# 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 sentences from (abstract) meaning representations (AMRs) are typically evaluated using automatic surface matching metrics that compare the generated texts to the texts that were originally given to human annotators to construct AMR meaning representations. However, besides well-known issues from which such metrics suffer (Callison-Burch et al., 2006; Novikova et al., 2017), we show that an additional problem arises when applied for AMR-to-text evaluation because mapping from the more abstract domain of AMR to the more concrete domain of sentences allows for manifold sentence realizations. In this work we aim to alleviate these issues and propose $\mathcal{MF}_\beta$, an automatic metric that builds on two pillars. The first pillar is the **principle of meaning preservation** $\mathcal{M}$: it measures to what extent the original AMR graph can be reconstructed from the generated sentence. We implement this principle by i) automatically constructing an AMR from the generated sentence using state-of-the-art AMR parsers and ii) apply fine-grained principled AMR metrics to measure the distance between the original and the reconstructed AMR. The second pillar builds on a **principle of (grammatical) form** $\mathcal{F}$, which measures the linguistic quality of the generated sentences, which we implement using SOTA language models. We show – theoretically and experimentally – that fulfillment of both principles offers several benefits for evaluation of AMR-to-text systems, including the explainability of scores.

## 1 Introduction

Abstract Meaning Representation (short: AMR) (Banarescu et al., 2013) aims at capturing the meaning of a sentence in a machine-readable graph format. For instance, the AMR in Figure 1 represents a sentence such as *"Perhaps, the parrot is telling herself a story?"*. Among other phenomena, AMR captures predicate senses, semantic roles, coreference and utterance type. In the example, tell-01 links to a PropBank (Palmer et al., 2005) predicate sense, and arg$_n$ labels indicate participant roles: *parrot* is both *speaker* (arg0) and *hearer* (arg2), *story* is the *utterance* (arg1).

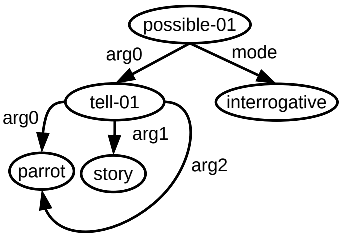


Figure 1: An AMR graph.

AMR-to-text generation is a task that has garnered lots of attention over the recent years (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 AMRs 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 employed in general NLG tasks. However, such metrics are suffer from several issues, for example, they are highly sensitive to the translations used for assessment which may easily lead to falsely confident conclusions about a metrics efficacy (Callison-Burch et al., 2006; Mathur et al., 2020).

Moreover, we find that this sensitivity to reference sentences is aggravated when evaluating AMR-to-text. The root cause lies in the fact that there are manifold ways to realize a sentence from a meaning representation. For example, in Figure 2, we see four candidate sentences (**i-iv**) generated from an AMR (left). In this case, one system generates **i**: *Maybe the cat is playing.* while another sys-

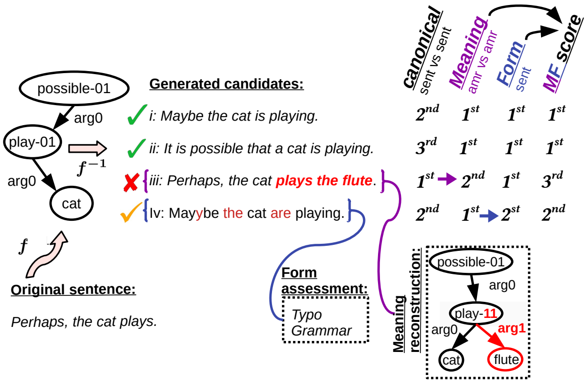


Figure 2: The ***Canonical*** evaluation matches n-grams from the sentences and assigns inappropriate ranks. Our metric $\mathcal{MF}_\beta$ fuses ***Meaning*** and ***Form*** assessment and better reflects the rankings of the generations.

tem generates **iii**: *Perhaps, the cat plays the flute.*. Clearly, **i** better captures the meaning contained in the gold graph (left side) compared to **iii**, which contains 'hallucinated' content – a severe issue in neural generation models that is hard to detect (Koehn and Knowles, 2017; Dušek et al., 2019; Nie et al., 2019; Logan et al., 2019; Wang and Sennrich, 2020). Now, when we use a Canonical surface matching metric (here: BLEU)[1], we evaluate the Original sentence against the Generated sentence. Yet, when comparing **i** and **iii** against the original sentence, the system that produces the hallucinating sentence (**iii**) is greatly rewarded (Δ: +36 BLEU points to the disadvantage of the systems that produce the meaning preserving sentences (**i**) (only 18 BLEU points) and (**ii**) (only 5 BLEU points).

In conclusion, we want to aim at a (better) metric that **measures meaning preservation** of the generated output towards the MR given as input; we do this by (re-)constructing an AMR from the generated sentence and comparing it to the input AMR. In Figure 2, Reconstruction is the result of parsing **iii**. We see that the reconstructed AMR is flawed, in the sense that it deviates from the original meaning representation. Specifically, **iii** misrepresents the sense of *play* (01 vs. 11) and hallucinates a semantic role (arg1) with filler *flute*. By contrast, when converting sentences (**i**, **ii**, or **iv**) to AMRs, we obtain flawless reconstructions. We will measure their preservation of ***Meaning*** using well-defined graph matching metrics.

However, Figure 2 also illustrates that assessing meaning preservation will not be sufficient to rate the quality of the generated sentence: sentence (**iv**) captures the meaning of the AMR perfectly – but its form is flawed: it contains a typo and wrong verb inflection, a common issue (especially) in low-resource text generation settings (Brussel et al., 2018; Koponen et al., 2019; Matusov, 2019). In order to rate both meaning and form of a generated sentence, we combine the score for meaning reconstruction with a score alled ***Form*** that allows us to **judge the sentence's grammaticality and fluency**. By this move we aim at an explainable and more suitable ranking with a combined ***MF*** score (last colum: 1st/2nd: **i**; 3rd: **iv**; 4th: **iii**).

