A Dynamic, Interpreted *CheckList* for Meaning-oriented NLG Metric Evaluation – through the Lens of Semantic Similarity Rating

Laura Zeidler and **Juri Opitz** and **Anette Frank**
Department of Computational Linguistics
Heidelberg University, Germany
{zeidler|opitz|frank}@cl.uni-heidelberg.de

Abstract

Evaluating the quality of generated text is difficult, since traditional NLG evaluation metrics, focusing more on surface form than meaning, often fail to assign appropriate scores. This is especially problematic for AMR-to-text evaluation, given the abstract nature of AMR. Our work aims to support the development and improvement of NLG evaluation metrics that focus on *meaning*, by developing a *dynamic CheckList* for NLG metrics that is *interpreted* by being organized around meaning-relevant linguistic phenomena. Each test instance consists of a pair of sentences with their AMR graphs and a human-produced *textual semantic similarity* or *relatedness* score. Our *CheckList* facilitates comparative evaluation of metrics and reveals strengths and weaknesses of novel and traditional metrics. We demonstrate the usefulness of *CheckList* by designing a new metric GRACO that computes lexical cohesion graphs over AMR concepts. Our analysis suggests that GRACO presents an interesting NLG metric worth future investigation and that meaning-oriented NLG metrics can profit from graph-based metric components using AMR.

1 Introduction

Abstract Meaning Representation (AMR, Banarescu et al. (2013)) has become popular in NLP, one of the reasons being that AMR captures the essence of a sentence's meaning, while abstracting away from syntactic idiosyncrasies. Especially AMR-to-text generation (Konstas et al., 2017; Song et al., 2018; Wang et al., 2020; Blloshmi et al., 2021) has received much attention for applications that require text generation from structured content. However, the evaluation of text generated from AMR has been argued to be unsatisfactory (Manning et al., 2020). Also, Opitz and Frank (2021) show that the syntactic diversity of sentences generated from AMR is challenging for traditional NLG metrics, especially when candidates differ from the reference in surface properties.

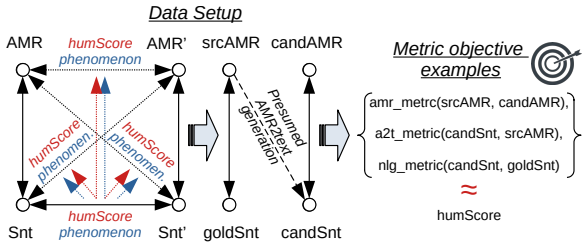

Figure 1: Our *CheckList* design for evaluating meaning-oriented NLG metrics against human semantic textual similarity and relatedness judgements – applicable to textual, meaning graph based and hybrid metrics.

Several metrics have been proposed that aim to rate the similarity of the meaning of sentences or phrases (Zhang et al. (2020); Opitz and Frank (2021); Zhao et al. (2019)). However, it is difficult to judge where exactly such a metric fails, making it hard for developers to further improve it. To address similar problems, Ribeiro et al. (2020) recently proposed a "task-agnostic methodology for testing NLP models" called *CheckList*. They argue that such a method should be used for testing NLP systems instead of solely relying on automatic metrics, which can overestimate a model's performance. Similar processes have been applied in early NLP research, e.g. with the TSNLP testsuite (Lehmann et al., 1996). Inspired by *CheckList*, in this work we aim to build a testsuite to enable systematic study and development of NLG evaluation metrics, with a focus on meaning.

Given the high variability of surface realizations that can be mapped into a single AMR graph, building reliable AMR-to-text NLG evaluation metrics is hard. Hence, it can be useful to construct a systematic *CheckList*, organized around diverse linguistic properties, to measure the performance of different metrics in an interpretable way. We frame our proposed CHECKLIST[1] and analyses derived

[1]The term *CheckList*, coined by Ribeiro et al. (2020), refers to their proposed methodology as well as concrete instantiations of such testsuites. We thus use the term *CheckList* (in

Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pages 157 - 172
July 14-15, 2022

from it in an AMR-to-Text NLG setting, and focus especially on a metric's capability to assess how well a specific meaning component of an AMR is reflected in its textual realization. We measure this using sentence pairs that differ in single linguistic aspects and measure how well various NLG metrics are able to rate such meaning differences. We compare the metric scores to human judgments from semantic textual similarity (STS) and relatedness datasets and analyze the metrics using our interpreted *CheckList* (an outline is shown in Fig. 1). Our contributions in this work are as follows:

i) We empirically identify properties relevant for rating the quality of generated sentences based on their meaning.
ii) We design an extensible, interpreted *CheckList* for evaluating NLG metrics, which offers 939 paired sentences with human judgements, covering 11 core linguistic phenomena.
iii) We propose a new metric GRACO to assess the semantic similarity of sentence pairs through the lens of AMR graphs.
iv) To showcase the potential of our approach, we provide an extensive comparative analysis of different types of NLG metrics, measuring their capacity of rating sentence similarity and relatedness according to linguistic differences.

2 Related Work

AMR-to-text evaluation Systems generating text from AMR graphs are typically evaluated using NLG metrics that were originally designed for other NLG tasks. BLEU (Papineni et al., 2002) or the CHRF(++) (Stanojević et al., 2015; Popović, 2015, 2016; Popov, 2017) metrics, e.g., are extensively used in MT. But May and Priyadarshi (2017) have shown that BLEU does not correspond well to human ratings of generations from AMR. Confirming this result, Manning et al. (2020) argue that existing automatic metrics fail to provide nuanced views on AMR-to-text generation quality. In an attempt to mitigate such issues, Opitz and Frank (2021) introduced a metric that combines meaning ($\mathcal{M}$) and form ($\mathcal{F}$) assessment in a weighted $\mathcal{MF}$ score, finding that system performances differ considerably in these two key quality aspects.

But to date, little is known about how different metrics measure meaning differences of generated sentences with regard to specific meaning alterations that may occur between a source and a reference. Our work provides a method and resources that can be used for performing such a detailed assessment for AMR-to-text generation metrics, and NLG evaluation metrics in general.

Checklist The current practice for evaluating NLP models is to assess their performance on unseen test data. Yet, summarizing performance in a single numerical score makes it difficult to assess where a model fails and how to fix remaining errors (Wu et al., 2019). Ribeiro et al. (2020) therefore proposed CHECKLIST, a methodology and tool for evaluating NLP systems based on the idea of *behavioural testing*, often used in software engineering. It aims at assessing specific capabilities of a system by testing whether inputs that feature specific properties will produce the expected output, without requiring knowledge of system's inner workings. This procedure is well-known in NLP, where before the rise of large-scale evaluation datasets, systems were tested and evaluated on so-called *testsuites* (Lehmann et al., 1996) that focused on specific *linguistic capabilities*. Ribeiro et al. (2020) adopted this approach to make their methodology applicable to many different NLP tasks. They evaluate multiple models on Sentiment Analysis, QA or Machine Reading Comprehension, showing that their method is beneficial in NLP: complementary to broad-scale evaluations, it can reveal specific points of failure, hence giving more detailed insight into a model's performance.

Semantic Textual Similarity (STS) Judging the similarity of texts is essential in tasks such as IR, text summarization or QA. But capturing semantic ambiguity, syntactic variance and paraphrasing is difficult. Hence, research started to investigate *Semantic* Textual Similarity (STS)[2], by tasking systems to judge the semantic similarity of sentences. Besides knowledge-based and distributional methods, neural methods have recently been proposed for STS estimation (Chandrasekaran and Mago, 2021). For example, S(entence)-BERT (Reimers and Gurevych, 2019) leverages pre-trained language models to predict STS scores, building on the insight of models that compute general sentence representations using paired sentence encoders (Conneau et al., 2017). These models outperform most traditional STS metrics, but lack interpretabil-

italics), to refer to our interpreted NLG testsuite.

[2]STS is a main component of SentEval and follow-up challenges, initiated by Conneau and Kiela (2018).

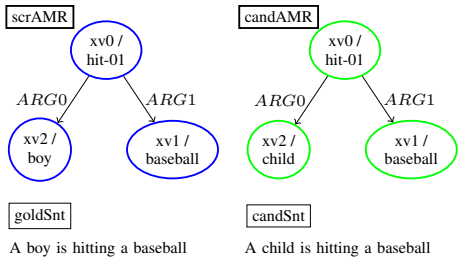

Figure 2: Example of a test case in our *CheckList* consisting of two sentence and AMR pairs. Drawn from the SICK dataset, with semantic relatedness score 4.4.

