Towards a Decomposable Metric for Explainable Evaluation of Text Generation from AMR

Juri Opitz
Dept. of Computational Linguistics
Heidelberg University
69120 Heidelberg
opitz@cl.uni-heidelberg.de

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

Abstract

Systems that generate natural language text from abstract meaning representations such as AMR are typically evaluated using automatic surface matching metrics that compare the generated texts to reference texts from which the input meaning representations were constructed. We show that besides well-known issues from which such metrics suffer, an additional problem arises when applying these metrics for AMR-to-text evaluation, since an abstract meaning representation allows for numerous surface realizations. In this work we aim to alleviate these issues by proposing $\mathcal{MF}_\beta$, a decomposable metric that builds on two pillars. The first is the **principle of meaning preservation** $\mathcal{M}$: it measures to what extent a given AMR can be reconstructed from the generated sentence using SOTA AMR parsers and applying (fine-grained) AMR evaluation metrics to measure the distance between the original and the reconstructed AMR. The second pillar builds on a **principle of (grammatical) form** $\mathcal{F}$ that measures the linguistic quality of the generated text, which we implement using SOTA language models. In two extensive pilot studies we show that fulfillment of both principles offers benefits for AMR-to-text evaluation, including explainability of scores. Since $\mathcal{MF}_\beta$ does not necessarily rely on gold AMRs, it may extend to other text generation tasks.

1 Introduction

Abstract Meaning Representation (AMR, Banarescu et al. (2013)) aims at capturing the meaning of a sentence in a machine-readable graph format. AMR captures, i.a., word senses, semantic roles and coreference. The AMR in Fig. 1 represents the sentence *Perhaps, the parrot is telling itself a story*. In this graph, tell-01 links to a PropBank (Palmer et al., 2005) frame, and arg$_n$ labels indicate partici-

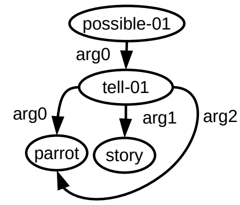

Figure 1: *"Perhaps, the parrot is telling itself a story"*.

pant roles: *parrot* is both *speaker* (arg0) and *hearer* (arg2), *story* is the *utterance* (arg1).

The task of AMR-to-text generation has recently garnered much attention (Song et al., 2017, 2018; Konstas et al., 2017; Cai and Lam, 2020b; Ribeiro et al., 2019). The output of AMR-to-text systems is typically evaluated against the sentence from which the AMR was created, using standard surface string matching metrics such as BLEU (Papineni et al., 2002) or CHRF(++) (Stanojević et al., 2015; Popović, 2015, 2016; Popov, 2017), as is standard in many NLG tasks. These metrics suffer from several issues, for example, they penalize paraphrases, are highly sensitive to outliers (Mathur et al., 2020), and lack interpretability (Sai et al., 2020).

Some of these issues get compounded when evaluating AMR-to-text. The core of the problem is that there are many ways to realize a sentence from a meaning representation. Fig. 2 shows four candidate sentences (**i-iv**) for a given AMR (left). One system generates **(i)**: *Maybe the cat is playing.* while another generates **(iii)**: *Perhaps, the cat plays the flute*. Clearly, **(i)** captures the meaning of the gold graph better than **(iii)**, which contains 'hallucinated' content – a well-known issue in neural generation (Logan et al., 2019; Wang and Sennrich, 2020). Yet, when using a canonical metric such as BLEU to evaluate sentences **(i)** and **(iii)** against the reference, the system that produces hallucinations **(iii)** is greatly rewarded (54 BLEU points) to the disadvantage of systems that yield meaning preserving

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1504–1518
April 19 - 23, 2021.

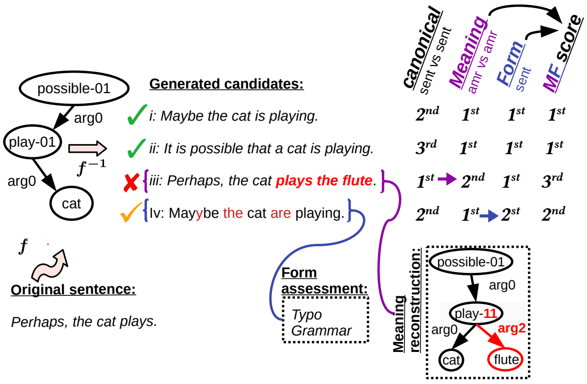

Figure 2: The ***Canonical*** evaluation matches n-grams from the sentences and assigns inappropriate ranks. Our metric $\mathcal{MF}_\beta$ fuses ***Meaning*** and ***Form*** assessment and better reflects the ranking of the generations.

sentences **(i)** (18 points) and **(ii)** (5 points).

This work aims at a (better) metric that **measures meaning preservation** of the generated output towards the MR given as input, by (re-)constructing an AMR from the generated sentence and comparing it to the input AMR. In Fig. 2, Reconstruction is the result of parsing **(iii)**. The reconstructed AMR exposes several meaning deviations (marked in red): it contains an alternate sense of *play* and contains an additional semantic role arg2 with filler *flute*. By contrast, when converting sentences **(i)**, **(ii)**, or **(iv)** to AMRs, we obtain flawless reconstructions. We will measure preservation of ***Meaning*** using well-defined graph matching metrics.

Figure 2 also illustrates that assessing meaning preservation is not sufficient to rate the quality of generations: (**iv**) captures the meaning of the AMR well – but its form is flawed: it suffers from wrong verb inflection, a common issue in low-resource text generation settings (Koponen et al., 2019).

In order to rate both meaning and form of a generated sentence, we combine the score for meaning reconstruction with a score called ***Form*** that **judges the sentence's grammaticality and fluency**. By these moves, we obtain a more suitable and explainable ranking with a combined ***MF*** score.[1] By clearly distinguishing between ***Meaning*** and ***Form***, our ***MF*** score (henceforth denoted by $\mathcal{MF}_\beta$) also aligns well with recent calls to achieve a clearer separation of these aspects in NLU (Bender and Koller, 2020).

Generally, our contributions are as follows:

(1) We propose two linguistically motivated principles that aim at a sound evaluation of AMR-to-text systems: the **principle of meaning preservation** and the **principle of (grammatical) form**.

(2) From these principles we derive and implement a (novel) $\mathcal{MF}_\beta$ **score for AMR-to-text generation**[2] which is composed of individual metrics for meaning and form aspects. $\mathcal{MF}_\beta$ allows users to modulate these two views on generation quality to vary their impact on the final metric score.

(3) We conduct two major pilot studies involving (English) text generations from a range of competitive AMR-to-text systems and human annotations. First we study the potential practical benefits of $\mathcal{MF}_\beta$ when evaluating systems, such as its prospects to offer interpretability of scores and finer-grained system analyses. The second study probes potential weak spots of $\mathcal{MF}_\beta$, e.g., its dependence on a strong AMR parser.

We consider $\mathcal{MF}_\beta$ as it stands as a suitable metric to enhance interpretability of generation scores.