Generally, our contributions are as follows:

- We propose two linguistically motivated principles that aim at a sound evaluation of AMR-to-text systems: (i) the principle of meaning preservation and (ii) the principle of (grammatical) form.
- From these complementary principles we derive and implement a (novel) $\mathcal{MF}_\beta$ **score for AMR-to-text generation** which composes its score based on individual measurements of meaning and form aspects. $\mathcal{MF}_\beta$ allows users to modulate these two views on generation quality with respect to their impact on the final metric score.
- We conduct two major pilot studies involving a range of competitive AMR-to-text generation systems and human annotations. In the first study, we investigate the potential practical benefits of $\mathcal{MF}_\beta$ when assessing systems, such as its prospects to offer interpretability of metric scores and finer-grained system analyses. In the second study, we assess potential weak spots of $\mathcal{MF}_\beta$ , for example, its dependence on a strong AMR parser.

We will release all code and data.

## 2 Related work

Many NLP tasks involve generation of text, e.g., from text to text as in machine translation or document summarization, or generation of sentences from structured content, such as data-to-text generation in its most general form (tables or graphs), or generation from meaning representations such as AMR or DRT. Traditionally, the performance of such systems has been evaluated with word n-gram matching metrics such as the popular BLEU

[1] With NIST geometric sentence probability smoothing (Chen and Cherry, 2014).

metric in MT (Papineni et al., 2002) or Rouge (Lin, 2004) in document summarization. Alternatively, researchers use character n-gram matching metrics such as crf (Stanojević et al., 2015; Popović, 2015, 2016; Popov, 2017). Yet, such metrics suffer from several well-known issues (Callison-Burch et al., 2006; Novikova et al., 2017; Mathur et al., 2020), for instance, they depend on symbolic matching, greatly penalizing equivalent generations that differ from the gold reference in surface form. The issues may become aggravated in settings where one maps from more abstract input to more concrete output, including, but not limited to AMR-to-text, Table-to-text (Liu et al., 2017; Parikh et al., 2020) or knowledge to text (Koncel-Kedziorski et al., 2019). Recently, unsupervised (Zhang* et al., 2020) or learned metrics (Sellam et al., 2020) based on contextual language models have been proposed. For example, the BERTSCORE (Zhang* et al., 2020) metric uses BERT (Devlin et al., 2019) to encode the candidate and the reference sentence and computes the score based on an a cross-sentence word-similarity alignment. This metric is computationally more expensive but tends to show higher agreement with human raters. Yet, all of these metrics have in common that they lack explainability and interpretability and are not well applicable for encoding an AMR graph.

First practical attempts of assessing sentence quality with semantic assessment 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).

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

Comparing sentences with surface matching metrics suffers from several well-known issues (see Section 1 and Section 2). Now, we will focus on another critical aspect of such metrics that is specific to tasks that map abstract input to natural language output (as in AMR-to-text). Equipped with this background, we start building our $\mathcal{MF}_\beta$ score which targets the alleviation of these issues.

**An issue of the typical evaluation setup that is specific to generating text from more abstract input** Let us first denote the process of creating AMRs from sentences as $parse \equiv abstractify \equiv f$ and the process of generating sentences from such abstract representation as $generation \equiv concretize \equiv f^{-1}$. When evaluating AMR-to-text generation approaches, researchers typically assess how well the generated sentence $s'$ matches the sentence $s$ from whence the original AMR was created: $s' = generate(parse(s))$ is matched against $s$. However, there is a problem with this approach, because we map from an abstract input to a more concrete output with a 'one-to-many mapping'. This means that means that there are possibly many different valid sentences. So, in order to really assess whether two produced sentences are both valid outputs from the same AMR structure, we need to perform this assessment in the AMR domain. This can be achieved by applying an inverse system that generates AMR from text (a parser). Put differently, consider that a system $f$ has generated $s'$ from an AMR $p = f(s)$, then we would like that a $metric : \mathcal{D} \times \mathcal{D} \rightarrow [0, 1]$ satisfies the following equivalence: $s \equiv s' \iff f^{-1}(s) = p \iff metric(s, s') = 1$. Two outputs are equivalent if they lead to the same abstract meaning construction. This also means that we can consider the actual source sentence as *distant*, i.e., we may never use it directly.

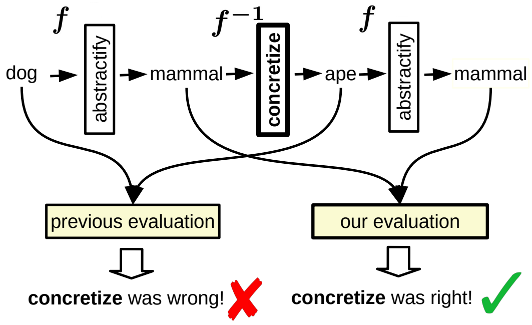


Figure 3: A critical issue and its alleviation.

This is exemplified in Figure 3, where, in an analogue to AMR-to-text, we see a (surjective) function that generates concrete objects from abstract objects (e.g., $mammal \rightarrow \{dog, mouse, cow\}$). Now, consider that we are given $mammal$ and are tasked with generating a single concrete instance. How can we assess whether our output is correct? We observe that this cannot safely happen by testing whether the output (e.g., $cow$) equals another instance of $mammal$ (e.g., $dog$). However, we can use $f^{-1}$ as a right-inverse function, re-applying the abstraction $f$ to convert the concrete instance back to an abstract object.

### 3.1 From principles to $\mathcal{MF}_\beta$

To alleviate the issue described above, we first introduce our

**Principle of meaning.** *Generated sentences should allow loss-less AMR reconstruction.*

This principle expresses a key expectation that we formulate for a system that generates NL sentences from abstract meaning representations. Namely, the generated sentence should reflect the meaning of the AMR.

However, this principle alone is not sufficient: we also expect the system to generate grammatically well-formed and fluent text. For example, the following system output: *Possibly, it(self) tells parrot a story.* contains relevant content expressed in the AMR of Figure 1, but it is neither grammatically wellformed, nor a natural and fluent sentence. This leads us to our

**Principle of form.** *Generated sentences should be syntactically well-formed, natural and fluent.*

In the style of the well-established $F_\beta$ score, 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, $\beta$ allows the user to gauge the evaluation towards $Form$ or $Meaning$, accounting for their specific application scenario. We anticipate that most users will prefer the harmonic mean ($\beta$ = 1), or giving $Meaning$ a higher emphasis compared to $Form$ (e.g., by setting $\beta$ = 0.5). However, in our experiments we will also consider extreme decompositions into $Meaning$-$only$ ($\beta \rightarrow 0$) or $Form$-$only$ ($\beta \rightarrow \infty$).