Phenomenon	Reference	AMR-to-text Generation
Antonymy	Flowers are so inconsistent !	flowers are so consistent .
Negation	My Drawing Number One .	not my picture number one .
Omission	the prince laughed , puzzled .	the prince laughed .
Passive	The wind blows them away .	they were blown away by wind .
Role Switch	The planet was inhabited by a conceited man .	the conceit man is inhabited by the planet .
more phenomena	hyponymy, co-hyponymy, partial synonymy, articles, subordinate clause types	

Table 1: (Modified) sentence pairs from AMR-to-text on the Little Prince AMR corpus.

ity. In our work we leverage STS and SentEval challenge datasets with human-rated semantic similarity (STS) and semantic relatedness (SICK) scores, to construct an interpreted *CheckList* that can be used to assess meaning-oriented NLG evaluation metrics, by evaluating them against human ratings.

3 An Interpreted Testsuite for Meaning-oriented NLG Evaluation Metrics

3.1 Aims and Method

The challenge of AMR-to-text NLG evaluation lies in the wide variability of sentences that can verbalize an abstract meaning representation. In our *CheckList*, we will consider human judgements of semantic textual similarity as a criterion for evaluating the adequacy of different NLG metrics for the AMR-to-text NLG evaluation task.

Specifically, we employ sentence pairs with human scores from the SICK and STS benchmarks[3] as test instances for our *CheckList* (cf. Fig. 2). We select pairs that *differ* by specific phenomena that can affect their semantic similarity, such as additional modifiers of a noun or verb, negation, or changes in the semantic roles of verb arguments. We parse such sentence pairs $S_{A,B}$ into pairs of AMR graphs $AMR_{A,B}$ that we manually validate.

Given such instances, we consider sentences S_A and S_B as a reference and candidate generation, and a pair of AMR and S as a sentence generated from an input AMR. For AMR_A we can take S_A as gold reference and S_B as a candidate generation; conversely, S_B can serve as a reference for AMR_B, and S_A as a candidate. We then interpret the human score for $S_{A,B}$ as a gold standard for a metric score that rates the appropriateness of S_B for AMR_A, given S_A as a reference, or S_A for AMR_B, given S_B as reference (see Fig. 1).

Following this rationale, our *CheckList* will offer curated input AMR graphs, their underlying sentences as references, and paired sentences from STS or SICK data points as candidate generations. The human scores serve as an objective to assess and compare various NLG evaluation metrics for their suitability in (A)MR-to-text evaluation tasks.

Aims Our *CheckList* is intended as a tool for researchers to build new or assess existing NLG metrics, regarding their ability to assess specific meaning aspects by comparing them to human judgements, thereby helping users to improve metrics, or better understand differences between metrics in meaning-oriented NLG evaluation in general and AMR-to-text generation in particular.

The suite is *interpreted* in two ways: by structuring the instances according to linguistic phenomena, and by pairing each sentence with its AMR graph, so that sentences can be compared at the textual and at the meaning representation level. Finally, the *CheckList* is conceived to be *dynamic*, by inviting developers to add new linguistic phenomena, test cases, and metrics.

Method To achieve this, we proceed as follows:

i) Empirical investigation We investigated sentences generated from the 'Little Prince Corpus'[4] using the AMR-to-text system of Song et al. (2018). We studied differences between the original and the generated sentences, to determine core phenomena that may influence the semantic similarity judgement of sentences generated from AMR towards their references. We distilled a list of phenomena shown in Table 1 that we further extended with phenomena observed in the STS and SICK datasets.

ii) Selection from STS and SICK Next, we select instances from the STS and Semantic Relatedness datasets (§5.1) that exhibit the phenomena identified in **i)**, and establish a suite of sentence

[3] https://github.com/facebookresearch/SentEval

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

pairs with their assigned human scores and respective AMRs. The data is structured into subsets exhibiting single phenomena, and is organized as an extensible *CheckList*.

iii) NLG metric scores & evaluation We implement scorers for various NLG metrics, and provide code to evaluate them via multiple measures to assess their strengths and weaknesses in view of phenomena captured in the CheckList. In addition, we propose a novel metric GRACO (§3.2) that constructs lexical cohesion graphs over tokens represented in the sentence's AMR, and compare it to existing metrics. The full range of functionalities to investigate NLG metrics is embedded into a CHECKLIST design (Ribeiro et al., 2020) (cf. A.1).

iv) Analysis and Interpretation We analyze the results and show how our *CheckList* enables systematic assessment of strenghts and weaknesses of NLG metrics when applied to outputs of AMR-to-text systems, taking into account the nature of different metrics in view of different phenomena.

3.2 Textual and AMR-based metrics

With our *CheckList* we aim at the evaluation of diverse metrics used in NLG and in semantic parsing, which we structure along two dimensions (cf. Table 2): metrics that evaluate candidate generations based on a) their textual (*tM*) vs. graph (*gM*) representations or both (hybrid, *hyM*), and b) whether the metric is based on symbolic as opposed to embedding representations. We don't include trained metrics, since their interpretation is difficult and would go beyond the current scope, but they can be evaluated on our *CheckList*, too. Table 6 provides an overview of characterizing traits of these metric types, which we will refer to in our analyses in §5.

Word/Char Ngram Matching Metrics Originally developed for MT evaluation, the BLEU (Papineni et al., 2002), Meteor (Lavie and Agarwal, 2007) and chrF++ (Popović, 2015) metrics have been increasingly used for evaluating NLG systems by comparing generated text to a reference on textual symbols. BLEU and Meteor compute overlap in word ngrams, while chrF++ extends the character ngram metric chrF by adding word ngrams.

Embedding-based Metrics BERTSCORE, proposed by Zhang et al. (2020), allows for reference-based evaluation using dense representations. Reference and candidate sentences are embedded with BERT to obtain contextualized representations for each token. A mapping between candidate and reference tokens is computed by greedy matching, based on cosine similarity of the encoding vectors. BERTSCORE shows a high correlation with human judgements for MT and Image Captioning tasks (Zhang et al., 2020). But while the metric is clearly meaning-based, it is focused on lexical meaning, and is not well equipped to capture word order and compositional meaning.

category	metric	gold information gldS	cndAMR	srcAMR
gM	S(^{2})match, W(W)LK	n	y	y
gM^{cndS}	S(^{2})match, W(W)LK	n	n	y
gM^{cndS}_{gldS}	S(^{2})match, W(W)LK	y	n	n
tM	BERTsc, Meteor, BLEU, chrF++	y	n	n
hyM	GRACO (this paper)	y	y	y

Table 2: Categorization of metrics into graph-based *gM*, text-based *tM* and hybrid *hyM* metrics, and their dependencies on gold information.

AMR Parse Evaluation Metrics While the previous metrics evaluate candidates against a reference at the *textual level* (tM), in our CheckList, we complement them by assessing similarity of meaning *structurally*, at the level of AMR graphs constructed from candidate and reference (gM).

We distinguish three potential setups: i) the metric is computed on manually rectified gold graphs (gM in Table 2); ii) an integrated parser component constructs an automatic candidate AMR *cndAMR* from the candidate sentence *cndSnt* to alleviate the requirement for a golden *cndAMR* (gM^{cndS} in Table 2); iii.) the parser constructs both *srcAMR* and *candAMR* from the reference and candidate sentence, i.e., we trade the dependency on a golden *srcAMR* against the dependency on a golden reference sentence (gM^{cndS}_{gldS} in Table 2). Variants ii) and iii) have also been used in the $\mathcal{M}$ ('Meaning') component of MF-score (Opitz and Frank, 2021). For simplicity, in this paper, we assume access to gold graphs and only consider gM, tM, and hyM metrics.

As AMR graph metrics, we use the canonical SMATCH (Cai and Knight, 2013), the recent S^{2}MATCH metric proposed by Opitz et al. (2020), and Weisfeiler-Leman based AMR graph similarity proposed by Opitz et al. (2021) that match contextualized AMR graphs.

