance parsers.

Limitation III: *Single-reference parses and ambiguity* Elaborating on *Limitation II* and recalling that AMR benchmarks have only single references, another caveat is that potentially correct metric behavior may be misinterpreted in our study. E.g., if a sentence allows two different interpretations, a metric might (correctly) yield a low score for the reference (different meaning), while the (reference-less) human rating may find the parse acceptable. This issue may also be mitigated by providing (costly) double annotation of AMR benchmark sentences.

To facilitate follow-up research, we release the annotated data. Our Little Prince annotations can be freely released, while AMR3 annotations require proof of LDC license.

9 Discussion and Conclusions

Main recommendations based on our study:

Recommendation I Besides micro aggregate scores we recommend using **a macro aggregate score** for parse evaluation (e.g., macro SMATCH, computed as an average over sentence scores): Commonly, only micro corpus statistics are used to compare and rank parsers. Yet, we found that macro (sentence-average) metrics can provide a valuable **complementary assessment** that can highlight important *additional* strengths of high-performance parsers.

Recommendation II We recommend conducting **more human evaluation of AMR parses**. With the available high-performance AMR parsers, it becomes more important to conduct manual analyses of parse quality. Our

study provides evidence that AMR parsing still has large room for improvement, due to small but significant errors. Since this may not be noticeable for (current) metrics when given a single human reference, **future work on parsing may profit from careful human acceptability assessments**.

T5 vs. BART: which parser to prefer? Next to AMR parser developers, this question mainly concerns potential users of AMR parsers. Fine-tuned T5 and BART are both powerful AMR parsers. We observe a slight tendency that researchers prefer BART, possibly since it achieves slightly better SMATCH scores than T5 on the AMR3 benchmark. But our work shows that differences between the systems are often finer than what can be assessed with structural overlap metrics (SMATCH), and both systems are generally strong but struggle with small but significant meaning errors.

In our study we found that when choosing between T5 and BART based AMR systems, **the choice might depend on the target domain**. Indeed, our results on Little Prince and AMR3 (mainly news) could indicate that **T5 may have an edge over BART when parsing literary texts**, and shorter sentences in general, while **BART has an edge over T5 when parsing longer sentences, and sentences from news sources**, especially if they are longer. However, it must be clearly noted, that we do not know (yet) whether this insight carries over to other types of literary texts.

Perhaps, if we presume that performance is carried over to other types of literary texts, a possible explanation can be found in the data these two large models were trained on. BART uses the same training data as RoBERTa (Liu et al., 2019), e.g., Wikipedia, book corpora and news. T5 leverages the colossal common crawl corpus (C4), that contains all kinds of texts scraped from the web. This *could* make T5 more robust to AMR domain changes, but less suitable for analysing longer sentences, since these may occur more frequently in BART's corpora that seem more normative.

Which AMR metric to use? Our findings do **not provide conclusive evidence** on this question, partly due to insufficient data size, partly due to the general difficulty of the task. WWLK-k3e2n seems slightly more useful for detecting parse acceptability and pairwise ranking on news, while SMATCH yields best ranking on Little Prince.

However, our work shows that **it can be useful to calculate more than one metric to compare parsers**. In particular, we saw that predictions of structural matching metrics differ considerably from graded semantic similarity-based metrics, such as the WWLK metric variants. This suggests that these two types can provide complementary perspectives on parsing accuracy. Metric selection may, of course, also be driven by users' specific desiderata, such as speed (SEMA, SEMBLEU, WLK), 1-1 alignment (SMATCH), n:m alignment (WWLK), or graded matching (SMATCH, WWLK). Overall, we see much **profit to gain from more research into AMR metrics**, and will now outline a direction that we believe is very interesting.

A direction for future research: Reference-less AMR metrics Recall that for human quality assessments a candidate graph is compared to a *sentence*, in lieu of a reference AMR. If this process can be approximated by a metric, we gain an important mechanism for assessing the quality of high-performance parsers: a measure that is cheap and not biased towards a single reference.

To date, referenceless AMR parse quality rating has been attempted by Opitz and Frank (2019); Opitz (2020). However, an unsolved issue is that this approach does not approximate a human quality assessment, but instead tries to project SMATCH score without using a reference, and we saw that SMATCH cannot well assess the impact of fine errors of high-performing parsers.

A worthwhile solution could be found in the exploitation of indirect tools: E.g., our human annotation indicated that significant, but small structural errors are sometimes due to coreference, which is known to be a hard task in general (Levesque et al., 2012) and for AMR in particular (Anikina et al., 2020). Therefore, e.g., one may profit from matching parses from a high-performance parser against the structures predicted by a strong coreference system, possibly with the help of a predicted AMR-to-text alignment (Blodgett and Schneider, 2021). Another promising route to take may be to invert approaches of Opitz and Frank (2021); Manning and Schneider (2021) who evaluate AMR-to-text generation without reliance on a reference by using a strong parser for back-parsing. It may be beneficial to use strong AMR-to-text systems to generate from candidate AMRs, and to match the generations against the source sentence using strong automatic text-to-text metrics.