### 3.2 Parameterizing meaning

We propose to measure $Meaning$ with a score range in $[0, 1]$ by (i) reconstructing the AMR with a state-of-the-art parser and computing the relative graph overlap of the reconstruction and the source AMR using graph matching. We call this a ResMatch. I.e., given a generated sentence $s'$ and source AMR $p$, we match $parse(s')$ against $p$ and compute $Meaning = amrMetric(parse(s'), p)$. This means that we have to decide upon $parse$ and $amrMetric$. We propose two potential settings.

**AMR reconstruction** To reconstruct the AMR using $parse$, we will be using the parser by Cai and Lam (2020a), henceforth denoted as GSII, as it constitutes the latest state-of-the-art AMR parser. Based on IAA estimates by Banarescu et al. (2013), this parser (80.3 Smatch F1[2]) is almost on-par with human agreement (estimates range between 0.71 and 0.83 Smatch F1).

**AMR metric for reconstruction assessment** To gain a single $Meaning$ score we propose to use S$^2$match (Opitz, 2020) that is based on the canonical AMR evaluation metric Smatch (Cai and Knight, 2013). It is essentially the same as Smatch except that it uses a graded match for concept nodes. This offers the potential to compensate for some unwanted noise in automatically generated text[3] or lexical deviations from the original sentence.

**Discussion** All in all, $\mathcal{MF}_\beta$ leaves researchers a lot of flexibility as to which $parse$r or $amrMetric$ they prefer. For our $parser$, we aimed at the possibly best one that achieves high IAA with humans. However, while this property makes it most suitable at first glance, we would also like to know whether the $parser$ is vulnerable to specific peculiarities of generated sentences. Moreover, we would like to have knowledge about the impact on the performance assessment of $\mathcal{MF}_\beta$ when we use another parsing system. Therefore, we will investigate these issues more closely in Section 5.1. With regard to the $amrMetric$, there certainly exist use-cases for other metrics, or custom metrics. For example, Anchiêta et al. (2019); Song and Gildea (2019) propose metrics that aim at faster evaluation by ablating the costly variable-alignment. This may prove valuable when one wants to apply $\mathcal{MF}_\beta$ on large corpora.[4]

### 3.3 Parameterizing form with LMs

Assessing the (related) aspects of sentence grammaticality and fluency is not an easy task (Heilman et al., 2014; Dickinson and Ragheb, 2015; Katinskaia and Ivanova, 2019). Recently, Lau et al. (2020) show that probability estimates based on language models can be used as an indicator for

[2] measured on the standard benchmarking corpus

[3] E.g., due to stemming mishaps of rare words: *bacteria* and *bacterium* are not allowed to match with Smatch, but S$^2$match considers them as similar and is more benevolent to such slight deviations.

[4] One may also fall back on graph metrics for other meaning representations, e.g., the evaluation of DRS (Kamp, 1981; Kamp and Reyle, 2013) is conducted similarly to Smatch (van Noord et al., 2018) but suffers from more complexity due to larger graphs. Liu et al. (2020) therefore develop a more efficient metric that circumvents the costly alignment.

measuring complex notions of form, measuring acceptability in context with LMs. Here, we want to measure $Form$ with respect to grammaticality and fluency. Therefore, we investigate the performance of state-of-the-art LMs for predicting these two aspects as rated by humans. Since gradedness of the acceptability of $Form$ is difficult to interpret, and we aim at producing a ratio score for $Form$ that we can feed into $\mathcal{MF}_\beta$ , we use a binary variable that reflects a threshold up to which a sentence is considered to be of acceptable form, or not. The $Form$ performance of the system then can be well interpreted as the ratio of sentences that it produced, which are judged to be of acceptable form.[5]

**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:

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

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 this score both for the generated sentence $mtp(s')$ and for the source sentence as reference $mtp(s)$, calculating a score of preference $prefScore = \frac{mtp(s')}{mtp(s')+mtp(s)}$. The decision on whether the 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 as being at least as good as the reference minus a tolerance, which we fix at 0.05. Finally, the corpus-level score for $Form$ reflects the ratio of sentences a system has produced, that are of acceptable form. This is inspired by Lau et al. (2020) except that the creation of the binary variable enables us to have obtain a corpus-level score for $Form$ that is interpretable by expressing a ratio in the range of [0,1], which is necessary to ensure sound $\mathcal{MF}_\beta$ calculation.[6]

**Form predictor selection** Similar to Lau et al. (2020), we consider GPT2 (Radford et al., 2019), distil GPT2 (Sanh et al., 2019) as well as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as a basis for assessing $Form$. We conduct experiments on the data by Gardent et al. (2017); Shimorina et al. (2017), which contains human fluency and grammaticality judgements for machine-generated sentences. Based on the results, we select GPT-2 as our basis for $Form$ assessment, since we find that it exhibits a good F1 score in binary prediction of fluency and grammaticality, and it shows slightly better performance compared with the other LMs. More details on this experiment can be found in Appendix 7.1.

### 3.4 Goals of our pilot studies

Our proposed $\mathcal{MF}_\beta$ metric for AMR-to-Text generation is aimed at offering a more balanced and justified assessment of generated sentences according to $Meaning$ and $Form$ than currently offered by standard surface-matching metrics. However, as detailed in §3.2 and §3.3, they depend on a number of hyper-parameters, such as the parser applied for $Meaning$ reconstruction or the used LMs for the assessment of $Form$.

To provide more insight into the properties of different modulations of the proposed decompositional $\mathcal{MF}_\beta$ metric and its possible dependence on the introduced parameters, we will conduct a series of pilot studies to better assess the potential benefits and weaknesses of $\mathcal{MF}_\beta$ when used to evaluate and rank AMR-to-text systems.

Specifically, we want to investigate i) to what extent $\mathcal{MF}_\beta$ aligns with other metrics in system scoring; ii) whether $\mathcal{MF}_\beta$ has the potential to explain its scores better than other metrics; iii) whether possible divergences in the assessment of system outputs are justified and in line with our principles for assessing $Meaning$ and $Form$.

Since any dependence on parameters that are subject to changes over time (such as LM capacity or AMR parsing performance) may be not desirable, an important task is to assess the effects of these factors on metric scores and system rankings. To investigate these questions, we conduct **two pilot studies**.