SMATCH is a *binary* triple overlap metric that assesses the structural similarity of candidate and reference AMRs, where a triple is a pair of AMR nodes connected by a labeled edge. S^{2}MATCH, by

contrast, computes a *graded* triple overlap score using the embedding similarity between the concept nodes of a triple pair, to reflect concept similarity in the overall AMR similarity score. Given a reference AMR for *'a kitten meows'*, S^{2}MATCH will assign a relatively high score for a candidate AMR for *'a cat meows'* that reflects high lexical similarity of *kitten* and *cat* in the overall score, while SMATCH will assign it a much lower score.

The Weisfeiler-Leman AMR metric comes in two variants: W(eisfeiler)L(eman)K(ernel) (WLK) compares contextualized AMR graphs structurally, while W(asserstein)WLK (WWLK) compares the contextualized AMR graphs in latent space, using an alignment-based Wasserstein distance. WWLK extends S^{2}MATCH beyond the lexical level, to capture *compositional* meaning similarity at the phrasal level, as between *'a young cat meows'* vs. *'a kitten meows'*.

Hybrid Metrics The above metrics take as input sentence pairs or AMR pairs. But a meaning-oriented NLG metric may profit from considering both explicit meaning structure as captured in AMR, and the textual level, to leverage knowledge from pretrained language models trained on text. We thus propose a **hybrid similarity metric** GRACO, which is based on *Lexical Cohesion Graphs* proposed by Sporleder and Li (2009). They construct an undirected graph from a text sequence where each node represents a content word, and compute edge weights between the lexical nodes using Normalized Google Distance (Cilibrasi and Vitanyi, 2007). By averaging the weights they derive a *connectivity* score for the graph. In their work they use the lexical cohesion graph of a given token sequence to predict whether it has an *idiomatic* as opposed to a *literal* meaning, depending on whether the presence of its subgraph in the overall graph raises or lowers the overall connectivity score.

We adapt Sporleder and Li (2009)'s approach to define a *hybrid metric* that measures the similarity of sentence pairs via their AMR graphs. We do this by building a lexical cohesion graph from the concept nodes present in a sentence's AMR. To do so, we align words from the sentence with concepts in the AMR graph using the JAMR (Flanigan et al., 2014a) alignment tool. The concepts are either represented using contextualized BERT embeddings or pretrained GloVe word embeddings. To compute edge weights, we follow Haagsma et al. (2018) and compute cosine similarity between nodes. We pursue two strategies. i) We follow Sporleder and Li (2009) and compute cosine similarity between all possible pairs of nodes of a single graph, creating a *fully connected* graph. Alternatively, ii) we compute a reduced graph that only takes into account edges connecting nodes that *differ* between the two sentences and their respective graphs (see Fig. 3). In case graph g_A differs from graph g_B in a single concept which is only present in g_A, the reduced graph g_B is empty, and we assign a connectivity score of 1 (consistent with anything).

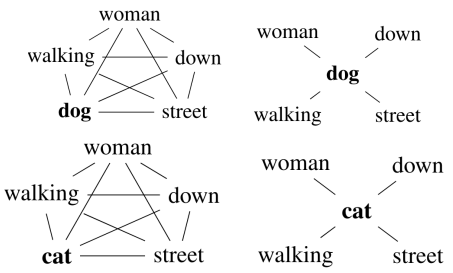

Figure 3: Two lexical cohesion graphs: fully connected (left) and reduced (right) for sentences S_A: *The woman is walking the* ***dog*** *down the street* – S_B: *The woman is walking the* ***cat*** *down the street.*

By applying this method to a pair of sentences S_A and S_B, we obtain their *connectivity scores* cs_A and cs_B, the average of their respective graphs' edge weights. From these we compute the GRACO Score (1) that rates the similarity of S_A and S_B by taking the difference between cs_A and cs_B to model their semantic difference – which we convert to a similarity score by subtracting it from 1.

$$\text{GRACO}Score = 1 - |cs_A - cs_B| \quad (1)$$

The resulting metric is hybrid by relying on the sentence's *AMR* to select text tokens for the connectivity graph – and represents nodes with *contextualized embeddings* in the BERT variant.

4 Semantic Phenomena

We consider structural and lexical phenomena that are likely to affect a sentence's meaning. Details and example AMRs are given in Appendix A.4.[5]

4.1 Structural Phenomena

Aspect Given its abstract nature, AMR does not represent aspect, hence present perfect and simple present are not distinguished in an AMR graph[6].

[5]AMR specifications follow Banarescu et al. (2019).

[6]This phenomenon was only found in the STS data.

Negation AMR represents negation with the feature `:polarity -`. Fig. 10 (A.4.1) shows sentence negation, with `polarity` attached to the matrix verb. Fig. 11 (A.4.1) shows an AMR that negates a constituent in a sentence. Both verb- and constituent negation are represented in the testsuite.

Omission or Hallucination of words or phrases is a recurring problem in NLG (Xiao and Wang, 2021) especially for AMR-to-text (Manning et al., 2020). We sampled three types involving *adjectives*, *adverbs*, *PPs*. In AMR, omission/hallucination is captured by (non-)existence of the corresponding structure (see Fig. 13, A.4.2).

Passive AMR does not distinguish active from passive voice: AMR graphs for active vs. passive sentences do not differ and do not reflect voice.

Semantic Role Switch describes cases where two verb arguments switch semantic roles. Fig. 15 (A.4.4) shows that the switch changes the `:ARG` roles of both arguments, involving two triples.

Subordinate Clauses In AMR, relative clauses can involve *inverse roles* if the relativizer is dependent on a verb. The AMR for *A boy who believes*, e.g., contains an inverse ARG0 role. Other types of relative clauses, *Noun Compound Expansions*, reveal a semantic relation between compound nouns. Such expansions can be expressed in various ways:

(1) a. *A man is playing a flute made of bamboo*
b. *A man is playing a bamboo flute*

(2) a. *A child is running in and out of the waves of the ocean*
b. *A child is running in and out of the ocean waves*

While the expansions in (1a, 2a) differ (*made of* vs. *of*), the two compound nouns in (1b) and (2b) are connected with same AMR relation `:part-of`, which reveals their semantic relation. The expansion in (1a), by contrast, emphasizes the process of the flute being *made*, which is reflected in its AMR (see Fig. 12, A.4.5). Hence, whenever we compare sentences that make use of a noun compound or an expansion of it, they may differ in their textual *and* their AMR representations, which can have implications for different types of metrics.

4.2 Lexical Phenomena

Articles AMR does not specify articles, so the sentence variants {*A*|*The*} *child is playing.* yield identical AMRs. I.e., it cannot distinguish sentences differing in definiteness of an article. Our *CheckList* includes pairs exhibiting such differences.

Antonymy denotes a relation of contrast that can apply to *adjectives*, *adverbs*, *nouns*, *prepositions* or *verbs*. In AMR, antonymy is either implicit for concept pairs or represented by negating a concept with `:polarity -` (Fig. 17 in A.4.7).

Note that human ratings in STS and SICK differ for antonymy and negation. While in STS, antonymy and negation are penalized with low similarity scores, this is different for SICK, which rates *semantic relatedness* of sentences. Pairs including a single opposing concept may yield higher scores than comparison to a random sentence. This must be observed when interpreting *CheckList* results.

Hypernymy and Hyponymy, and the derived **Co-Hyponymy** relation, while known from WordNet, are not explicitly expressed between AMR concepts. They form the basis for inferential relations between sentences and play an important role in judging NLG quality from a semantic view. Often, a candidate may differ from its reference sentence by resorting to a superordinate, less specific concept, but may combine it with a differentiating modifier, yielding an equivalent meaning. Equivalence of compositional meaning is difficult to capture for word-based and lexical NLG metrics, and is even more challenging for metrics based on structured meaning representations. **Co-Hyponymy**, however, involves contrast and interferes with **Antonymy** and **Negation**.

(Partial) Synonymy We distinguish *total* and *partial* synonymy. In the former, linguistic expressions are interchangeable without restriction, while in the latter this may hold in a context given their denotative meaning, may not hold when considering their connotative meaning (Edmonds and Hirst, 2002). Examples are *lie – untruth*, or *task – job*. While the former type is unproblematic for meaning-oriented, lexical NLG metrics, the latter is not, as it requires judging contextual conditions. Since AMR specifies abstract concepts, choosing contextually adequate synonyms is a challenge, and contextualized metrics may have an advantage.

