AMR4NLI: Interpretable and robust NLI measures from semantic graphs

Juri Opitz[❀] **Shira Wein**[✿] **Julius Steen**[❀] **Anette Frank**[❀] **Nathan Schneider**[✿]

[❀]Heidelberg University [✿]Georgetown University
opitz.sci@gmail.com {steen,frank}@cl.uni-heidelberg.de
{sw1158,nathan.schneider}@georgetown.edu

Abstract

The task of natural language inference (NLI) asks whether a given premise (expressed in NL) entails a given NL hypothesis. NLI benchmarks contain human ratings of entailment, but the meaning relationships driving these ratings are not formalized. Can the underlying sentence pair relationships be made more explicit in an interpretable yet robust fashion? We compare semantic structures to represent premise and hypothesis, including *sets of contextualized embeddings* and *semantic graphs* (Abstract Meaning Representations), and measure whether the hypothesis is a semantic substructure of the premise, utilizing interpretable metrics. Our evaluation on three English benchmarks finds value in both contextualized embeddings and semantic graphs; moreover, they provide complementary signals, and can be leveraged together in a hybrid model.

1 Introduction

Natural language inference (NLI) and textual entailment (TE) assess whether a hypothesis ($\mathcal{H}$) is entailed by a premise ($\mathcal{P}$). Systems have various interesting applications, e.g., the validation of automatically generated text (Holtzman et al., 2018; Honovich et al., 2022). Recent systems make use of neural networks to encode $\mathcal{H}$ and $\mathcal{P}$ into a vector and thereupon make a prediction (Jiang and de Marneffe, 2019). While this can provide strong results when such systems are trained on large-scale training data, the overall decision process is not transparent and may rely more on spurious cues than on informed decisions (Poliak et al., 2018).

We aim to develop more transparent alternatives for NLI prediction, and therefore compare representations and metrics to predict entailment. Figure 1 gives an intuition of how 5 different sentences overlap in meaning. Representing each sentence with a semantic structure, we assume that, by and large, the semantic elements of an entailed sentence should be contained within the premise.

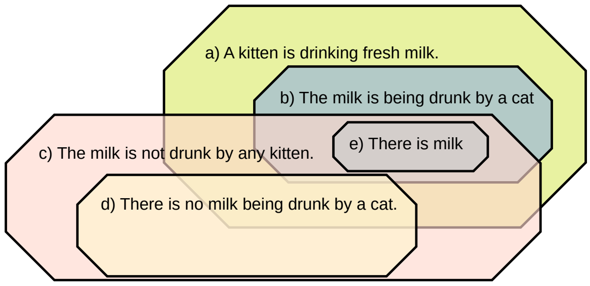

Figure 1: Semantic (sub-)structure analysis shows that 4 of 25 candidate relations are true entailment relations: b) is entailed by a). d) is entailed by c). e) is entailed by a), b), and c).

These considerations trigger three interesting research questions that we will investigate in this paper: RQ1. *How to characterize a semantic structure?* RQ2. *How to determine/measure what is a substructure?* RQ3. *Is there a suitable and interpretable structure and measure that help to make NLI judgments more robust, or more accurate?*

To assess RQ1, we test three options: token sets, sets of contextualized embeddings, or graph-based meaning representations (MRs). As a meaning representation, we select Abstract Meaning Representation (AMR; Banarescu et al., 2013), using automatic AMR parses of the NLI sentences. To assess RQ2, we test different types of metrics that are designed or adapted to measure entailment on the selected structures, inspired from research on, e.g., MT evaluation and MR similarity. One of our key goals is to investigate whether it is possible to accurately capture relevant semantic substructure relationships via meaning representations. Finally, we show that we can positively answer all aspects of RQ3: First, besides their enhanced interpretability, unsupervised semantic graph metrics are more robust and generalize better than fine-tuned BERT. Second, importantly, we show that they are high-precision NLI predictors, a property that we exploit to achieve strong NLI results with a simple decomposable hybrid model built from a fine-tuned BERT on the one hand, and a semantic graph score on the other. Code and data are available at `https://github.com/flipz357/AMR4NLI`.

2 Related work

Textual entailment Automatic approaches for this task date back to, at least, Dagan et al. (2006), who introduced a shared task for entailment classification. Since then, we can distinguish many different kinds of systems for addressing the task (Androutsopoulos and Malakasiotis, 2010), for instance, based on logics (Bos and Markert, 2005) or string- and tree-similarity (Zhang and Patrick, 2005), or graph matches of semantic frames and syntax (Burchardt and Frank, 2006) that aim in a similar direction as us. Recent releases of large-scale training corpora, such as SNLI (Bowman et al., 2015), or MNLI (Williams et al., 2018) can be exploited for supervised training of strong classifiers, e.g., by fine-tuning a BERT language model (Devlin et al., 2019). However, trained systems tend to suffer from the 'Clever Hans' effect and fall prey to spurious cues (Niven and Kao, 2019; Jin et al., 2020), such as position (Ko et al., 2020) or even gender (Sharma et al., 2021). This can lead to undesired and peculiar NLI system behavior. Poliak et al. (2018) show that supervised NLI systems can make many correct predictions solely based on $\mathcal{P}$, without even seeing $\mathcal{H}$. In our work, we want to test more transparent ways of rating entailment.