In **the first pilot study**, we want to assess the relation of $\mathcal{MF}_\beta$ to the conventionally applied string

[5] This aligns with insights drawn from judging well-formedness of automatically generated or translated sentences using binary preference-based rating (Zopf, 2018; Mathur et al., 2020).

[6] This means that the $Form$ score for a single sentence with $accept > 0.5$ equals 1.0. However, when such an assessment of a single sentence would be required, we may fall back on the $prefScore$ (+/- tol) as a realistic assessment of form.

matching metrics when ranking state-of-the-art systems, and its potential advantages. For instance, we are interested whether $\mathcal{MF}_\beta$ can justify potential differences in rankings and if it succeeds in disentangling $Form$ and $Meaning$.

In **the second pilot study**, we investigate a potential Achilles' heel of $\mathcal{MF}_\beta$ , namely its dependence on a parser and a LM. Therefore, we (i) investigate the effects of using another parser and (ii) we assess a potential remedy for this problem by using parse quality control. Finally, (iii), we validate the binary predictions by $Form$ in a small annotation study conducted by a native speaker.

## 4 Pilot study I: Assessing potential advantages of $\mathcal{MF}_\beta$

### 4.1 Setup

**Data and canoncial metrics** We retrieve the test predictions of several state-of-the-art AMR-to-text generation systems on LDC2017T10, which has served as the main testing grounds over the recent years: (i) densely connected graph convolutional networks (Guo et al., 2019); (ii) the system of Ribeiro et al. (2019) 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) by introducing a multi-task loss; finally, we obtain predictions of two system variants of Manuel et al. (2020) that fine-tune LMs and encode linearized graphs using (v) a large and (vi) a medium-sized model. 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 align with the sentence-matching setup that was previously used for evaluation of AMR-to-text. Along with BLEU, we display Meteor and chrf++ scores, since these three metrics are the most commonly used ones. Additionally, we calculate the recently proposed BERTSCORE (Zhang* et al., 2020) based on RoBERTa-large (Liu et al., 2019). The results are displayed in Table 1, col. 3-6. $\mathcal{MF}_\beta$ scores (col. 7-12) are divided into the core $Meaning$ (RESMATCH using GSII) and $Form$ scores, and the combined $\mathcal{MF}_\beta$ scores with $\beta = 1$ (harmonic) vs. $\beta = 0.5$, giving higher weight to $Meaning$.

**RESMATCH upper-bound approximation** As an upper-bound approximation for RESMATCH we propose parsing a gold sentence $s$ and comparing the result $m_s$ against the gold AMR $m_{gold}$: $apprUB$ = metric(parse($s$),$m_{gold}$). Essentially, this is the same score as used in canonical parser evaluation. This means that we would not expect the reconstruction parse $m'$ of $s'$ to score higher than had we applied $parse$ to the original sentence: $metric(parse(s'), m') \leq metric(parse(s), m_{gold}) = apprUB$, where $s'$, $s$ the generated and original sentence, $parse(s')$ the reconstructed AMR $m'$, $m_{gold}$ the original AMR.[7]

### 4.2 Enhanced interpretability of system rankings with $\mathcal{MF}_\beta$

**Surface matching metrics are not very discriminative and lack interpretability** Table 1 shows that the baseline metrics tend to agree with each other on the ranking of systems, but there also exist 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 properties of metrics as such, e.g., Meteor allowing inexact word matching of synonyms, in general, the underlying factors are difficult to assess, since the score differences between the systems with switched ranks are rather small, and none of these metrics can hardly provide us with meaningful interpretation for their score that would extend beyond shallow surface statistics. Therefore, these metrics cannot give us much intuition about why and when one system may be preferable over the other.

**$\mathcal{MF}_\beta$ yields more discriminative rankings** We assess the $\mathcal{MF}_\beta$ score with harmonic mean ($\beta =$ 1; Table 1, col. 11) and emphasis on $Meaning$ ($\beta = 0.5$; Table 1, col. 12). We see that, while the overall rankings stay similar, the $\Delta$s between system scores tend to grow. E.g., BERTSCORE assigns only 1.3 and BLEU 6.0 points difference between their selected best and worst systems, while $\mathcal{MF}_{\beta=0.5}$ assigns 8.1 points and $\mathcal{MF}_{\beta=0.5}$ more than 15 points in difference.

**$\mathcal{MF}_1$ and $\mathcal{MF}_{0.5}$ align well with BERTSCORE** Table 2, which shows the correlation of metrics, indicates that $\mathcal{MF}_\beta$ score is quite similar to BERTSCORE with respect to assigned rankings

[7] 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 $s'$ of $s$. Since we are interested in a very rough upper bound estimation, we abstract from such cases in our present work.

|  |  |  |  |  |  | $Meaning$ |  |  | $Form$ | $\mathcal{MF}_1$ | $\mathcal{MF}_{0.5}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  | abbrev. | BLEU | Meteor | chrF++ | BERTsc. | RESMATCH |  |  | - | - | - |
|  |  |  |  |  | F1 | P | R | F1 | Eq. 3.3 | Eq. 1 | Eq. 1 |
| $apprUB$ | - | - | - | - | - | 83.1 | 80.1 | 81.5 | 100 | 84.6 | 89.8 |
| Ribeiro et al. (2019) | R'19 | $27.9_{(5)}$ | 33.2 | - | $92.7_{(4)}$ | 76.5 | 67.7 | $71.9_{(6)}$ | $51.6_{(5)}$ | $60.1_{(5)}$ | $66.6_{(5)}$ |
| Guo et al. (2019) | G'19 | $27.6_{(6)}$ | - | 57.3 | $92.4_{(7)}$ | 78.2 | 70.0 | $73.9_{(3)}$ | $47.1_{(7)}$ | $57.5_{(7)}$ | $66.3_{(6)}$ |
| Wang et al. (2020a) | Wb'20 | $27.3_{(7)}$ | - | - | $92.6_{(6)}$ | 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 | 59.4 | $92.7_{(4)}$ | 78.1 | 69.2 | $73.4_{(5)}$ | $51.9_{(4)}$ | $60.3_{(4)}$ | $67.0_{(4)}$ |
| Manuel et al. (2020)-M | Mb'20 | $33.0_{(2)}$ | 37.3 | 63.1 | $93.9_{(2)}$ | 79.4 | 68.7 | $73.7_{(4)}$ | $\mathbf{74.0}_{(1)}$ | $\mathbf{73.9}_{(1)}$ | $\mathbf{73.8}_{(1)}$ |
| Manuel et al. (2020)-L | M'20 | $33.0_{(2)}$ | **37.7** | 63.9 | $\mathbf{94.0}_{(1)}$ | **80.8** | 69.2 | $74.5_{(2)}$ | $69.8_{(2)}$ | $72.1_{(2)}$ | $73.5_{(2)}$ |
| Wang et al. (2020b) | W'20 | $\mathbf{33.9}_{(1)}$ | 37.1 | **65.8** | $93.7_{(3)}$ | 80.3 | **70.9** | $\mathbf{75.3}_{(1)}$ | $55.7_{(3)}$ | $64.0_{(3)}$ | $70.3_{(3)}$ |