2 Fusing meaning and form into $\mathcal{MF}_\beta$

While current NLG metrics lack *interpretability* and mainly focus on the form of generated text (Sai et al., 2020), in this work we emphasize the *meaning aspect* in NLG evaluation, which is most clearly dissociated from form when generating text from structured inputs such as AMR. At the same time, form and wording of the generated text cannot be ignored, as we want such systems to produce *natural and well-formed sentences*. Equipped with this two-fold objective, we start building our $\mathcal{MF}_\beta$ score which aims at a *balanced combination* of both quality aspects: **meaning and form**.

2.1 From principles to $\mathcal{MF}_\beta$

In a first step we introduce our

Principle of meaning $\mathcal{M}$. *Generated sentences should allow loss-less AMR reconstruction.*

This principle expresses a key expectation for a system that generates NL sentences from abstract meaning representations. Namely, the generated sentence should reflect the meaning of the AMR. So, in order to assess whether a generated sentence $s' = f^{-1}(m)$ is a valid generation for the input AMR m, rather than matching s' against a

[1]See Fig. 2: $1^{st}/2^{nd}$ rank: **i**; 3^{rd} rank: **iv**; 4^{th} rank: **iii**.

[2]We make code available at https://github.com/Heidelberg-NLP/MFscore.

reference sentence s, we perform this assessment **in the abstract MR domain**, by applying an inverse system f that *parses* the generated text back to an AMR $m' = f(s') = f(f^{-1}(m))$. I.e., we desire a $metric : \mathcal{D} \times \mathcal{D} \rightarrow [0, 1]$ that satisfies: $s \equiv s' \iff m = m' \iff metric(s, s') = 1$. Two texts are equivalent iff their meaning abstractions denote the same meaning. In case $f(s')$ yields an AMR $m' \neq m$, we can still determine the degree to which s' preserves the meaning of AMR m by measuring the distance between m and m' by standard AMR metrics, e.g., $Smatch(m, m')$.

Note that computing $Smatch(m, m')$ does not depend on a reference sentence, because the comparison is conducted purely in the abstract domain. This is mathematically more appealing for the evaluation of AMR-to-text, since it solves the problem that one abstract representation may result in various (valid) surface realizations (cf. Appendix A.1). Finally, we also do not necessarily need to rely on a gold graph m, but can instead set $m = f(s)$, i.e., the parse of the reference sentence. This means that future application of $\mathcal{M}$ to other kinds of text generation tasks is straightforward.

However, the principle $\mathcal{M}$ alone is not sufficient: we also expect the system to generate grammatically well-formed and fluent text. For example, s': *Possibly, it(self) tells parrot a story.* contains relevant content expressed in the AMR of Fig. 1, but it is neither grammatically well-formed, nor a natural and fluent sentence. This leads us to our

Principle of form $\mathcal{F}$. *Generated sentences should be syntactically well-formed, natural and fluent.*

In the style of the well-established F_β score (van Rijsbergen, 1979), we fuse these two principles into the $\mathcal{MF}_\beta$ score:

$$\mathcal{MF}_\beta = (1+\beta^2)\frac{Meaning \times Form}{(\beta^2 \times Meaning) + Form} \quad (1)$$

Here, $Form$ and $Meaning$ are expressed as ratios that will be more closely described in the following subsection. β allows users to gauge the evaluation towards $Form$ or $Meaning$, depending on specific application scenarios. Users may prefer the harmonic mean (β = 1) or may give $Meaning$ double weight compared to $Form$ (e.g., β = .5).[3] In our experiments we consider extreme decompositions into $Meaning$-$only$ ($\beta \rightarrow 0$) or $Form$-$only$ ($\beta \rightarrow \infty$).

[3]Generally, $Form$ receives β times as much importance compared with $Meaning$.

2.2 Parameterizing meaning

We measure $\mathcal{M}$ or $Meaning$ (Meaning Preservation) with a score range in $[0, 1]$ by reconstructing the AMR with a SOTA parser and computing the relative graph overlap of the reconstruction and the source AMR using graph matching. We call this ReSmatch. Given a generated sentence s' and source AMR m, we match $parse(s')$ against m by computing $amrMetric(parse(s'), m)$. This means that we have to decide upon $parse$ and $amrMetric$. We propose two potential settings.

AMR reconstruction To reconstruct the AMR with $parse$, we use the latest state-of-the-art AMR parser by Cai and Lam (2020a). With 80.3 Smatch F1, this parser is almost on-par with human agreement (estimated at 0.71–0.83 Smatch F1 in Banarescu et al. (2013)). We henceforth call it GSII.

Assessing $\mathcal{M}$ with AMR metrics To obtain a score for $\mathcal{M}$ we propose to use S^2match (Opitz et al., 2020) – a variant of Smatch (Cai and Knight, 2013) that performs a *graded match for concept nodes*. This offers the potential to compensate for noise in automatically generated text or minor lexical deviations from the original sentence.

Discussion Comparing to references by matching their meaning graphs has the prospect of offering interpretability and explanations, by detecting redundant or missing meaning components in the generations. In our studies, we will see that this assessment can be conducted by computing a *single graph overlap score* (e.g., S^2match F1), or along *multiple dimensions of meaning*, such as SRL, coreference or WSD (Damonte et al., 2017). Generally, $\mathcal{MF}_\beta$ gives researchers the flexibility of choosing a $parser$ or $amrMetric$ to their liking. In this work, we choose the best current $parser$ that achieves high IAA with humans. Yet, we would also like to know whether the $parser$ is vulnerable to specific peculiarities of generated sentences, or how using another parser affects the scores. We will investigate these issues more closely in §4.1.

2.3 Parameterizing form with LMs

Assessing sentence grammaticality and fluency is not an easy task (Heilman et al., 2014; Katinskaia and Ivanova, 2019). Recently, Lau et al. (2020); Zhu and Bhat (2020) show that probability estimates based on language models can be used as an indicator for measuring complex notions of form and for measuring acceptability in context. For our $\mathcal{MF}_\beta$ score we desire an interpretable ratio as input, which we base on LM predictions as follows.

Binary form assessment Given a specific candidate generation s', we use a binary variable to assess whether s' is of satisfactory form. For this, we first calculate the mean token probability:[4]

$$mtp(\cdot) = \frac{1}{n}\sum_{j=1}^{n} P(tok_j|ctx_j), \qquad (2)$$

where ctx_j is different for uni-directional LMs ($ctx_j = tok_{1...j-1}$) and bi-directional LMs ($ctx_j = tok_{1...j-1,j+1...n}$). We compute mtp for the generated sentence s' and the reference s and calculate a preference score $prefScore = \frac{mtp(s')}{mtp(s')+mtp(s)}$. The decision of whether the $Form$ of a generated sentence s' is acceptable is then calculated as

$$accept = \begin{cases} 1, & \text{if } prefScore \geq 0.5 - tol \\ 0, & \text{otherwise,} \end{cases}$$

where tol is a tolerance parameter. Less formally, a sentence is considered to have an acceptable surface form in relation to its reference if its form is estimated to be at least as good as the reference minus a tolerance, which we fix at 0.05. I.e., the corpus-level $Form$ score reflects the ratio of generated sentences that are of acceptable form.[5]

Predictor selection We consider GPT-2 (Radford et al., 2019), distil GPT-2 (Sanh et al., 2019), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as a basis for assessing $Form$. We conduct experiments on WebNLG (Gardent et al., 2017; Shimorina et al., 2017), which contains human fluency and grammaticality judgements for machine-generated sentences. We find that GPT-2 performs best: it discriminates sentences of poor and perfect fluency and grammaticality with an F1 score of approximately 0.8, and shows marginally better performance compared to the other LMs (see Appendix A.2 for the experiment details). We thus select GPT-2 as our LM for $Form$ assessment.