5 Interpreted Evaluation of NLG Metrics

5.1 Datasets and Statistics

We sampled 939 sentence pairs, each differing in a single phenomenon from SICK (877) and STS (62)[7], parsed them into AMRs using the parser of Raffel et al. (2019) and manually corrected them.[8]

[7]Distributions of phenomena and human scores in A.3.2.
[8]Manual correction was performed by two of the authors.

STS (Semantic Textual Similarity). Since the first SemEval STS task (Agirre et al., 2012), a total of 15,459 sentence pairs were created in follow-up challenges. Each sentence pair is annotated for semantic similarity on a Likert scale from 5: "completely equivalent" to 0: "on different topics".

SICK: Sentences Involving Compositional Knowledge by Marelli et al. (2014) contains 10,000 English sentence pairs, annotated for *semantic relatedness* and *entailment*. Pairs were normalized, expanded using specific linguistic phenomena, and finally paired with one another. Due to this process, pairs often differ by single linguistic phenomena, making them well suited for our aims. The sentence pairs were rated for semantic relatedness on a five-point Likert scale, from 1: "completely unrelated" to 5: "very related".

Since the annotations on SICK and STS are not equivalent, they will be analyzed separately.

5.2 Experimental Setup

Metrics All metrics except GraCo use existing implementations. To enhance comparability between metrics, we standardize and normalize the scores of every metric and the annotated human scores (see A.3.3 for details on both).

Evaluation metrics for metric performance We compute **i) Correlations** of the metric scores with the human scores using *Spearman's rho*. **ii) Pairwise Ranking scores** for all metrics, where for each phenomenon we consider all possible combinations of pairs (x, y) and (x', y'). A metric m scores one point if the relation between the predicted scores $m(x, y)$ and $m(x', y')$ for the given pairs corresponds to the relation between their human scores $h(x, y)$ and $h(x', y')$. If for instance $h(x, y) < h(x', y')$, metric m earns one point if

$$m(x, y) < m(x', y') \quad \wedge$$
$$|m(x, y) - m(x', y')| > \tau$$

where τ is a threshold we define as the fifth percentile of all scores. We define $m(x, y) = m(x', y')$ if $|m(x, y) - m(x', y')| \leq \tau$. **iii)** Mean Average score and its **Mean Absolute Deviation** (MAD) from the human score over test cases.

5.3 Hypotheses

We state hypotheses on how various metrics are expected to perform for selected phenomena.[9]

H1: *gM* vs. *tM* AMR metrics are less sensitive to surface variation than textual metrics. This can be beneficial when variations have a mild impact on human judgements of similarity (*Passive*, *Articles*), but may have adverse effects when the impact is high. This may happen with *Antonymy*, if the metric cannot capture relevant differences in lexical meaning, as in SMATCH.

We expect BERTScore to compete with *gM* metrics, due to its contextualized representations. In general we expect all AMR metrics to have an advantage over textual metrics, except for BERTSCORE, in detecting *Switched Roles*, since they explicitly represent argument roles.

H2: Impact of small substrings or subgraphs Irrespective of differences in human judgement for *Antonymy, Co-hyponymy* and *Negation* between SICK vs. STS (cf. §4), metrics can differ in how strongly a contrast at token or concept level affects a pair's overall rating. In such cases only few triples may differ between sentence pairs, so we don't expect S(2)MATCH to reflect strong drops in human score. W(W)LK may fare better, as its kernel can capture a wider context of a given node. BERTScore faces similar problems when small text portions cause a strong contrast, but its contextualization may reflect the impact of neighboring words, an effect that could be shared with W(W)LK.

While all prior metrics compute scores over the entire sentences, GRACO^{red} only considers local subgraphs restricted to *differing* nodes. We expect this to be beneficial for phenomena like *Negation*.

H3: Capturing (dis)similarity We expect S^{2}MATCH and W(W)LK to perform closer to human judgement than SMATCH for sentences that differ by semantically similar or closely related words, e.g., with *Partial Synonymy* or *Hyponymy*. The same should hold true for Meteor as opposed to BLEU and chrF++, since it accounts for synonyms and paraphrases. W(W)LK is expected to capture compositional similarity (*young cat – kitten*) better than S^{2}MATCH, which is purely lexical. But S^{2}MATCH and WWLK could perform worse for *Antonymy*, since antonyms tend to be close to each other in latent space (Samenko et al., 2020).

5.4 Results and Analyses

Results are displayed in Tables 3 and 4 for SICK.[10] Fig. 4 displays an aggregated view of correlations between the metric scores and human scores for

[9] Due to space restrictions, we only discuss a selection, which we mark with ✓Hx vs. ✗Hx if (un)supported by results.

[10] STS results are seen in Tables 7, 8 and Fig. 5, in A.2.

	Antonymy	Article	Co-Hyp.	Hyponymy	Negation	Omission	Part. Syn.ymy	Passive	Sem. Roles	Sub. Clauses	Overall
Ann. Score	0.614	0.977	0.628	0.863	0.597	0.86	0.941	0.976	0.6	0.963	0.789
BLEU	0.672 ± *0.19*	0.772 ± 0.21	0.775 ± 0.22	0.72 ± 0.18	0.582 ± 0.2	0.645 ± 0.23	0.734 ± 0.22	0.108 ± 0.87	0.298 ± 0.3	0.579 ± 0.38	0.611 ± 0.28
chrF++	0.796 ± 0.2	0.865 ± 0.11	0.794 ± 0.2	0.779 ± 0.12	0.846 ± 0.25	0.728 ± 0.14	0.798 ± 0.15	0.339 ± 0.64	0.669 ± 0.12	0.733 ± 0.23	0.75 ± 0.22
Meteor	0.421 ± 0.24	0.605 ± 0.37	0.444 ± 0.22	0.669 ± 0.26	0.46 ± *0.16*	0.466 ± 0.39	0.808 ± 0.18	0.258 ± 0.72	0.415 ± 0.19	0.408 ± 0.56	0.482 ± 0.33
BERTSCORE	0.868 ± 0.26	0.953 ± 0.04	0.854 ± 0.24	0.86 ± *0.08*	0.749 ± 0.17	0.813 ± *0.08*	0.925 ± **0.04**	0.512 ± 0.46	0.726 ± 0.16	0.783 ± 0.18	0.805 ± 0.17
SMATCH	0.793 ± 0.22	0.998 ± **0.02**	0.833 ± 0.22	0.83 ± **0.07**	0.921 ± 0.32	0.844 ± **0.06**	0.829 ± 0.12	0.995 ± **0.03**	0.647 ± *0.11*	0.917 ± 0.09	0.877 ± **0.14**
S^2MATCH	0.793 ± 0.22	0.998 ± **0.02**	0.838 ± 0.23	0.831 ± **0.07**	0.921 ± 0.32	0.844 ± **0.06**	0.829 ± 0.12	0.995 ± **0.03**	0.647 ± *0.11*	0.917 ± 0.09	0.877 ± **0.14**
WLK	0.575 ± **0.16**	0.989 ± *0.03*	0.586 ± **0.16**	0.539 ± 0.32	0.791 ± 0.2	0.782 ± 0.1	0.614 ± 0.33	0.993 ± **0.03**	0.525 ± **0.1**	0.896 ± 0.11	0.745 ± *0.16*
WWLK	0.76 ± 0.21	0.996 ± *0.03*	0.736 ± *0.19*	0.721 ± 0.16	0.644 ± **0.15**	0.685 ± 0.18	0.734 ± 0.21	0.994 ± **0.03**	0.936 ± 0.34	0.907 ± 0.1	0.774 ± **0.14**
GRACO$_{gl}$	0.952 ± 0.36	1.0 ± **0.02**	0.97 ± 0.34	0.963 ± 0.11	0.974 ± 0.38	0.926 ± 0.13	0.975 ± *0.05*	0.936 ± 0.06	0.998 ± 0.4	0.992 ± **0.03**	0.961 ± 0.2
GRACO$^{red}_{gl}$	0.883 ± 0.35	1.0 ± **0.02**	0.942 ± 0.32	0.933 ± 0.09	0.381 ± 0.23	0.277 ± 0.59	0.951 ± *0.05*	0.93 ± 0.06	1.0 ± 0.4	0.853 ± 0.16	0.711 ± 0.26
GRACO	0.952 ± 0.34	0.969 ± 0.04	0.959 ± 0.33	0.949 ± 0.11	0.942 ± 0.35	0.935 ± 0.11	0.965 ± *0.05*	0.938 ± *0.05*	0.985 ± 0.38	0.946 ± *0.04*	0.948 ± 0.19
GRACO^{red}	0.875 ± 0.32	1.0 ± **0.02**	0.91 ± 0.29	0.915 ± 0.11	0.497 ± 0.24	0.447 ± 0.43	0.937 ± 0.06	0.92 ± 0.07	0.92 ± 0.39	0.865 ± 0.14	0.755 ± 0.23

Table 3: Avg. normalized score & mean abs. deviation (most indicative, lower is better) from human score for SICK.