Table 1: Main metric results.

|  | BLEU | BERTsc | Resm. F1 | Form | $\mathcal{MF}_1$ | $\mathcal{MF}_{0.5}$ |
|---|---|---|---|---|---|---|
| BLEU | 100 | 94.6 / 83.6 | 79.3 / 75.7 | 77.2 / 84.6 | 81.8 / 84.7 | 89.1 / 88.3 |
| BERTsc |  | 100 | 63.4 / 48.7 | 90.0 / 95.5 | 93.2 / **95.5** | **96.7** / 91.9 |
| Resm. F1 |  |  | 100 | 39.9 / 42.9 | 46.4 / 42.9 | 61.3 / 57.1 |
| Form |  |  |  | 100 | 99.6 / 1.0 | 96.4 / 96.4 |
| $\mathcal{MF}_1$ |  |  |  |  | 100 | 98.2 / 96.4 |
| $\mathcal{MF}_{0.5}$ |  |  |  |  |  | 100 |

Table 2: Correlation matrix presenting Pearson's / Spearman's $\rho$ x100 of system scores over metric pairs.

(96.7 Pearson's $\rho$ with $\beta = 0.5$ and 93.2 Pearson's $\rho$ with $\beta = 1$). Interestingly, it appears that this is mostly due to $Form$, which exhibits, in contrast to RESMATCH, a very good agreement with BERTSCORE ($Form$: 90 Pearson's $\rho$, RESMATCH: 63.4 Pearson's $\rho$). However, $Form$ differs from the other metrics in the aspect that it assigns greater $\Delta$s among some systems, which indicates that some systems are capable to produce sentences of significantly improved form.

At this point, it is also important to recall that $Form$, in contrast to the other metrics, does not match two inputs, instead it bases its decisions solely on the generated sentences without matching their tokens against a reference. Thus, the high agreement with BERTSCORE could support the view that BERTSCORE may be more form-orientated than perhaps one would assume (Mehri and Eskenazi, 2020). However, that does not mean that BERTSCORE ignores the meaning, a conclusion that is supported by an even better correlations to $\mathcal{MF}_\beta$ , i.e., when we factor in some $Meaning$ into our $\mathcal{MF}_\beta$ score.

**(Decomposing) $\mathcal{MF}_\beta$ can provide explanations for system strengths** Before, we have seen that BERTSCORE incorporates both aspects, $Meaning$ and $Form$ without separating them. However, because it intermingles these two aspects in a way that is hardly transparent, it cannot provide us with an insight into whether the systems have different strengths with respect to $Form$ and $Meaning$. Here, it would be important that $Form$ and $Meaning$ are disentangled, as much as possible, so that they can provide complementary views on our problem that could explain different system rankings. That our metric indeed captures such complementary views is supported by the correlation statistic in Table 2, where we see that RESMATCH indeed appears to measure some different properties than the other metrics, since it exhibits the lowest average agreement compared with respect to all other metrics. Therefore, we may conclude that the weak correlation of $Meaning$ and $Form$ points towards an achievement of a key goal of this work: the disentanglement of $Form$ and $Meaning$, and that different systems have a tendency to be better in one aspect than the other (W'20 slightly favors $Meaning$, achieving first place in this aspect, while $M'20$ favors $Form$, Table 1); in Section 5.2 we will see that the latter (M'20) indeed appears to produce sentences of considerably better form).

**Using RESMATCH based on Damonte et al. (2017) leads to interpretable rankings** RESMATCH, when parameterized with fine-grained AMR metrics by Damonte et al. (2017), gives us deeper insight into the performance differences of competitive systems, with respect to specific semantic aspects.

The results are shown in Table 3. For example, when researchers aim at high quality for generation of named entities, they might better rely on the system ranked last in the overall ranking (R'19), which improves upon the best overall system by 3.4 points in NER recall and 1.9 points in F1 NER.

Furthermore, we see that the third best system according to all main metrics (Mb'20), may be less suited for correct negation generation. In this aspect, it lags behind the overall fourth best system

|  | 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 3: Fine-grained results using $\mathcal{MF}_0$ parameterized with the metrics proposed by Damonte et al. (2017).

C'20 by -5.8 points negation F1. We provide a full example, where RESMATCH explains a meaning negation error, in Figure 4 in the Appendix 7.2.

In sum, the system of W'20 appears to be the 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. However, this systems suffers in $Form$ performance, ranging much lower compared to the M'20 systems, which is why it is ranked only third place when using $\mathcal{MF}_\beta$ (and BERTSCORE), c.f. Table 1. In Section 5.2, we will conduct a native speaker study to assess whether this lack in $Form$ performance is really as great as indicated by our $Form$ score. Nevertheless, researchers who want to focus completely on the $Meaning$ may set $\beta = 0$ which discounts the form factor completely. Our evaluation shows that these researchers then may want to prefer the W'20 system for generation.

Finally, we see that the fine-grained metrics of Damonte et al. (2017) enhance our $Meaning$ component with the capacity to provide interpretation for system ranks. Additionally, in the Appendix of this paper, we provide very detailed examples of AMR reconstructions that lead to *different rankings of single candidate sentences*: in one case, RESMATCH explains SRL confusion (Appendix 7.3), in another aspect confusion (Appendix 7.4).