Discussion While the reconstruction of meaning does not depend on the reference sentence, we do make use of it, in $prefScore$, for better assessment of $Form$. One reason is that when assessing the form of a sentence s' that contains rare words, the 'raw' $mtp(s')$ may be too pessimistic and may not well relate to the quality of the form. Generally, the mtp (or any LM probability) itself is not well interpretable and hardly allows comparison to the mtp of other sentences (e.g., if they are about a different topic). However, by relating the mtp of the generated sentence to the mtp of a (same-topic) reference, we gain three advantages: first, we do not, a-priori, penalize generations that contain rare words. Second, we obtain an interpretable corpus-level ratio (rate of sentences that are of acceptable form). This is important, since sound $\mathcal{MF}_\beta$ calculation ideally requires two interpretable ratios as input. Third, by avoiding any string matching, we still keep form and meaning aspects clearly distinct.

2.4 Goals of our pilot studies

Our main aim is to establish, with the proposed $\mathcal{MF}_\beta$ score for AMR-to-text generation, **i)** a **balanced** and **interpretable assessment** of generated text according to ***Meaning* and *Form***. Yet, as detailed in §2.2 and §2.3, both components depend on a number of **ii) hyperparameters**, such as the parser applied for $Meaning$ reconstruction, or the LM used for $Form$ assessment. These parameters may also be subject to change over time. It is thus important to assess the effects of such factors on metric scores and system rankings. We investigate both aspects of $\mathcal{MF}_\beta$ in **two pilot studies**.

In **the first study**, in §3, we aim to assess the prospects of $\mathcal{MF}_\beta$ when ranking SOTA systems. We will see that $\mathcal{MF}_\beta$ can explain system performance differences by disentangling $Form$ and $Meaning$, an asset that no other metric can offer.

The second study, in §4, investigates the impact of $\mathcal{MF}_\beta$'s dependence on a parser and a LM. We i) investigate the effects of using different parsers, ii) assess the potential suitability of $\mathcal{MF}_\beta$ for other text generation tasks, by ablating the human gold graph from the evaluation and using $\mathcal{MF}_\beta$ to evaluate generated text vs. reference text, and iii) validate the LM's binary predictions for $Form$ in a manual annotation study.

3 Study I: Assessing interpretability

Setup: data & metrics for system ranking We obtain test predictions of several state-of-the-art *AMR-to-text generation systems* on LDC2017T10, the main benchmark for this task: (i) densely con-

[4]We use the mean (instead of the product) because Bryant and Briscoe (2018) find that basing decisions on the mean works well in practice when assessing possible corrections of grammatical errors.

[5]I.e., the $Form$ score for a single sentence with $accept \geq 0.5 - tol$ equals 1.0. If a precise assessment for a single sentence is needed, we can fall back on $prefScore$ (+/- tol).

	abbrev.	BLEU	METEOR	chrF++	BERTsc. F1	*Meaning* ReSMATCH P	*Meaning* ReSMATCH R	*Meaning* ReSMATCH F1	*Form* - %acc.	$\mathcal{MF}_1$ - Eq. 1	$\mathcal{MF}_{0.5}$ - Eq. 1
apprUB	-	-	-	-	-	83.1	80.1	81.5	100	89.8	84.6
Ribeiro et al. (2019)	R'19	$27.9_{(5)}$	$33.2_{(7)}$	$58.7_{(6)}$	$92.7_{(4)}$	76.5	67.7	$71.9_{(6)}$	$51.6_{(5)}$	$60.1_{(5)}$	$66.6_{(5)}$
Guo et al. (2019)	G'19	$27.6_{(6)}$	$33.7_{(6)}$	$57.3_{(7)}$	$92.4_{(7)}$	78.2	70.0	$73.9_{(3)}$	$47.1_{(7)}$	$57.5_{(7)}$	$66.3_{(6)}$
Wang et al. (2020a)	Wb'20	$27.3_{(7)}$	$34.1_{(5)}$	$59.3_{(5)}$	$92.6_{(6)}$	79.6	65.0	$71.5_{(7)}$	$49.5_{(6)}$	$58.5_{(6)}$	$65.7_{(7)}$
Cai and Lam (2020b)	C'20	$29.8_{(4)}$	$35.1_{(4)}$	$59.4_{(4)}$	$92.7_{(4)}$	78.1	69.2	$73.4_{(5)}$	$51.9_{(4)}$	$60.3_{(4)}$	$67.0_{(4)}$
Mager et al. (2020)-M	Mb'20	$33.0_{(2)}$	$37.3_{(2)}$	$63.1_{(3)}$	$93.9_{(2)}$	79.4	68.7	$73.7_{(4)}$	**$74.0_{(1)}$**	**$73.9_{(1)}$**	**$73.8_{(1)}$**
Mager et al. (2020)-L	M'20	$33.0_{(2)}$	**$37.7_{(1)}$**	$63.9_{(2)}$	**$94.0_{(1)}$**	**80.8**	69.2	$74.5_{(2)}$	$69.8_{(2)}$	$72.1_{(2)}$	$73.5_{(2)}$
Wang et al. (2020b)	W'20	**$33.9_{(1)}$**	$37.1_{(3)}$	**$65.8_{(1)}$**	$93.7_{(3)}$	80.3	**70.9**	**$75.3_{(1)}$**	$55.7_{(3)}$	$64.0_{(3)}$	$70.3_{(3)}$

Table 1: Main metric results.

nected graph convolutional networks (Guo et al., 2019); (ii) Ribeiro et al. (2019)'s system that uses a dual graph representation; two concurrently published models (iii) based on graph transformers (Cai and Lam, 2020b; Wang et al., 2020a) and (iv) a model based on graph transformers that uses reconstruction information (Wang et al., 2020b) in a multi-task loss; finally, we obtain predictions of two system variants of Mager et al. (2020) that fine-tune LMs and encode linearized graphs using (v) a large and (vi) a medium-sized LM. We true-case all sentences and parse them with GSII.

To put the results of $\mathcal{MF}_\beta$ into perspective, we display the scores of several metrics that have been previously used for AMR-to-text: BLEU, METEOR, CHRF++. We also calculate BERTscore (Zhang et al., 2020) with RoBERTa-large (Liu et al., 2019).[6] Results are displayed in Table 1, col. 3-6. $\mathcal{MF}_\beta$ scores (col. 7-12) are divided into $Meaning$ (ReSMATCH using GSII) and $Form$ scores (based on GPT-2), and composite $\mathcal{MF}_\beta$ scores with $\beta = 1$ (harmonic) and $\beta = 0.5$ (double weight on $\mathcal{M}$).