	Ant.my	Art.	CoHyp	Hyp	Neg	Omiss	P.Syn	Pass	SRL	Sb.Cl	Ovll
BLEU	0.492	0.34	*0.54*	0.419	0.433	0.459	0.391	*0.335*	0.469	0.321	0.424
chrF++	0.5	0.342	0.523	0.437	0.441	*0.489*	0.435	0.303	*0.562*	0.336	0.367
Meteor	**0.538**	0.35	**0.564**	**0.494**	0.441	0.435	**0.524**	0.322	0.438	0.365	0.463
BERTSc	0.483	0.36	0.505	*0.469*	0.473	**0.523**	0.435	0.31	0.406	0.355	0.47
SMATCH	0.485	0.357	0.486	0.402	0.408	0.456	0.399	**0.349**	0.406	0.364	0.579
S^2MATCH	0.484	0.357	0.474	0.395	0.408	0.456	0.399	**0.349**	0.406	0.364	0.578
WLK	*0.516*	*0.375*	0.509	0.413	0.429	0.471	0.349	**0.349**	0.469	0.363	*0.628*
WWLK	0.485	0.357	0.456	0.439	0.449	0.47	0.396	**0.349**	0.469	0.357	**0.636**
GRACO$_{glo}$	0.489	**0.385**	0.469	0.436	0.458	0.415	0.296	0.302	0.219	0.368	0.511
GraCo$^{red}_{glo}$	0.437	0.367	0.509	0.406	*0.496*	0.405	0.402	0.305	0.188	0.378	0.553
GRACO	0.473	0.292	0.497	0.411	0.428	0.46	*0.485*	0.321	**0.625**	**0.46**	0.449
GraCo^{red}	0.433	0.367	0.481	0.416	**0.505**	0.418	0.444	0.327	0.219	*0.384*	0.565

Table 4: Pairwise ranking scores for the SICK test cases.

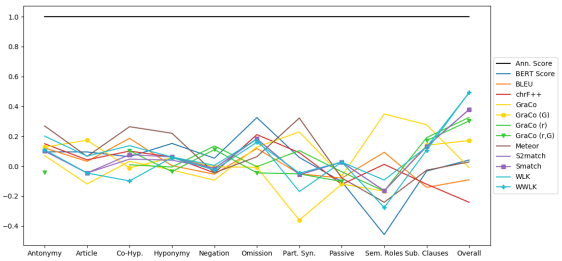

Figure 4: Spearman's rho correlation between metric and human scores for SICK. Broken lines indicate phenomena where no correlation coefficient could be computed due to identical metric scores for all instances.

individual phenomena. Finally, Table 5 presents a summary for all metrics and the phenomena they perform best or 2nd best on, according to our three evaluation metrics: ranking score, MAD and correlation to human judgement scores.

The *gM* metrics W(W)LK show best overall performance, sharing 1st place with S(^{2})MATCH in SICK and obtaining first place in pairwise ranking, and we see top places being achieved for 4-5 phenomena (✓ H1, ✓ H3). But S(^{2})MATCH produce very similar scores across the board (✗ H3).

Among symbolic *tM* metrics, *Meteor* performs best in ranking score, and *chrF++* for MAD. BERTSCORE performs better than symbolic *tM* metrics overall, except for ranking score for STS, where it only fails on *Aspect* (✓H1). But it falls behind *gM* and most *hyM* metrics in *overall* scores. GRACO performance varies across phenomena and its variants. It occupies 1st and 2nd places in ranking score for *Neg* in SICK in the *reduced* variant, where the drop in avg score and MAD is striking (✓H2). For other phenomena, the performance aligns with the other *gM* metrics. This suggests that the connectivity score captures most lexical phenomena well – while for *SRL* this is evidently not sufficient (✓H1).

Beyond tendencies in overall results, we now focus on observations for single phenomena.

While *gM* generally outperform *tM* metrics, this doesn't necessarily hold for Meteor: it outperforms *gM* for phenomena reflecting lexical-semantic relations for SICK (Table 4, Fig. 4). The spike in correlation for *Part. Syn.* is expected, as Meteor accounts for synonyms and paraphrases (✓H3). This may also explain its superior performance for *(Co-)Hyponymy*. But its high performance for *Antonymy* is surprising (✗H3).

S^2MATCH performing very similar to SMATCH is most likely due to a high threshold for allowing a soft match. GRACO was designed to better represent semantic contrast between sentences and their AMR graphs. We can see this reflected in a large drop of MAD for GRACO^{red} in *Negation*. In comparison, for *Antonymy* we only see a relatively small drop in MAD. This is because, for *Negation*, GRACO^{red} produces a bigger contrast between the connectivity scores as one of them is 1 for the empty graph. For *Antonymy* the scores are closer, since both graphs have neighbors. Another factor could be the proximity of antonyms in embedding space, which suggests that a threshold, similar to S^2MATCH, could be beneficial.

We also observe that GRACO using BERT outperforms GRACO$_{glo}$ in *Part.Syn, SRL, SubCl* (Table 4, Fig. 4). This is unexpected since neither of them uses AMR relations. This could be explained by the contextualized node embeddings that see context at textual level–combined with connectivity graphs

	Best & 2nd Best Ranking Scores	**Best** & 2nd Best MAD	**Highest** & 2nd Highest Correlation w/ Human
BLEU	Passive, Co-Hyp.	Antonymy	Co-Hyp., SRL
chrF++	Omission, SRL		Omission
Meteor	**Co-Hyp.**, **Antonymy**, **Part. Synonymy**, **Hyp.**	Negation	**Part. Synonymy**, **Antonymy**, **Co-Hyp.**, **Hyp.**
BERTSc	**Omission**, Hyp.	**Part. Synonymy**, Omission, Hyp.	**Omission**, Hyp.
SMATCH	**Passive**	**Article**, **Passive**, **Omission**, **Hyp.**, SRL	**Passive**
S^2MATCH	**Passive**	**Article**, **Passive**, **Omission**, **Hyp.**, SRL	**Passive**
WLK	**Passive**, Article, Antonymy	**Passive**, **SRL**, **Antonymy**, **Co-Hyp.**, Article	**Passive**, Antonymy
WWLK	**Passive**	**Passive**, **Negation**, Article, Co-Hyp.	**Passive**
GRACO$_{glo}$	**Article**	**Article**, **Sub. Clauses**, Part. Synonymy	**Article**
GraCo$^{red}_{glo}$	Negation	**Article**, Part. Synonymy	Negation
GRACO	**SRL**, **Sub. Clause**, Part. Synonymy	Sub. Clauses, Part. Synonymy, Passive	**SRL**, **Sub. Clauses**, Part. Synonymy
GraCo^{red}	**Negation**, Sub. Clause	**Article**	**Negation**, Sub. Clauses

Table 5: **Best** & 2^{nd} Best Metric Performances in Ranking Score, MAD, Corr. with Human Scores for SICK dataset.

		textual level					graph level				
Type	Metric	words	chars/pieces	lexicon	dense	contextual	concepts	sem. edges	sim. edges	dense	contextual
tM	BLEU	+	-	-	-	+					
	chrF++	+	+	-	-	+					
	Meteor	+	-	+	-	-					
	BERTScore	-	+	-	+	+					
gM	SMATCH						+	+	-	-	-
	S^2MATCH						+	+	-	+	-
	WLK						+	+	-	-	+
	WWLK						+	+	-	+	+
hyM	GRACO$_{glo}$	+	-	-	-	-	+	-	+	+	-
	GRACO$^{red}_{glo}$	+	-	-	-	-	+	-	+	+	-
	GRACO	+	-	-	-	-	+	-	+	+	+
	GRACO^{red}	+	-	-	-	-	+	-	+	+	+

Table 6: Characterization of the used textual (*tM*), graph-based (*gM*) and hybrid (*hyM*) metrics in terms of textual and graph-level properties. **textual level**: word/char/lexicon-based; **graph-level**: semantic vs. similarity edges; **both levels**: dense = embedding-based representation; contextual = contextualized representation.

that look at the sentence only via AMR nodes.