**The gap to the $apprUB$ indicates ample room for improvement of AMR-to-text systems** All metrics, including the surface matching metrics, e.g., BLEU or BERTSCORE, have a *mathematical upperbound*, which is 100 points. However, this upper-bound is not well interpretable since we cannot expect a system to score 100 points and estimation of true upper-bounds is extremely costly. RESMATCH, however, has an *interpretable upperbound (approximation)*: $apprUB$. It shows researchers that there is room for improvement of AMR-to-text generation systems (the gap to the best system according to RESMATCH (W'20), is more than 6 points in F1 and almost 10 points in recall).

Form, being disentangled from the distant source sentence, also shows that for most systems there is much room for the generation of wellformed and fluent sentences.

## 5 Pilot study II: Assessing vulnerabilities of $\mathcal{MF}_\beta$

$\mathcal{MF}_\beta$ has two apparent vulnerabilities: first, it depends on a parser for reconstruction. Here, we have used the state-of-the-art parser that is on par with human IAA. However, we cannot exclude the possibility that it introduces unwanted errors in the evaluation scores of $\mathcal{MF}_\beta$ .

Second, the $Form$ component is based on a LM and we have seen that it can change system rankings, even when it is discounted (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 one hand our LM was carefully selected, and other metrics (e.g., BERTSCORE) also heavily depend on LMs. Yet, on the other hand, we cannot exclude the possibility that the changed rankings are unjustified.

In this pilot study, we investigate these weak spots more closely by first assessing the outcome of $\mathcal{MF}_\beta$ when using another parser and discussing a mitigation of parser errors using a parse-quality control mechanism. Then we discuss the result of a human annotation study to assess whether the provided rankings by $Form$ were really justified.

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

**Using another parser** In this experiment we assess RESMATCH's robustness against using a different parser. This is an important point, since the metric and rankings could change with the parser and/or users may have reasons to use different parsers for the reconstruction. Here, we would hope, that the difference of using one competitive parser over the other will not be too extreme. To investigate this issue, we use GPLA (Lyu and Titov, 2018), a neural graph-prediction system that jointly predicts latent alignments, concepts and relations. We select GPLA because it constitutes a technically quite distinct approach compared to GSII.

The results are shown in Table 4, in the columns

|  | ReSmatch F1 |  |  | ranks ReSmatch |  |  | ranks $\mathcal{MF}_{0.5}$ |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | GPLA | GSII | GSII$^{\blacklozenge}$ | GPLA | GSII | GSII$^{\blacklozenge}$ | GPLA | GSII | GSII$^{\blacklozenge}$ |
| *apprUB* | 76.2 | 81.5 | 86.4 | 0 | 0 | 0 | 0 | 0 | 0 |
| R'19 | 70.1 | 71.9 | 80.1 | 7 | 6 | 6/7 | 5 | 4 | 5 |
| G'19 | 72.2 | 73.9 | 81.7 | 3 | 3 | 3 | 6 | 6 | 6 |
| Wb'20 | 70.2 | 71.5 | 80.1 | 6 | 7 | 6/7 | 7 | 7 | 7 |
| C'20 | 70.4 | 72.2 | 80.5 | 5 | 5 | 5 | 4 | 5 | 4 |
| Mb'20 | 70.5 | 73.7 | 82.1 | 4 | 4 | 1/2 | 2 | 1 | 1 |
| M'20 | 72.5 | 74.5 | 81.5 | 2 | 2 | 4 | 1 | 2 | 2 |
| W'20 | 73.1 | 75.3 | 82.1 | 1 | 1 | 1/2 | 3 | 3 | 3 |

Table 4: ReSmatch using different parsers (GPLA and GSII) or using a parser and high-quality filtering (GSII$^{\blacklozenge}$).

labeled with GPLA and GSII, without a ◆. We see that ReSmatch $^{GPLA}$ and ReSmatch $^{GSII}$ tend to agree in the majority of the rating (F1: Spearman's $\rho$ = 0.95, Pearson's $\rho$ = 0.96, p<0.001). When considering $\mathcal{MF}_\beta$ with $\beta = 0.5$, the vulnerability further decreases (Spearman's $\rho$ = 0.96, Pearson's $\rho$ = 0.99, p<0.001). Thus, we may conclude that ReSmatch exhibits some vulnerability towards using any of these two quite different parsers, but the extent of this vulnerability does not appear critical.

While we see that using GPLA has little effect on the ranks, we see that the nominal scores can differ substantially (e.g., W'20 73.1 F1 using GPLA and 75.3 F1 using GSII). However, we see that the increments are almost uniform. Therefore, we conjecture that there does not exist a system which got unfairly treated by parameterizing our metric with another parser. An unfair treatment could have arisen, e.g., if a parser unjustifiably generates overtly bad AMR reconstructions to specific systems. In such a case, the score increments would not be uniform. Hence, these increments are very likely to stem from the fact that we simply used a better parser, which is more benevolent to all generation systems.

**More quality control: parse quality assessment** An assessment for the reconstruction quality of single parses would allow researchers to get confidences for the provided scores by $\mathcal{MF}_\beta$ or one could conduct the evaluation only on a subset of generations where we are ensured that the quality of the parse reconstruction lies above a certain level. To assess the potential of such a solution, we use a parse quality estimation system (Opitz and Frank, 2019; Opitz, 2020). We then filter all tuples of generated sentences where the estimated quality of the parse lies above 95% F1 score. This leaves us with 169 tuples, on which we run the evaluation.

The results are given in Table 4, in the columns labeled with a ◆. With high-quality parses ensured, the ReSmatch ranking of systems changes slightly (GSII vs. GSII$^{\blacklozenge}$: Pearson' $\rho = 0.92$, Spearman's $\rho = 80.0$), as well as the ranking of $\mathcal{MF}_\beta$ (GSII vs. GSII$^{\blacklozenge}$: Pearson' $\rho = 0.95$, Spearman's $\rho = 0.96$). However – even though the evaluation data were changed by the filtering step – the tendency of $\mathcal{MF}_\beta$ in discriminating better systems from worse systems stays stable: over all settings, the two groups containing the highest-scored three systems and the lowest-scored four systems do not change.

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

In Section 4.2, we have seen that the $Form$ component of $\mathcal{MF}_\beta$ can impact the system rankings. We also saw that it tends to be in large agreement with BERTscore (not in the absolute scores, but in the rankings). However, BERTscore is mostly used in MT and therefore we would like to assess if the scores provided by $Form$ are really justified when evaluating AMR-to-text.