As an upper-bound approximation for ReSMATCH we propose parsing a gold sentence s and comparing the result against the gold AMR m: $apprUB$ = metric(parse(s),m).[7]

3.1 Interpretability of system rankings

Surface matching metrics lack differentiation and interpretability Table 1 shows that the baseline metrics tend to agree with each other on the ranking of systems, but there are also differences, for example, BERTscore and METEOR select M'20 as the best performing system while BLEU and CHRF++ select W'20. While certain differences may be due to individual metric properties, e.g., METEOR allowing inexact word matching of synonyms, the underlying factors are difficult to assess, since the score differences between systems with switched ranks are small, and none of these metrics can provide us with a meaningful interpretation of their score that would extend beyond shallow surface statistics. Hence, these metrics cannot give us much intuition about why and when one system may be preferable over another.

$Meaning$ vs. $Form$: How $\mathcal{MF}_\beta$ explains system performance We have seen that current metrics cannot provide us with convincing explanations as to why, e.g., W'20 should be preferred over M'20 (BLEU), or M'20 over W'20 (BERTscore). $\mathcal{MF}_\beta$ score, however, tells a story about how these systems differ, highlighting their complementary strengths by disentangling $Meaning$ and $Form$ (Bender and Koller, 2020): W'20 displays the highest ReSMATCH score, i.e., AMRs constructed from its generations recover a maximum of the meaning contained in the input AMR. M'20, by contrast, outperforms all systems in $Form$ score. Looking at $\mathcal{MF}_1$, the harmonic mean of both, both systems still occupy leading ranks, but W'20 falls back to 3rd rank, due to its weaker $Form$ score.

Hence, given our metric principles, a user who cares about faithfulness to meaning, but less about fluency, should select W'20 (with higher ReSMATCH compared to M'20 by Δ=1 point) – a user who desires a system that preserves meaning well but also produces sentences of decent form, should select M(b)'20 (with $\mathcal{MF}_{0.5}$ and $\mathcal{MF}_1$

[6] BERTscore computes an F1-score over a *cosim*-based alignment of the contextual embeddings of paired sentences.

[7] This is the score of canonical parser evaluation. I.e., we would not expect the reconstruction m' of s' to score higher than had we applied $parse$ to the original sentence: $metric(m', m) \leq metric(parse(s), m)$ = $apprUB$. This is an idealization, as we can imagine cases where the original sentence s is more complex and thus more difficult to parse to an AMR than a simpler generated paraphrase s'. Since we are interested in a very rough upper bound estimation, we abstract from such cases in our present work.

score differences against W'20 of Δ=3.5 points and Δ=8 points). Overall, $\mathcal{MF}_\beta$ mostly agrees with BERTscore in the rankings of the teams. However, $\mathcal{MF}_\beta$'s larger score differences between the systems, due to $Form$, are striking, prompting us to investigate the $Form$ predictions in closer detail (§4.2). We will see that using a different $Form$ predictor as well as a manual native speaker annotation clearly support our assessment of $Form$.

3.2 On the quest for deeper explanation and interpretation

```
----------------------original sent--------------------------
Costa added that insurgents have been holding significant
amounts of opium .
-----------------------original AMR--------------------------
            (a / add-01
               :ARG0 (p / person
                  :name (n / name
                     :op1 "Costa"))
               :ARG1 (h / hold-01
                  :ARG0 (i / insurgent)
                  :ARG1 (o / opium
                     :quant (a2 / amount
                        :ARG1-of (s / significant-02)))))
----------Candidate A---------------Candidate B--------------
Costa added the insurgents         Costa added that
to hold a significant              the insurgents have
amounts of opium .                 held a significant
                                   amount of opium .
------------------------BLEU score---------------------------
           37.7             >>              22.6
---------------------Reconstructions------------------------
(c0 / add-02                  (c0 / add-01
   :ARG0 (c2 / person            :ARG0 (c2 / person
      :name (c4 / name               :name (c5 / name
         :op1 "Costa")                  :op1 "Costa"))
   :ARG1 (c1 / insurgent)        :ARG1 (c1 / hold-01
   :ARG2 (c3 / hold-01              :ARG0 (c4 / insurgent)
      :ARG0 c1                      :ARG1 (c3 / opium
      :ARG1 (c5 / opium                :quant (c6 / amount
         :quant (c6 / amount              :ARG1-of (s / sign.-02
            :ARG1-of (s / sign.-02              )))))
               )))))
-------------------------ReSmatch F1-------------------------
          82.9              <<              100.00
-------------------------------------------------------------
```

Figure 3: Explainable $Meaning$ score (re-)ranking.

ReSmatch can also provide us with **explanations for single-sentence (re-)rankings**. An example is shown in Figure 3. Here, the gold reference (both sentence and AMR) indicates that a person named Costa *adds* (as a communicative act[8] that some insurgents have been holding large amounts of opium. However, system generation A (which is higher ranked by Bleu) chooses a different sense of *add*, add-02, which represents the action as an operation[9], which results in an incoherent or nonsensical meaning representation where the person Costa *adds* (in the operational sense) the insurgent (as thing being added) to a circumstance to the effect that the insurgents hold a significant amount of opium. By contrast, system generation B preserves more of the gold AMR's meaning and clearly expresses that Costa performs an *act of communication* when he *adds* something. ReSmatch ($\mathcal{MF}_{\beta\to 0}$) is able to detect the meaning differences and assigns candidate B a significantly higher score than A, in fact, an S^{2}match score of 1.00.

ReSmatch, when parameterized with fine-grained AMR evaluation metrics of Damonte et al. (2017), can also facilitate deeper insight into **how well system generations reflect or violate specific meaning aspects**. E.g., we can investigate a system's capacity to properly reflect negation (NEG); to generate correct surface forms for NEs (NER); assess how well a system captures coreference between entities (Coref); and whether or not the predicate-argument structures (SRL) of generated sentences appropriately reflect the source meaning. We apply these fine-grained AMR metrics to the ReSmatch scores of systems displayed in Table 1 (see Appendix A.3), and observe, e.g., that R'19, which ranks last in the overall ranking, improves upon the best overall system by 3.4 points in NER recall and 1.9 points in F1. The analysis also corroborates that W'20 excels among competitors with best scores for coreference, SRL and negation, i.e., the more global aspects of sentence meaning. Such information can be valuable for researchers for deeper system analysis and for practitioners aiming for specific use cases.

4 Study II: Assess vulnerability of $\mathcal{MF}_\beta$

$\mathcal{MF}_\beta$ has two apparent vulnerabilities: first, it depends on a parser for reconstruction. We have used a SOTA parser that is on par with human IAA. Yet, we cannot exclude the possibility that it introduces unwanted errors in computing $\mathcal{MF}_\beta$ scores.

Second, the $Form$ component is based on a LM and we have seen that it can change system rankings, even when it is discounted.[10] On the one hand, our LM was carefully selected, and other metrics such as BERTscore also heavily depend on LMs.

[8] Sense add-01 w/ roles: Arg0: *Speaker*; Arg1: *Utterance*.

[9] Sense add-02 w/role set: Arg0: *adder*; Arg1: *thing being added*; Arg2: *thing being added to*; Arg3: *resulting sum*.

[10] In Table 1, both $\mathcal{MF}_\beta$ with $\beta = 0.5$ and $\beta = 1.0$ slightly disagree with the ranks assigned by $Meaning$ only.