Overall we see surprising effects with GRACO: i) by restricting connectivity to local subgraphs for contrasting elements, it yields strong performance for *Negation*; ii) it only focuses on AMR nodes, but the contrast with GRACO$_{glo}$ suggests that the contextualization helps to assess surface differences underlying *SRL* and *SubCl*. The insights from GRACO could trigger ideas for improving a *tM* metric like BERTSCORE, by computing it under a similar AMR lens, and handling *Negation* in similar ways. It also suggests studying the use of BERT embeddings in WWLK, and seeking ways of integrating a comparable mechanism for *Negation*. As for *tM* metrics, it came as a surprise to find Meteor keep 1st rank for lexical relations ((Co-)Hyp; (Partial)Syn, Antonymy), beyond BERTSCORE.

6 Conclusion

We introduced an extensible *CheckList* for meaning-oriented NLG metrics that allows for comparison of a wide range of NLG metrics. It is interpreted by way of offering test cases grouped by linguistic phenomena. Our analyses showcase how *CheckList* can be used to compare metrics, to reveal their strengths and weaknesses. They align with a number of hypotheses, but also show surprising effects, opening avenues to further improve NLG evaluation metrics. We propose a novel, hybrid similarity metric GRACO that builds cohesion graphs over contextualized AMR concept nodes. The metric can focus on contrastive subgraphs, which yields strong correlation with human judgements for negation. With regard to current practice in AMR-to-text evaluation, we find evidence that meaning-oriented graph-based metrics present advantages over typical text-based metrics, confirming the findings of Opitz and Frank (2021); Manning et al. (2020). Therefore we recommend to include graph metrics or hybrid graph- and textual metrics into AMR-to-text evaluation protocols. Our data and code will be publicly available.[11] We welcome contributions to grow it.

Acknowledgements

We thank the anonymous reviewers for useful feedback and suggestions.

[11] https://github.com/Heidelberg-NLP/NLG-CHECKLIST

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

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

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2019. Abstract meaning representation (amr) 1.2.6 specification.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Rexhina Blloshmi, Michele Bevilacqua, Edoardo Fabiano, Valentina Caruso, and Roberto Navigli. 2021. SPRING Goes Online: End-to-End AMR Parsing and Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 134–142, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2).

Rudi L Cilibrasi and Paul MB Vitanyi. 2007. The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3):370–383.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. 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.

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

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

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

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2018. The other side of the coin: Unsupervised disambiguation of potentially idiomatic expressions by contrasting senses. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 178–184, Santa Fe, New Mexico, USA. 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.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels

of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP - test suites for natural language processing. In *COLING 1996 Volume 2: The 16th International Conference on 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.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Juri Opitz, Angel Daza, and Anette Frank. 2021. Weisfeiler-leman in the bamboo: Novel AMR graph metrics and a benchmark for AMR graph similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.

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

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

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

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

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.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Igor Samenko, Alexey Tikhonov, and Ivan P. Yamshchikov. 2020. Synonyms and antonyms: Embedded conflict. arXiv:2004.12835.

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.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. 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.

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

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 *CheckList*'s functionalities and resources

As described in §3.1, *CheckList* contains the selected sentence pairs as well as the corresponding AMR structures and their human score grouped by linguistic phenomena in `json` format. It further includes the assigned scores for the test instances as well as code to run the implementation for the following metrics:

- BLEU
- Meteor
- chrF++
- BERTSCORE
- SMATCH
- S^2MATCH
- WLK
- WWLK

Output. The *CheckList* can be run from the command line, printing an overview of the data used, accompanied by statistics concerning human judgement for each phenomenon. These statistics include the mean, median, standard deviation and standard error of the human scores. Finally, it will output tables displaying the overall results of the *CheckList* (hereby, we use the evaluation measures that were also applied in the paper). If a metric were to be tested, it would furthermore print the correlation of that metric with the others in decreasing order.
The results for the phenomena are summarized in individual text files. These files once more list the statistics about the human score and then display the average scores of all metrics for that very phenomenon. Finally, each test case is listed, including the sentences as well as their AMR structures and the scores assigned to it by the metrics and the annotator.

A.2 STS Results

Table 7 and 8 and Fig. 5 demonstrate the results on the test cases selected from the STS data set. Table 9 shows a summary of metrics yielding Best and 2nd Best Results.

	Article	Aspect	Co-Hyponymy	Hyponymy	Omission	Overall
Ann. Score	0.967	1.0	0.282	0.647	0.77	0.647
BLEU	0.358 ± 0.61	0.155 ± 0.84	0.674 ± 0.48	0.58 ± 0.2	0.508 ± 0.27	0.503 ± 0.45
chrF++	0.661 ± 0.31	0.521 ± 0.48	0.661 ± 0.39	0.683 ± *0.12*	0.707 ± 0.14	0.654 ± 0.29
Meteor	0.385 ± 0.58	0.557 ± 0.44	0.313 ± **0.2**	0.462 ± 0.3	0.407 ± 0.36	0.408 ± 0.33
BERTSCORE	0.863 ± 0.1	0.824 ± 0.18	0.838 ± 0.56	0.761 ± *0.12*	0.801 ± **0.07**	0.816 ± 0.26
S^2MATCH	1.0 ± **0.03**	1.0 ± **0.0**	0.779 ± 0.5	0.737 ± 0.13	0.785 ± *0.09*	0.83 ± 0.21
SMATCH	1.0 ± **0.03**	1.0 ± **0.0**	0.779 ± 0.5	0.737 ± 0.13	0.785 ± *0.09*	0.83 ± 0.21
WLK	1.0 ± **0.03**	1.0 ± **0.0**	0.459 ± *0.25*	0.426 ± 0.23	0.733 ± 0.11	0.659 ± **0.15**
WWLK	1.0 ± **0.03**	1.0 ± **0.0**	0.689 ± 0.41	0.587 ± **0.1**	0.612 ± 0.19	0.732 ± *0.2*
GRACO_{gl}	1.0 ± **0.03**	0.859 ± 0.14	0.936 ± 0.65	0.963 ± 0.32	0.957 ± 0.19	0.94 ± 0.34
$\text{GRACO}_{gl}^{reduced}$	1.0 ± **0.03**	0.875 ± 0.12	0.924 ± 0.64	0.949 ± 0.3	0.322 ± 0.45	0.782 ± 0.39
GRACO	0.978 ± *0.05*	0.876 ± 0.12	0.969 ± 0.69	0.949 ± 0.3	0.961 ± 0.19	0.949 ± 0.35
$\text{GRACO}^{reduced}$	1.0 ± **0.03**	0.904 ± *0.1*	0.957 ± 0.67	0.939 ± 0.29	0.51 ± 0.26	0.841 ± 0.35

Table 7: Avg. normalized score & mean abs. deviation (most indicative, lower is better) from human score for STS

	Article	Aspect	Co-Hyponymy	Hyponym	Omission	Overall
BLEU	0.389	*0.52*	0.17	0.504	0.573	0.218
chrF++	*0.611*	0.1	*0.68*	0.653	0.511	0.403
Meteor	0.556	0.22	0.35	0.636	0.52	0.625
BERTSCORE	**0.722**	0.1	**0.75**	**0.785**	**0.689**	0.537
SMATCH	0.333	**1**	0.305	0.603	*0.591*	0.682
S^2MATCH	0.333	**1**	0.305	0.603	*0.591*	0.682
WLK	0.333	**1**	0.32	0.603	0.582	**0.748**
WWLK	0.333	**1**	0.67	*0.769*	0.582	*0.712*
GRACO_{gl}	0.333	0.1	0.655	0.62	0.316	0.579
$\text{GRACO}_{gl}^{reduced}$	0.333	0.1	0.665	0.587	0.538	0.52
GRACO	0.278	0.1	0.36	0.554	0.493	0.417
$\text{GRACO}^{reduced}$	0.333	0.1	0.36	0.669	**0.689**	0.443