Acknowledgements

We are grateful to three anonymous reviewers for their valuable comments that have helped to improve this paper. We are also thankful to Julius Steen for valuable discussions.

References

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 Publishg.

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

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, Melbourne, Australia. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12564–12573.

Austin Blodgett and Nathan Schneider. 2021. Probabilistic, structure-aware algorithms for improved variety, accuracy, and coverage of AMR alignments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3310–3321, Online. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online. Association for Computational Linguistics.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Shibhansh Dohare, Harish Karnick, and Vivek Gupta. 2017. Text summarization using abstract meaning representation.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Friedrich Wilhelm Levi. 1942. *Finite geometrical systems: six public lectues delivered in February, 1940, at the University of Calcutta*. University of Calcutta.

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

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Emma Manning and Nathan Schneider. 2021. Referenceless parsing-based evaluation of AMR-to-English generation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 114–122, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Juri Opitz. 2020. AMR quality rating with a lightweight CNN. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 235–247, Suzhou, China. Association for Computational Linguistics.

Juri Opitz, Angel Daza, and Anette Frank. 2021. Weisfeiler-Leman in the Bamboo: Novel AMR Graph Metrics and a Benchmark for AMR Graph Similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.

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. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2021. Towards a decomposable metric for explainable evaluation of text generation from AMR. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1504–1518, Online. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2022. Sbert studies meaning representations: Decomposing sentence embeddings into explainable amr meaning features. *arXiv preprint arXiv:2206.07023*.

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

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

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Leonardo Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. 2019. Enhancing AMR-to-text generation with dual graph representations. 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 3183–3194, Hong Kong, China. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. 2011. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9).

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. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic Neural Machine Translation Using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. 2019. Wasserstein weisfeiler-lehman graph kernels. In *Advances in Neural Information Processing Systems*, volume 32, pages 6436–6446. Curran Associates, Inc.

Sarah Uhrig, Yoalli Garcia, Juri Opitz, and Anette Frank. 2021. Translate, then parse! a strong baseline for cross-lingual AMR parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 58–64, Online. Association for Computational Linguistics.

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

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2020. Improving AMR parsing with sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2501–2511, Online. Association for Computational Linguistics.

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

A Appendix

A.1 Sentence length vs. score

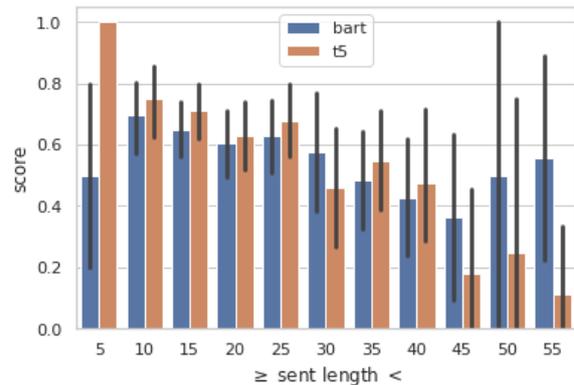See Figures 6, 7. For total sentence length distribution see Figure 8.

Figure 6: **Sentence length vs. human acceptability** on all annotated data. 55 includes all sentences longer than 55 tokens. See Figure 8 for occurences of different sentence lengths.
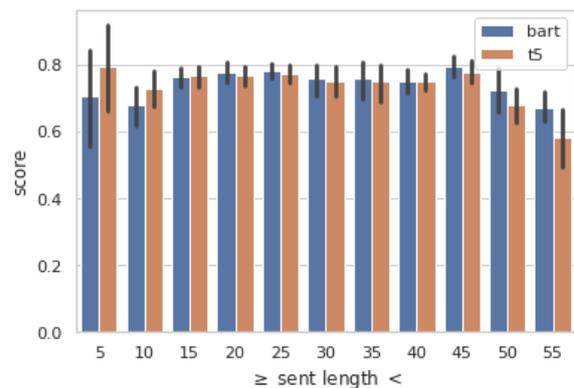

Figure 7: **Sentence length vs. Smatch** on all annotated data. 55 includes all sentences longer than 55 tokens. See Figure 8 for occurences of different sentence lengths. Other metrics look similar.
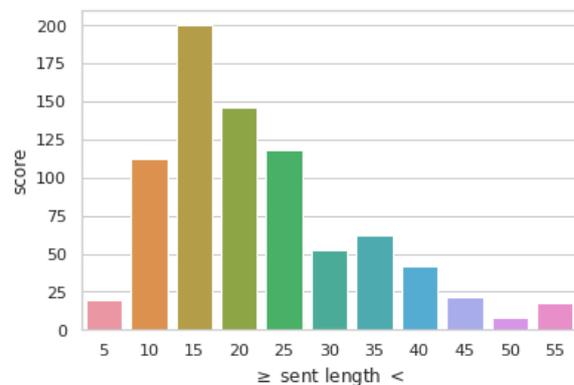

Figure 8: Sentence length occurrences. 55 includes all sentences longer than 55 tokens.