On the other hand, we cannot exclude the possibility that the changed rankings are unjustified.

Our next studies investigate these weak spots more closely. First, in §4.1, we assess the outcome of $\mathcal{MF}_\beta$ when using another parser and assess its potential portability to other text generation tasks by ablating the human gold graph and evaluate generated text against reference *text*. In §4.2 we conduct a human annotation study to assess whether the provided $Form$ rankings are justified.

4.1 The parser: Achilles' heel of $\mathcal{MF}_\beta$?

Using another parser In this experiment we assess the robustness of RESMATCH against using different parsers. This is important, since the metric and rankings could change with the parser. Here, we would hope that the difference of using one competitive parser over another will not be too extreme, especially with regard to system rankings. To investigate this issue, we apply two alternative parsers: i) GPLA (Lyu and Titov, 2018), a neural graph-prediction system that jointly predicts latent alignments, concepts and relations, and ii) TTSA (Groschwitz et al., 2018), a neural transition-based parser that converts dependency trees to AMR graphs using a typed semantic algebra. We select GPLA and TTSA since they constitute technically quite distinct approaches compared to GSII.

The results are shown in Table 2 (columns labelled GPLA, TTSA and GSII). All variants tend to agree in the majority of their rankings[11] (e.g., RESMATCH ^{GPLA} vs. RESMATCH ^{GSII} F1: Spearman's ρ = 0.95, Pearson's ρ = 0.96, p<0.001). When considering $\mathcal{MF}_{\beta=0.5}$, the agreement further increases (e.g., $\mathcal{MF}_{0.5}^{GPLA}$ vs. $\mathcal{MF}_{0.5}^{GSII}$: Spearman's $\rho = 0.95$, Pearson's $\rho = 0.99$, p<0.001).

However, while using TTSA or GPLA instead of GSII has little effect on the ranks, the absolute scores can differ (e.g., W'20 70.5 F1 w/ TTSA, 73.1 F1 w/ GPLA and 75.3 F1 w/ GSII). Yet, we find that none of the generation systems are unfairly treated by our main parser GSII since we observe (mostly uniform) increments from TTSA to GPLA and from GPLA to GSII. An unfair treatment could arise, e.g., if GSII generates bad AMR reconstructions for specific NLG systems but not so for others. However, we do not observe such tendencies.

Hence we assume that GSII's score increments stem from the fact that GSII yields better reconstructions for all systems. In future work, we plan to explore parse quality control (Opitz and Frank, 2019; Opitz, 2020) or ensemble parsing (van Noord and Bos, 2017), to gain more detailed information on the quality of the meaning reconstructions.

Ablating the gold graph? Yes, we can. In lack of a gold standard for the automatic reconstructions, we elicit some indirect answers and insight about the parser's quality, by considering the following question: *What is the effect on system rankings when we replace the input gold graphs with automatic parses of the distant source sentence?* If this effect is large, this will give us reasons to worry, as it would indicate that the parser is less reliable than expected given its high IAA with humans. On the other hand, if we only see a minor effect, this may increase the trust in our parser and indicate that $\mathcal{MF}_\beta$ **could be confidently applied for explainable evaluation in other generation tasks** (such as MT or summarization), where we do not have gold AMRs, and would have to parse both generated *and* reference sentences.

The results of this experiment are displayed in Table 2: our standard setup is displayed in columns labeled GSII and the results of the setup where we replace the gold input graph with an automatic parse is indicated by GSII$^\blacklozenge$. When considering RESMATCH scores, we see only one switched rank between Mb'20 and G'20 (3–4). However, note that the absolute F1 score Δ between these two systems is overall very small (GSII: 0.2; GSII$^\blacklozenge$: 0.5). Overall, the scores do not tend to differ much when the gold graph is ablated, we observe rather small (mostly positive) changes in system scores (GSII $\rightarrow$ GSII$^\blacklozenge$): 0.1 / 1.2 / 0.4 (min/max/avg). In sum, we conclude from this experiment that ablating the gold graph does not have a major effects on the scores and rankings. And when considering the $\mathcal{MF}_{\beta=0.5}$ score, the ranking stays fully stable (the same holds true for $\mathcal{MF}_{\beta=1}$.

Discussion We have shown that metric rankings are fairly robust to using different parsers and that we do not necessarily depend on gold AMR graphs to compute the measure. This offers prospects for **using** $\mathcal{MF}_\beta$ **for an explainable assessment of systems that perform other kinds of text generation**. In order to measure $\mathcal{M}$, a parser could be applied to both the generated and the reference text, to measure their agreement in the domain of ab-

[11] We observe one switch of ranks for TTSA-GPLA and GPLA-GSII and 2 rank switches for TTSA–GSII in RESMATCH, and no rank switch for TTSA-GSII and one switch for TTSA-GPLA and GPLA-GSII, for $\mathcal{MF}_{0.5}$.

	ReSmatch F1				ranks ReSmatch				ranks $\mathcal{MF}_{0.5}$			
	TTSA	GPLA	GSII	GSII$^{\blacklozenge}$	TTSA	GPLA	GSII	GSII$^{\blacklozenge}$	TTSA	GPLA	GSII	GSII$^{\blacklozenge}$
$apprUB$	73.7	76.2	81.5	86.4	0	0	0	0	0	0	0	0
R'19	66.9	70.1	71.9	72.3	7	7	6	6	5	5	5	5
G'19	69.7	72.2	73.9	73.7	3	3	3	4	6	6	6	6
Wb'20	67.3	70.2	71.5	71.6	6	6	7	7	7	7	7	7
C'20	69.1	70.4	72.2	73.4	4	5	5	5	4	4	4	4
Mb'20	68.9	70.5	73.7	74.2	5	4	4	3	1	2	1	1
M'20	69.8	72.5	74.5	75.1	2	2	2	2	2	1	2	2
W'20	70.5	73.1	75.3	75.4	1	1	1	1	3	3	3	3

Table 2: Analysis of our metric using different parsers (GPLA, TTSA GSII) or ablating the gold parse by comparing the parsed generation against the parse (distant) source sentence (GSII$^{\blacklozenge}$).

stract meaning representation. This would in turn offer means for conducting fine-grained meaning analysis of generation tasks where the reference is a natural language sentence (e.g., in MT).

Note, however, that AMR, as of now, does not capture some facets of meaning that may be of interest in some generation tasks. For instance, it does not capture tense or aspect. However, what we have investigated as a *potential weakness* of $\mathcal{MF}_\beta$, namely the necessity to select a meaning parser, can also be viewed as a *potential strength*. E.g., Donatelli et al. (2018) show how tense and aspect can be captured with AMR. This indicates that $\mathcal{MF}_\beta$ can indeed be used for a tense and aspect analysis of generated text – if we parameterize it with a dedicated parser. Finally, if output and reference do not consist of single sentences, it may be apt to use a parser that constructs MRs for discourse (e.g., *DRS* (Kamp, 1981)).

In summary, we conclude that $\mathcal{MF}_\beta$, our proposed metric that aims to assess text generation quality by decomposing it into *form* and *meaning* aspects, is broadly applicable. However, different parser parametrizations may have to be considered in light of the specific nature of a generation task.