Table 8: Pairwise ranking scores for the STS test cases

	Best & 2nd Best Ranking Scores	**Best** & 2nd Best MAD	**Highest** & 2nd Highest Correlation w/ Human
BLEU	Aspect		
chrF++	Co-Hyponymy, Article	Hyponymy	Article
Meteor		**Co-Hyponymy**	
BERTSc	**Hyponymy**, **Co-Hyponymy**, **Article**, **Omission**	**Omission**, Hyponymy	**Hyponymy**, **Article**, **Co-Hyponymy**, **Omission**
SMATCH	**Aspect**, Omission	**Aspect**, **Article**, Omission	Omission
S^2MATCH	**Aspect**, Omission	**Aspect**, **Article**, Omission	Omission
WLK	**Aspect**	**Aspect**, **Article**, Co-Hyponymy	
WWLK	**Aspect**, Hyponymy	**Aspect**, **Article**, **Hyponymy**	Hyponymy, Co-Hyponymy
GRACO$_{glo}$		**Article**	
GraCo$^{red}_{glo}$		**Article**	
GRACO		Article	
GraCo^{red}	**Omission**	**Article**, Aspect	

Table 9: Overview over **Best** and 2nd Best Metric Performances in Ranking Score, MAD and Corr. to Human Scores for the STS dataset.

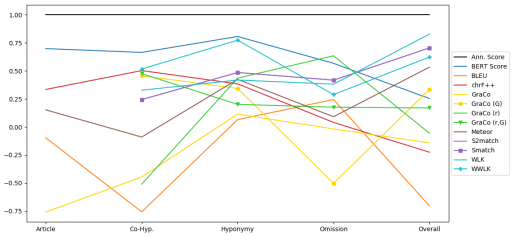

Figure 5: Spearman's rho correlation between metric and human scores for STS. *Aspect* is not included since all annotated scores are 1.

A.3 Experimental Settings

A.3.1 Generating sentences from the *Little Prince* AMR corpus.

We investigated sentences generated from AMRs from the 'Little Prince Corpus'[12] using the AMR-to-text system of Song et al. (2018). We used their pretrained *G2S_silver_2m* model and validated it on test data from Song et al. (2018), with a difference of -0.35 points BLEU score. For the 'Little Prince', consisting of 1,562 sentences, we obtained a BLEU score of 13.5.

constructional	lexical	SICK	STS	SICK	STS
Negation		156	-		
Omission		155	15		
Passive		78	-		
Aspect		-	10		
Semantic Roles		8	-		
Subordinate Clauses		69	-		
	Antonymy			157	-
	Article			77	6
	Hyponymy			116	11
	Co-Hyponymy			35	20
	Partial Synonymy			26	-
		466	25	411	37
Overall				877	62

Table 10: Number of SICK and STS test cases grouped by linguistic phenomena

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

A.3.2 Data Statistics

The following figures show the distribution of the human human scores in the *CheckList* for the individual linguistic phenomena. SICK and STS are displayed separately.
Fig. 7 further displays the sentence length distribution for SICK and STS.

A.3.3 Implementation details of metrics

Here, we list the hyperparameters and libraries employed for the metrics used in the CheckList.

For the text-based metrics, we employ NLTK's implementation for **BLEU**, where we add the method4 smoothing function (Bird et al., 2009)[13]; for **chrF++** use the sentence-level implementation by Popović (2015), and for **Meteor** the Version 1.5 implementation by Denkowski and Lavie (2014).

For Zhang et al. (2020)'s embedding-based metric **BERTSCORE**, we employ the implementation provided by Huggingface[14].

As for graph-based metrics, we made use of the implementations of **SMATCH** and the refined **S^{2}MATCH** provided by Opitz et al. (2020). For S^{2}MATCH we defined a cut-off threshold of 0.9, so that only concepts with a cosine similarity above that threshold would be granted a soft match. Further, the coefficient by which the similarity of differing senses is multiplied was set to 0.95.

For WLK and WWLK we employ the implementation by Opitz et al. (2021) without any additional hyperparameters.

For the implementation of the GRACO, we used the AMR Alignment tool from JAMR (Flanigan et al., 2014b) to align words from the sentence with concepts in the AMR structure. For concepts that have been successfully

[13] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[14] https://huggingface.co/metrics/bertscore

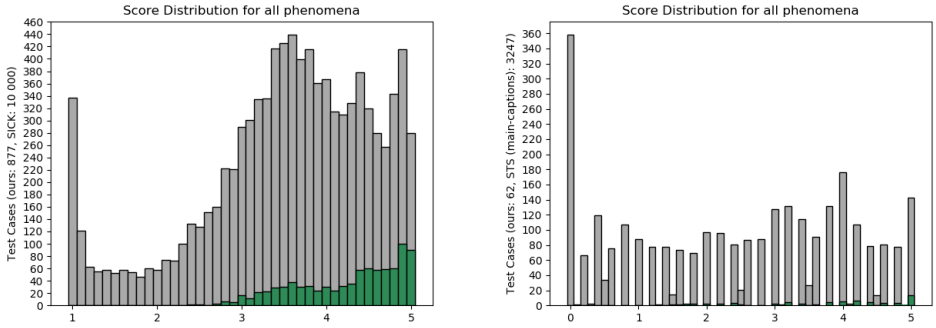

Figure 6: Score distribution for the test cases in the *CheckList* (green) grouped by SICK (left) and STS (right) test cases alongside the distribution of the whole datasets (grey)

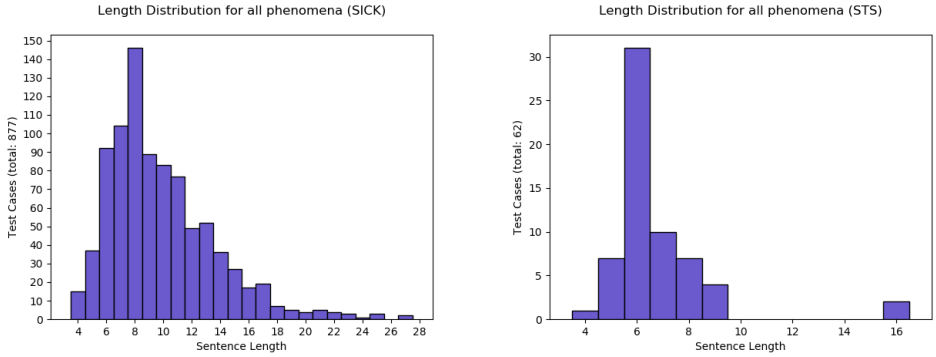

Figure 7: Sentence length distribution for the test cases in the *CheckList* grouped by SICK (left) and STS (right) test cases

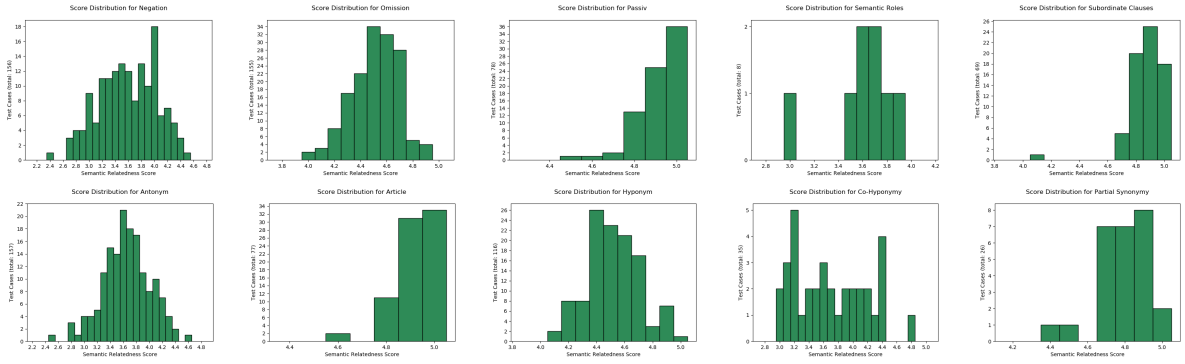

Figure 8: Score distributions for SICK per phenomenon: top: a.) Negation, b. Omission, c. Passive, d. Sem. Roles, e. subord. Clauses; bottom: f. Antonymy, g. Article, h. Hymonymy, i. Co-Hyponymy, j. Partial Synonymy.