**Human annotation** To investigate this, we ask a native speaker of English to annotate 50 paired sentences of M'20 and W'20 with respect to their structural well-formedness, considering only grammaticality and fluency. The annotator was explicitly asked to not consider whether a sentence 'makes sense', by presenting the *Green ideas sleep furiously* example as free from structural error. We give more details on this annotations and provide examples in 7.5. The annotator agreed in 42 of 50 pairs with the preference as predicted by GPT-2, which is a significant result (binomial test p<0.000001). Additionally, we manually examine several produced sentences. We find that the M'20 and Mb'20 generations indeed appear considerably better on the surface level, compared to the generations of all other systems. For instance, the best system on the meaning level, W'20, frequently produces inflection mishaps: *Their hopes for entering the heat is already in-sight*, while we find little of such violations with M'20 (here: *Their hopes for entering the heat are already in sight*). We also find adverbial errors to varying degrees, 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 of these sentences are not of perfect form (correct: *mostly*), but the second sentence 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 5: $Form$ scores of systems when using a different LM.

**Using a different LM** The human study indicates that GPT-2 was mostly right when it favors one sentence over the other, with respect to fluency and grammaticality. However, when considering that there is a recent trend to build systems that are based on fine-tuning LMs, we need to assess whether they may be favored (too) much if $Form$ is parameterized with a same or a highly similar LM compared to the LM these systems use for tuning. We find such a case in M’20: on one hand, they did not fine tune the same GPT-2 which we used for $Form$ prediction, but they fine-tuned its siblings GPT-2-medium and GPT-large, which may share great structural similarities. Therefore, we also use BERT for our $Form$ prediction. The results (Table 5) support the unambiguous 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 best three systems. Note that this insight can be provided by $\mathcal{MF}_\infty$, but it cannot be carved out by using the conventional metrics, since they prohibit us from disentangling $Form$ and $Meaning$.

## 6 Conclusion

We proposed $\mathcal{MF}_\beta$ -score, a linguistically motivated metric for evaluation of text generation from (abstract) meaning representation. The metric is built on two pillars: $Form$, which measures grammaticality and fluency of the produced sentences and $Meaning$, which assesses how much meaning of the input AMR is reflected in the produced sentence. We saw that $\mathcal{MF}_\beta$ allows for a fine-grained system performance assessment that goes beyond what surface matching metrics can provide. Specifically, the $\beta$-parameter allows researchers to decompose the metric in either of the two parts, paving the way for custom gauging and selection of text generation systems. We observed that $\mathcal{MF}_\beta$ score could potentially be interpreted as BERTSCORE but offers the possibility to factorize and focus on the meaning aspects disentangled from form properties, and bears the potential for score interpretability via fine-grained semantic system assessment. Conversely, and in sharp contrast to BERTSCORE, the $Form$ component of $\mathcal{MF}_\beta$ enables an assessment of grammaticiality and fluency that does not rely on a match of the generated sentences against their references, and thus offers an assessment independent of lexical alignment.

A critical hyper-parameter of our metric is the dependency on the parser used for meaning reconstruction. To alleviate this issue, we used the latest state-of-the-art parser in our experiments. Additionally, we investigated this dependency by trying out a different parser and controlling for parse-quality. Our studies show that the absolute scores tend to increment when a better parser or only high-quality parses are used, but the ranking of systems stays quite stable. In future work, we want to investigate more ways of reconstruction quality control, e.g., using ensemble parsing. Furthermore, while benchmarking of systems needs deeper exploration, we consider the usage of $\mathcal{MF}_\beta$ scores to obtain better diagnostics and explainability of generated texts another interesting use case.

## 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.

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.

Laura Van Brussel, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Deng Cai and Wai Lam. 2020a. Amr parsing via graph-sequence iterative inference. *arXiv preprint arXiv:2004.05572*.

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.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. 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.

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

Markus Dickinson and Marwa Ragheb. 2015. On grammaticality in the syntactic annotation of learner language. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 158–167, Denver, Colorado, USA. 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.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, Brussels, Belgium. 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.

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 discourse representation. *Formal methods in the study of language*, (135).

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. 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 2284–2293, Minneapolis, Minnesota. 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.

Yi-hsiu Lai, Hsiu-hua Pai, et al. 2009. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, 22(5):465–475.

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.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2020. Dscorer: A fast evaluation metric for discourse representation structure parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4554, Online. Association for Computational Linguistics.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2017. Table-to-text generation by structure-aware seq2seq learning. *arXiv preprint arXiv:1711.09724*.

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.

Mager Manuel, Fernandez Astudillo Ramón, Naseem Tahira, Sultan Md Arafat, Lee Young-Suk, Florian Radu, and Roukos Salim. 2020. Gpt-too: A language-model-first approach for amr-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, USA. Association for 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.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Rik van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018. Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

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. *arXiv preprint arXiv:2005.12187*.

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.

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.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

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, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Roi Reichart and Ari Rappoport. 2010. Tense sense disambiguation: A new syntactic polysemy task. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 325–334, Cambridge, MA. 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.

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 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, 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.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association* *for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. 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*.

Markus Zopf. 2018. Estimating summary quality with pairwise preferences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1687–1696, New Orleans, Louisiana. Association for Computational Linguistics.

# 7 Appendices

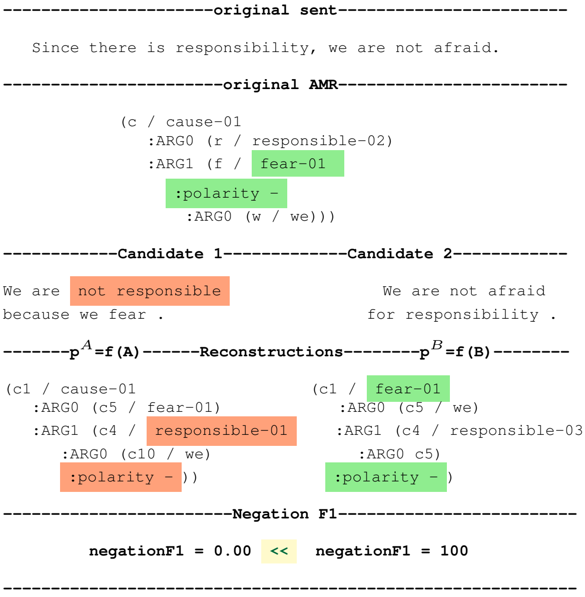
## 7.1 Form predictor selection experiment