4.2 The $Form$ component of $\mathcal{MF}_\beta$

In §3.1, we have seen that the $Form$ aspect of $\mathcal{MF}_\beta$ can change system ranks. Notably, it has promoted M'20 as the best generation system, outranking W'20 (in agreement with BERTscore), whereas W'20 is selected by Bleu or ReSmatch. Now, we aim to investigate whether these impactful decisions of the $Form$ component were justified.

Human annotation We ask a native speaker of English to rate 50 paired generations of M'20 and W'20, considering only grammaticality and fluency.[12] We give more detail and provide examples in Appendix A.6. The annotator agreed in 42 of 50 pairs with the preference predicted by GPT-2 (a significant result: binomial test p<0.000001). We find that the M'20 and Mb'20 generations are considerably better on the surface level, compared to generations of all other systems. For instance, the best system according to $Meaning$, W'20, frequently produces inflection mishaps: *Their hopes for entering the heat is already in-sight*, while we find few such violations with M'20 (here: *Their hopes for entering the heat are already in sight*). We also find errors with adverbials, e.g., W'20 writes *They are the most indoor training at home*, while M'20 writes *They are most trained indoors at home.* Arguably both sentences are not perfect but the second is substantially more well-formed.

	R'19	G'20	Wb'20	C'20	Mb'20	M'20	W'20
GPT-2	$51.6_{(4)}$	$47.1_{(6)}$	$49.5_{(5)}$	$51.9_{(4)}$	$74.0_{(1)}$	$69.8_{(2)}$	$55.7_{(3)}$
BERT	$43.4_{(6)}$	$40.6_{(7)}$	$50.4_{(4)}$	$44.7_{(5)}$	$71.4_{(1)}$	$71.0_{(2)}$	$55.9_{(3)}$

Table 3: $Form$ scores when using a different LM.

Using a different LM The human study indicates that GPT-2 is accurate to 84% when favoring one sentence over the other, with respect to fluency and grammaticality. However, when considering that there is a trend to building systems based on fine-tuned LMs, we need to assess whether they may be favored (too) much if $Form$ is parameterized with a same or a highly similar LM to the one used by the NLG model. We find such a case in M'20: while it was not fine-tuned with the same GPT-2 that we used for $Form$ assessment, they fine-tuned their model with its siblings GPT-2-medium and GPT-large, which may share structural similarities. Therefore, we also use BERT for

[12] The annotator was explicitly instructed not to consider whether a sentence 'makes sense', by presenting the *Green ideas sleep furiously* example as free from structural error.

$Form$ assessment. The results in Table 3 support the conclusion from the human annotation: by large margins, both M'20 and Mb'20 deliver generations that are of significantly improved form and both agree on the group of the three best systems. Note that this insight can be provided by $\mathcal{MF}_\infty$, but it cannot be carved out by conventional metrics, since these do not disentangle $Form$ and $Meaning$.

5 Related work

Traditionally, the performance of NLG systems has been evaluated with word n-gram matching metrics such as the popular BLEU metric in MT (Papineni et al., 2002) or Rouge (Lin, 2004) in document summarization. Yet, such metrics suffer from several well-known issues (Novikova et al., 2017; Nema and Khapra, 2018; Sai et al., 2020). E.g., due to their symbolic matching strategy they cannot account for paraphrases. Recently, unsupervised (Zhang et al., 2020) or learned metrics (Sellam et al., 2020; Zhou and Xu, 2020) based on contextual language models have been proposed. For example, BERTscore (Zhang et al., 2020) uses BERT (Devlin et al., 2019) to encode candidate and reference and computes a score based on a cross-sentence word-similarity alignment. Compared with BLEU, it is computationally more expensive but tends to show higher agreement with humans. However, *all* of the aforementioned metrics return scores that are hardly interpretable and we cannot tell what exactly they have measured.

These problems carry over to the evaluation of AMR-to-text generation: May and Priyadarshi (2017) find that BLEU does not well correspond to human ratings of generations from AMR, and Manning et al. (2020) show through human analysis that none of the existing automatic metrics can provide nuanced views on generation quality. Our proposal $\mathcal{MF}_\beta$ takes a first step to address these issues by aiming at a clear separation of form and meaning, as called for by Bender and Koller (2020).

First attempts of assessing semantic generation quality have been examined in MT using semantic role labeling (Lo, 2017) or WSD and NLI (Carpuat, 2013; Poliak et al., 2018), in-between lies SPICE that evaluates caption generation via inferred semantic propositions (Anderson et al., 2016). Just like $\mathcal{MF}_\beta$, SPICE relies on automatic parses (a dependency parse of the caption and a scene graph predicted for the image) to evaluate content overlap of image and caption. Thus, SPICE is a direct precursor of an NLG metric in V&L that relies on automatically produced structured representations. Our work extends this previous work by showing ways of probing potentially harmful effects of incorporating automatic parsing components.

6 Conclusion

We propose $\mathcal{MF}_\beta$ score, a new metric for evaluation of text generation from (abstract) meaning representation. The metric is built on two pillars: $Form$ measures grammaticality and fluency of the produced sentences and $Meaning$ assesses to what extent the meaning of the input AMR is reflected in the produced sentence. We show that $\mathcal{MF}_\beta$ has the potential to yield fine-grained performance assessment that go beyond what conventional metrics can provide. Using its β-parameter, $\mathcal{MF}_\beta$ can be decomposed into complementary views – $Meaning$ and $Form$ – paving the way for custom gauging and selection of NLG systems. We have seen that $\mathcal{MF}_\beta$ corresponds well to BERTscore when rankings systems, but overcomes its opaqueness by disentangling $Meaning$- and $Form$-related quality aspects. In sharp contrast to BERTscore, the $Form$ component of $\mathcal{MF}_\beta$ dispenses with string matching against reference sentences, offering an assessment independent of lexical alignment.

An important hyperparameter of our metric is the required AMR parsing component for meaning reconstruction. We investigate the impact of its choice by choosing alternative high-performing parsers. Our study shows that absolute metric scores tend to increment when using a better parser, while system rankings are quite stable. Furthermore, we outline the potential of $\mathcal{MF}_\beta$ to extend to further text generation tasks, by ablating the human gold graph from the evaluation, such that the metric score can be computed from candidate and reference text alone. Since benchmarking of systems needs deeper exploration, we recommend $\mathcal{MF}_\beta$ score to obtain better diagnostics and explainability of text generation systems, including, but not limited to (A)MR-to-text.

Acknowledgments

We are grateful to three anonymous reviewers for their valuable comments that have helped to improve this paper. This work has been supported by the *Deutsche Forschungsgemeinschaft (DFG)* through the project *ExpLAIN*, FR 1707/4-1 as part of the RATIO Priority Program (SPP-1999).

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pages 382–398, Cham. 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.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253, New Orleans, Louisiana. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020a. 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.

Deng Cai and Wai Lam. 2020b. Graph transformer for graph-to-sequence learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7464–7471. AAAI Press.

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.

Marine Carpuat. 2013. A semantic evaluation of machine translation lexical choice. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–10, Atlanta, Georgia. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Jonas Groschwitz, Matthias Lindemann, Meaghan Fowlie, Mark Johnson, and Alexander Koller. 2018. AMR dependency parsing with a typed semantic algebra. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1831–1841, Melbourne, Australia. Association for Computational Linguistics.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019. Densely connected graph convolutional networks for graph-to-sequence learning. *Transactions of the Association for Computational Linguistics*, 7:297–312.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.