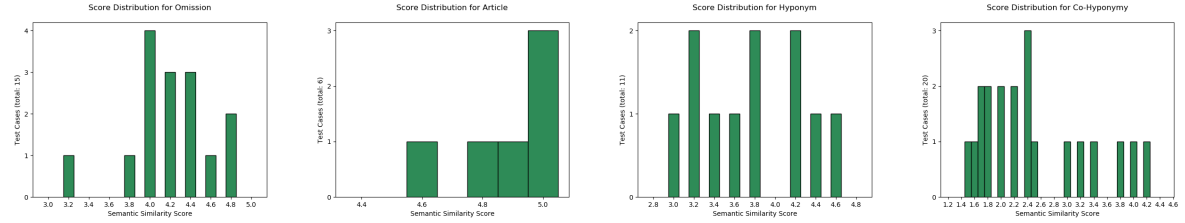

Figure 9: Score distributions for STS per phenomenon: b. Omission, g. Article, h. Hymonymy, i. Co-Hyponymy.

aligned, we experimented with contextualized BERT word embeddings, for which we use the `bert-large-uncased` model with a dimensionality of 1024 (Devlin et al., 2019), and 300 dimensional pretrained GloVe word embeddings (Pennington et al., 2014). In case GloVe may not have seen some inflected word, the embedding of its lemma will be used instead (the lemmata are obtained using the spacy lemmatizer and the `en_core_web_sm` model). If neither the token nor its lemma is contained in the vocabulary, we generate a zero vector representing an unknown token.

For standardization, given a metric predicts $s = \{s_1, ...s_n\}$, where n is the size of the data, we define the standardized score for an example i as $s'_i = \frac{s_i - mean(s)}{std(s)}$. Given s as above, the normalized score for an example i is defined as $s'_i = \frac{s_i - min(s)}{max(s) - min(s)}$.

A.4 Phenomena

A.4.1 Negation

We display two types of negation. In Fig. 10 the whole sentence is negated since `polarity` is attached to the matrix verb. Fig. 11 shows an AMR where only one constituent in a coordinated sentence is negated.

```
(xv0 / exercise-01
      :ARG0 (xv1 / man)
      :polarity - )
```

Figure 10: AMR for the sentence *The man is not doing excercises.* Semantic relatedness score: 3.8

```
(xv0 / and
      :op1 (xv1 / walk-01
            :ARG0 (xv3 / child))
      :op2 (xv2 / pull-up-07
            :ARG1 (xv5 / jeep-01)
            :polarity - )
```

Figure 11: AMR for the sentence *A child is walking and a jeep is not pulling up.* Semantic relatedness score: 3.5

A.4.2 Omission and Hallucination

Fig. 13 demonstrates the AMR of the sentence *The man is cautiously operating a stenograph*. The adverb is realized by the use of the role `:manner`. The sentence *The man is operating a stenograph* would look the same, except that the red-colored branch would not exist. Since concepts can be described in various ways, some words may be represented by more than one branch which would lead to more than two triples that don't have a counterpart. The omission of a prepositional phrase often resembles the omission of adjectives or adverbs, especially for phrases that can be realized by so-called "none-core-roles" such as `destination`, `location` or `medium`, hence, within one branch. As described in section A.3, prepositions, however, can be realized in various ways. The omission of a prepositional expression might therefore concern only one branch, but can also concern multiple branches like in Fig. 14.

A.4.3 Passive

Since AMR aims to capture the events of a sentence and not necessarily its *point of view*, AMR structures of an active-passive sentence pair do not differ at all.

A.4.4 Semantic and Syntactic Role Switch

The AMRs in Fig. 15 show that semantic and syntactic role switch is expressed by switching the `:ARG` roles. This results in the pair of AMRs differing in two triples.

A.4.5 Subordinate Clauses

In §4.1 we already discussed *inverse roles* for relative clauses when the relativizer is dependet on a verb. For attributive adjectives on the other hand, AMR structures should look the same. This is demonstrated by the AMR representations for *A black bird is sitting on a dead tree* and *A bird, which is black, is sitting on a dead tree* in Fig. 16. Fig. 12 displays a sentence pair featuring a noun compound expansion.

A.4.6 Article

Banarescu et al. (2013) specifically state that "AMR does not represent inflectional morphology for tense and number, and [...] omits articles".

A.4.7 Antonomy

In Fig. 17, we see two AMR graphs for a sentence pair exhibiting an antonymous relation between *young* and *old*. The antonymy is realized by mapping the differing concepts to the variable `xv3` respectively.

Another way of realizing antonymy between adjectives in an AMR graph is adding the feature

```
(xv0 / play-11                          (xv0 / play-11
     :ARG0 (xv2 / man)                       :ARG0 (xv2 / man)
     :ARG1 (xv1 / flute                      :ARG1 (xv1 / flute
          :consist-of (xv3 / bamboo))             :ARG1-of (xv3 / make-01
                                                       :ARG2 (xv4 / bamboo)))
```

Figure 12: AMR structures for the sentence pair *A man is playing a bamboo flute – A man is playing a flute made of bamboo* Semantic relatedness score: 4.9

```
(xv0 / operate-01
     :ARG0 (xv2 / man)
     :ARG1 (xv1 / stenograph)
          :manner (xv3 / cautious-02))
```

Figure 13: Gold AMR for the sentence *A man is cautiously operating a stenograph.* Semantic relatedness score: 4.5

```
(xv0 / attack-01
     :ARG0 (xv2 / dog
          :mod (xv3 / brown))
     :ARG1 (xv1 / animal)
     :location (xv4 / in-front-of
          :op1 (xv5 / man))
```

Figure 14: Gold AMR for the sentence *The brown dog is attacking an animal in front of the man.*

`:polarity` – to the branch of the adjective's concepts which inverts its meaning.

A.4.8 Hyperonymy, Hyponymy and Co-Hyponymy

An AMR structure of two sentences displaying a sub- or superset relation would differ merely in the concepts mapped to the corresponding variable as demonstrated in Fig. 18. This is also true for co-hyponymy.

```
(xv0 / follow-02                    (xv0 / follow-02
     :ARG0 (xv1 / turtle)                :ARG0 (xv2 / fish)
     :ARG1 (xv2 / fish))                 :ARG1 (xv1 / turtle))
```

Figure 15: AMR structures of the sentence pair *The turtle is following the fish. – The fish is following the turtle.* Semantic relatedness score: 3.8

```
(xv0 / sit-01
     :ARG1 (xv1 / bird
          :ARG1-of (xv3 / black-04))
     :ARG2 (xv2 / tree
          :ARG1-of (xv4 / die-01)))
```

Figure 16: AMR structure for the sentence pair *A black bird is sitting on a dead tree. – A bird, which is black, is sitting on a dead tree.* Semantic relatedness score: 5.0

```
(xv0 / talk-01                      (xv0 / talk-01
     :ARG0 (xv1 / man                    :ARG0 (xv1 / man
          :mod (xv3 / young))                 :mod (xv3 / old))
     :ARG2 (xv2 / leaf))                 :ARG2 (xv2 / leaf))
```

Figure 17: AMR structures for the sentence pair *A young man is talking to a leaf. – An old man is talking to the leaf.* Semantic relatedness score: 3.915

```
(xv0 / run-02                       (xv0 / run-02
     :ARG0 (xv2 / squirrel)              :ARG0 (xv2 / animal)
     :ARG1 (xv1 / circle))               :ARG1 (xv1 / circle))
```

Figure 18: AMR structures for the sentence pair *A squirrel is running in circles. – An animal is running in circles.* Semantic relatedness score: 4.4