To estimate how well they are able to assess $Form$, we make use of human-assigned scores for data 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[8] 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.[9]

The results are displayed in Table 6 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. Based on this, we select GPT2 for assessing $Form$, since it provides the best score on average, outperforming the other systems in grammaticality prediction.

| | 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 6: 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.



```
----------------------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 4: Explained negation confusion.

## 7.2 ResMatch explains negation error

In Figure 4, 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.

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

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

### 7.3 ReSmatch explains SRL error

Figure 5 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.

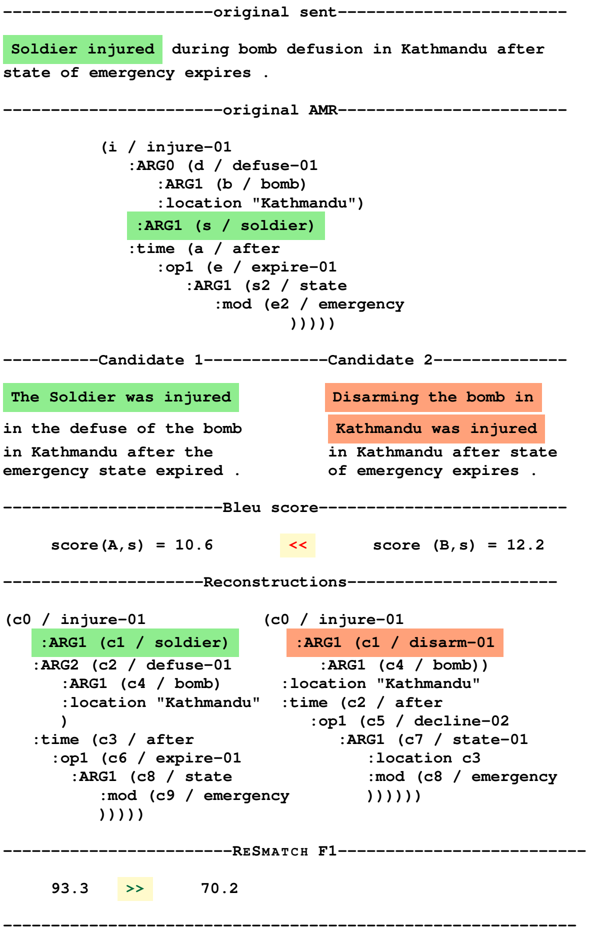
We see that the surface matching metric cannot explain its scores (beyond superficial statistics) and delivers a ranking that does not appropriately reflect the performance of the generation systems. However, ReSmatch shows that the gold parse and the parse of candidate A agree with each other in the 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.

### 7.4 ReSmatch explains aspect error

Here, we inspect a concrete example, where we see that ReSmatch can explain aspect confusion. Aspect is a complex phenomenon and an active area of NLP research (Reichart and Rappoport, 2010; Donatelli et al., 2018; Fan et al., 2018) and cognitive research (Tajiri et al., 2012; Lai et al., 2009).

In Figure 6, the gold sentence and AMR clearly capture that at the time *when the person heals*, the person feels forced to hurt themselves again. This aspect is well captured by candidate generation B: it states that at the time when they were healing, the person cut themselves, reflected in the AMR reconstruction as `<cut, :time, heal>`. Candidate generation A, on the other hand, misses this aspect, stating that at the time of the cut someone gets something, (reflected in the AMR reconstruction as `<cut, :time, get>`. Bleu, however, erroneously assigns a higher score to A (which misses this temporal aspect) than to B (which correctly captures the temporal aspect). On the other hand, ReSmatch is able to correct the wrong ranking and delivers an explanation, too.



```
----------------------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
in the defuse of the bomb         Kathmandu was injured
in Kathmandu after the            in Kathmandu after state
emergency state expired .         of emergency expires .

-----------------------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 5: Explainable re-ranking of single candidate sentences: SRL confusion.

### 7.5 Annotation study for form assessment

**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 label on a 11 point Likert scale where each number, starting from zero, indicates the amount of grammatical or fluency issues as assessed by the

```
----------------------original sent------------------------

      I am addicted,  when ever one heals  I cut again

-----------------------original AMR------------------------

           (m / multi-sentence
              :snt1 (a / addict-01
                 :ARG1 (i / i))
              :snt2 (c / cut-01
                 :ARG0 i
                 :mod (a2 / again)
                  :time (h / heal-01
                    :ARG1 (o / one))))

----------Candidate 1-------------Candidate 2--------------

I 'm an addiction , i cut         Addiction . again , i cut
again  when one gets one  .        when my one was healing  .

------------------------Bleu score------------------------

            14.4              >>              9.0

---------------------Reconstructions----------------------

(c0 / and                      (c0 / multi-sentence
   :op1 (c1 / addict-01           :snt1 (c1 / addict-01
      :ARG1 (c3 / i))                :mod (c3 / again))
   :op2 (c2 / cut-02              :snt2 (c2 / cut-01
      :ARG0 c3                       :ARG0 (c4 / i
      :mod (c4 / again)                 :part (c6 / one))
       :time (c5 / get-01             :time (c5 / heal-01
         :ARG0 c3                    :ARG1 c6)))
         :ARG1 (c6 / one)
         :quant 1)))

-------------------------ReSmatch F1-----------------------

           59.4               <<                86.7

-----------------------------------------------------------
```

Figure 6: Explainable re-ranking: Aspect confusion.

native speaker. Additionally, the human was asked to provide a correction.

**Examples of sentences of bad form.** See Figure 7.

```
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 .
----> our LM based prediction: 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 .
----> our LM based prediction: 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 .
----> our LM based prediction: acceptable

Sys (M’20): Fengzhu Xu won many championships in international competition before .
Corr (human): Fengzhu Xu won many championships in international competitions before .
----> our LM based prediction: acceptable
```

Figure 7: Sentences of with flawed form, i.e., containing grammatical or fluency errors, as assessed and corrected by the native speaker. ---> refers to the binary acceptability prediction that we used to determine the ratio of sentences that a system produces, which are of acceptable form.