Hans Kamp. 1981. A theory of truth and semantic representation. In J. Groenendijk, Th. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematisch Centrum. Reprinted in "Meaning and the Dynamics of Interpretation - Selected Papers of Hans Kamp" (Kamp (2013)) p. 329 -369.

Anisia Katinskaia and Sardana Ivanova. 2019. Multiple admissibility: Judging grammaticality using unlabeled data in language learning. In *Proceedings of* the 7th Workshop on Balto-Slavic Natural Language Processing, pages 12–22, Florence, Italy. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Maarit Koponen, Leena Salmi, and Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, 33(1-2):61–90.

Jey Han Lau, Carlos S Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colourless green ideas sleep? sentence acceptability in context. *arXiv preprint arXiv:2004.00881*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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. *CoRR*, abs/1907.11692.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

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.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. In *Proc. of ACL*, pages 1846–1852, Online. Association for Computational Linguistics.

Emma Manning, Shira Wein, and Nathan Schneider. 2020. A human evaluation of AMR-to-English generation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4773–4786, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Jonathan May and Jay Priyadarshi. 2017. SemEval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545, Vancouver, Canada. Association for Computational Linguistics.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Rik van Noord and Johan Bos. 2017. The meaning factory at SemEval-2017 task 9: Producing AMRs with neural semantic parsing. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 929–933, Vancouver, Canada. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. 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 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, Anette Frank, and Letitia Parcalabescu. 2020. Amr similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8(0):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.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 513–523, New Orleans, Louisiana. Association for Computational Linguistics.

Alexander Popov. 2017. Word sense disambiguation with recurrent neural networks. In *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 25–34, Varna. INCOMA Ltd.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2020. A survey of evaluation metrics used for nlg systems. *arXiv preprint arXiv:2008.12009*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Anastaisa Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: report on human evaluation. Technical report, Technical report, Université de Lorraine, Nancy, France.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. *arXiv preprint arXiv:2005.03642*.

Tianming Wang, Xiaojun Wan, and Hanqi Jin. 2020a. Amr-to-text generation with graph transformer. *Transactions of the Association for Computational Linguistics*, 8(0):19–33.

Tianming Wang, Xiaojun Wan, and Shaowei Yao. 2020b. Better amr-to-text generation with graph structure reconstruction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3919–3925. International Joint Conferences on Artificial Intelligence Organization. Main track.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9717–9724. AAAI Press.

Wanzheng Zhu and Suma Bhat. 2020. Gruen for evaluating linguistic quality of generated text. *arXiv preprint arXiv:2010.02498*.

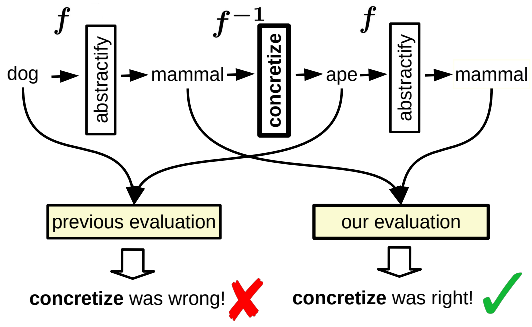

Figure 4: A critical issue and its alleviation.

A Appendices

A.1 On the soundness of comparing generated sentences in the AMR domain

First, we provide a simple example for our argument (it is safer to compare texts generated from AMR in the AMR domain) and then a simple proposition together with its proof. The example is displayed in Fig. 4, where, similar to AMR-to-text, we see a (surjective) function that generates concrete objects from abstract objects (e.g., $mammal \rightarrow \{dog, mouse, cow\}$). Now, imagine we are given $mammal$ and are tasked with generating a single concrete instance. How can we assess whether our output is correct? We cannot safely assess this by testing whether the output (e.g., cow) is the same as another instance of $mammal$ (e.g., dog). Instead, we can re-apply the abstraction f to cow and conduct the comparison safely in the abstract domain.

Proposition. *a) The canoncical AMR-to-text evaluation setup, that matches generated sentence s' to distant source sentence s, is not well defined. b) This issue can be alleviated by grounding the evaluation in the AMR domain by re-appling parse, abstaining from direct use of s (thereby using AMR-to-text generation as a right inverse function).*

Proof. Let X be a set of concrete objects (e.g., sentences) and f a (surjective) function from X to Y (e.g., 'sent-to-AMR'), where Y contains abstract objects (e.g., AMRs), s.t. $|Y| < |X|$. Then, using $f^{-1} : Y \rightarrow X$ (e.g., 'AMR-to-sent')) as right-inverse is well-defined: $f \circ f^{-1} = id_Y$ (Proposition b), but using it solely as left-inverse (as done in previous evaluation) does not guarantee a well-defined result: $f^{-1} \circ f \neq id_X$ (Proposition a). □

A.2 Form predictor selection experiment

To estimate how well they are able to assess $Form$, we make use of human-assigned scores for data

	F1 score			
	grammaticality		fluency	
LM	poor/perfect	all	poor/perfect	all
GPT2	**0.80**	**0.74**	**0.80**	0.71
GPT2-distill	0.79	0.73	0.76	0.70
BERT	**0.80**	0.72	**0.80**	**0.72**
RoBERTa	0.66	0.72	0.69	**0.72**

Table 4: Results for assessing the $Form$ score prediction (corpus-level) of different LMs for NLG-generated sentences against humans judgements (separated by grammaticality and fluency); all: all 12k generated sentences vs. 'poor/perfect': the 5k instances of best/worst generations in both grammaticality and fluency.

	Reentrancies			SRL			negation			NER		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
$apprUB$	72.1	60.7	65.9	77.7	73.5	75.5	88.6	70.5	78.5	82.2	80.1	81.1
R'19	63.7	50.3	56.2	71.1	62.4	66.4	72.1	50.6	59.5	82.2	**70.7**	**76.0**
G'19	66.9	52.9	59.1	73.7	64.9	69.0	75.0	51.5	61.1	78.6	68.9	73.5
Wb'20	67.6	51.5	58.4	75.1	63.6	68.9	74.3	49.7	59.6	86.5	60.3	71.0
C'20	66.1	52.4	58.4	73.4	64.8	68.8	78.3	54.2	64.1	80.8	67.2	73.4
Mb'20	65.9	53.2	58.9	74.3	65.7	69.8	70.6	45.5	55.3	82.6	69.4	75.4
M'20	67.9	53.3	59.7	**76.4**	66.5	71.1	73.7	53.9	62.3	**82.8**	68.3	74.9
W'20	**68.8**	**55.7**	**61.6**	76.1	**68.1**	**71.9**	**79.2**	**55.1**	**65.0**	82.4	67.3	74.1

Table 5: Fine-grained results using $\mathcal{MF}_0$ parameterized with metrics proposed by Damonte et al. (2017).

from the WebNLG task as provided by Gardent et al. (2017). It contains grammaticality and fluency judgments by humans for more than 2000 machine-generated sentences. We report the F1 score, both for grammaticality and fluency, by converting the human assessment scores to *accept* predictions, and using them as a gold standard to evaluate the LM-based *accept* predictions over (i) all 12k sentence pairs[13] and (ii) only the 5k sentence pairs where both grammaticality and fluency where either rated as 'perfect' (max. score) or 'poor' (min. score) by the human.[14]

The results are displayed in Table 4 and show (i) that the LMs lie very close to each other with respect to their capacity to predict fluency and grammatically, and (ii) that both fluency and grammaticality can be predicted fairly well.

A.3 ResMatch with fine-grained meaning metrics

Using Damonte et al. (2017)'s metric suite for fine grained semantic system analysis, we obtain fine-grained results with respect to various meaning aspects of system performance. The results are shown in Table 5.

In sum, the system of W'20 appears to be the

[13] This includes all generated sentences from a given input, as provided by Gardent et al. (2017); Shimorina et al. (2017)

[14] The ratings are based on a 3-point Likert scale.

```
-----------------------original sent------------------------
   Since there is responsibility, we are not afraid.
-----------------------original AMR-------------------------
            (c / cause-01
               :ARG0 (r / responsible-02)
               :ARG1 (f / fear-01
                  :polarity -
                   :ARG0 (w / we)))
------------Candidate 1-------------Candidate 2------------
We are not responsible                  We are not afraid
because we fear .                    for responsibility .
-------p^A=f(A)------Reconstructions--------p^B=f(B)--------
(c1 / cause-01                  (c1 / fear-01
   :ARG0 (c5 / fear-01)            :ARG0 (c5 / we)
   :ARG1 (c4 / responsible-01      :ARG1 (c4 / responsible-03
      :ARG0 (c10 / we)                :ARG0 c5)
      :polarity -))                :polarity -)
------------------------Negation F1--------------------------

         negationF1 = 0.00  <<   negationF1 = 100
-------------------------------------------------------------
```

Figure 5: Explained negation confusion.

clear winner in most aspects of meaning. This is intuitive, since the system has been trained with an auxiliary signal that provides information on how well an AMR can be reconstructed from the generated sentence.

A.4 ReSmatch explains negation error

In Figure 5, both systems struggle to fully capture the meaning of the original AMR $f(s)$. However, the system based on GPT medium (Mb'20) erroneously assesses that *we are not responsible* and *we fear*. However, quite the opposite is true: the gold graph and gold sentence states that *there is responsibility* and *there is no fear*. This important facet of meaning is better captured by C'20. The reconstruction shows that it reflects the gold negated concepts much better and does not distort facts that are core to the meaning. In consequence, the negation F1 is zero for the left sentence with the distorted facts and maximum for the sentence that sticks true to the facts.

A.5 ReSmatch explains SRL error

Figure 6 shows an example, were ReSmatch ranks two generated candidate sentences differently compared to Bleu. In this case, gold sentence and gold AMR both express that there is some soldier who tried to defuse a bomb and got injured in the process. Clearly, candidate generation A captures the meaning better, in fact, it captures it almost perfectly. However, since the surface text deviates from the gold sentence, Bleu overly penalizes this generation and assigns a very low score of 10.6 points. In contrast, candidate B matches the surface slightly better (12.2 points), but distorts the meaning: it does not contain any information about the soldier and states that *Disarming was injured*, which is grammatically correct, but semantically wrong, or even non-sense.

We see that the surface matching metric cannot explain its scores (beyond superficial statistics) and delivers a ranking that does not appropriately reflect the performance of the generation systems. However, ReSmatch shows that the gold parse and the parse of candidate A agree with each other in the central *ARG1*-role of the main predicate *injure-01*: *it is the soldier who got injured*. On the other hand, in the reconstruction of the AMR of candidate B, the *ARG1* argument is filled differently: *it is the disarmament that gets injured*.

This assessment allows ReSmatch to increment the score for generation A by a large margin, from 10.6 (Bleu) to 93.3 points (ReSmatch), expressing substantial agreement in meaning with the gold. The score for the candidate generation B also gets incremented – but it gets incremented much less, only to 70.2 points, expressing good to mediocre agreement. Thus, by detecting the SRL confusion, ReSmatch re-ranks the candidate generation such that the resulting ranking is more appropriate.

```
----------------------original sent-----------------------
Soldier injured during bomb defusion in Kathmandu after
state of emergency expires .
-----------------------original AMR-----------------------
          (i / injure-01
             :ARG0 (d / defuse-01
                :ARG1 (b / bomb)
                :location "Kathmandu")
             :ARG1 (s / soldier)
             :time (a / after
                :op1 (e / expire-01
                   :ARG1 (s2 / state
                      :mod (e2 / emergency
                              )))))
----------Candidate 1-------------Candidate 2--------------
The Soldier was injured           Disarming the bomb in
                                  Kathmandu was injured
in the defuse of the bomb         in Kathmandu after state
in Kathmandu after the            of emergency expires .
emergency state expired .
-----------------------BLEU score-------------------------
     score(A,s) = 10.6     <<      score (B,s) = 12.2
--------------------Reconstructions-----------------------
(c0 / injure-01               (c0 / injure-01
   :ARG1 (c1 / soldier)          :ARG1 (c1 / disarm-01
   :ARG2 (c2 / defuse-01            :ARG1 (c4 / bomb))
      :ARG1 (c4 / bomb)          :location "Kathmandu"
      :location "Kathmandu"      :time (c2 / after
      )                             :op1 (c5 / decline-02
   :time (c3 / after                   :ARG1 (c7 / state-01
     :op1 (c6 / expire-01                 :location c3
       :ARG1 (c8 / state                  :mod (c8 / emergency
          :mod (c9 / emergency            ))))))
          )))))
------------------------ReSmatch F1-----------------------
                  93.3     >>     70.2
----------------------------------------------------------
```

Figure 6: Explained SRL confusion.

A.6 Annotation study for form assessment

```
Sys (W'20): He also said that our athletes do n't very use of competition under strong sunlight .
Corr (human): He also said that our athletes are not very used to competition under strong sunlight .
---->  not acceptable

Sys (W'20): Sheng Chen , the 6 th position of Hubei province , who was totally scored 342.60 at 342.60 points this year ,
is a temporary position .
Corr (human): Sheng Chen , the 6 th position of Hubei province , who has totally scored 342.60 points this year ,
is in a temporary position .
---->  not acceptable

Sys (W'20): The Chinese competitors are Lan Wei and Sheng Chen , qualify semi - final .
Corr (human):  The Chinese competitor Lan Wei and Sheng Chen qualify for the semi - final .
---->  acceptable

Sys (M'20): Fengzhu Xu won many championships in international competition before .
Corr (human): Fengzhu Xu won many championships in international competitions before .
---->  acceptable
```

Figure 7: Sentences of flawed form. `--->` refers to the binary acceptability judgment (Eq. 2.3).

Annotator and annotation The English native speaker (UK) annotated 50 paired sentences of M'20 and W'20. They were presented in shuffled order and the annotator was tasked with assigning a nominal number, starting from zero, that indicates the amount of grammatical or fluency issues as assessed by the native speaker. Additionally, the human was asked to provide a correction.

Examples of sentences of flawed form. See Figure 7.