Metrics and meaning representations In part due to the reduced dependence on spurious cues, unsupervised/zero-shot metrics are found in evaluation of MT (e.g., BERTscore (Zhang et al., 2020), BLEURT (Sellam et al., 2020)), and NLG faithfulness checks (Honovich et al., 2022). Through the lens of abstract meaning representation (Banarescu et al., 2013), systems perform explainable sentence similarity (Opitz et al., 2021b; Opitz and Frank, 2022b), NLG evaluation (Opitz and Frank, 2021; Manning and Schneider, 2021), cross-lingual AMR analysis (Wein and Schneider, 2021, 2022; Wein et al., 2022), and search (Bonial et al., 2020; Müller and Kuwertz, 2022; Opitz et al., 2022). Leung et al. (2022) discuss different use-cases of embedding-based and MR-based metrics.

3 Method

3.1 Underlying research hypotheses

RH1: Semantic substructure analysis with asymmetric metrics can predict entailment We aim to study the entailment problem through analysis of semantic structure of $\mathcal{P}$ and $\mathcal{H}$. To perform such analysis, we need a metric that can measure the degree to which $\mathcal{H}$-structure is contained in the $\mathcal{P}$-structure. Therefore, we hypothesize that an *asymmetric metric* is preferable. Note that asymmetric metrics of complex objects like sets or graphs tend to be under-studied in NLP.[1]

RH2: Meaning representations are suitable semantic structures Semantic structures for $\mathcal{P}/\mathcal{H}$ should (ideally) hold facts that make them true. In this work we explore three options to build such structures for $\mathcal{H}/\mathcal{P}$: i) the set of text tokens, ii) the set of (contextual) embeddings obtained from them, and iii) graph-structured MRs. It is the latter that we hope will represent the facts best: A token set holds 'facts' in their surface form, which can be lossy in morphologically rich languages or with paraphrases. Contextual embedding sets, on the other hand, are powerful meaning representations, but hardly offer interpretability. An MR-structure is semantically more explicit, and is defined to represent a sentence's meaning through its parts.

3.2 Implementation

Preliminaries Let us define a

$$metric_T^{\mathcal{D}} : \mathcal{D} \times \mathcal{D} \rightarrow [0, 1] \quad (1)$$

where 1 implies true entailment. With the parameter $\mathcal{D}$ we denote the metric domain (i.e., text with $metric^{text}$ or MR with $metric^{graph}$). The type parameter T specifies whether the metric is symmetric ($metric_{sym}$), or asymmetric ($metric_{asym}$).

3.3 Text metrics: $metric^{text}$

Token metrics Given a set of tokens from $\mathcal{H}$ and from $\mathcal{P}$, our asymmetric $metric^{text}_{asym}$ calculates a

[1] Indeed, most metrics used in NLP are *naturally symmetric* (e.g., cosine distance). Others fuse two asymmetric metrics into, e.g., an F1 score from precision and recall (Popović, 2015; Zhang et al., 2020). Alternatively, they are inherently asymmetric but enforce symmetry via balancing with an inversely correlated metric, e.g., BLEU (Papineni et al., 2002) focuses on precision but tries to factor in recall via a 'brevity penalty'. Even in related cases, where using an asymmetric metric seems intuitive, we find that sometimes symmetric metrics being used instead, e.g., Ribeiro et al. (2022) design a baseline for assessing faithfulness of automatically generated summaries with a symmetric F1 score using an AMR metric.

unigram *precision*-score:

$$\text{TokP} = |\mathcal{H}|^{-1} \cdot |toks(\mathcal{H}) \cap toks(\mathcal{P})|, \quad (2)$$

which is known to be a simple but strong predictor baseline for NLI-related tasks such as faithfulness evaluation in generation (Lavie et al., 2004; Banerjee and Lavie, 2005; Fadaee et al., 2018) (the most closely related 'BLEU-1' is used in many papers to assess system outputs). By switching $\mathcal{H}$ and $\mathcal{P}$ in Eq. 2, we calculate TokR, and based on these a symmetric $metric^{text}_{sym}$ TokS via harmonic mean.

BERTscore (Zhang et al., 2020) is a contextual embedding metric that calculates a greedy match between BERT embeddings of two texts, in our case: hypothesis $E^{\mathcal{H}} := embeds(\mathcal{H})$ and premise $E^{\mathcal{P}} := embeds(\mathcal{P})$. For our asymmetric $metric^{text}_{asym}$, we calculate a precision-based score:

$$\text{BertScoP} = |E^{\mathcal{H}}|^{-1} \sum_{e \in E^{\mathcal{H}}} \max_{e' \in E^{\mathcal{P}}} e^T e'. \quad (3)$$

Symmetric $metric^{text}_{sym}$ BertS is calculated as harmonic mean of BertScoP and BertScoR, the latter being obtained by switching $\mathcal{H}$ and $\mathcal{P}$ in Eq. 3.

3.4 MR Graph metrics: $metric^{graph}$

We study the following (a)symmetric MR metrics.

GTok Emulating TokP and TokS, we introduce GTokS and GTokP via Eq. 2 applied to two bags of graphs' node- and edge-labels.

Structural matching with Smatch (Cai and Knight, 2013) aligns triples of two graphs for best matching score, and returns precision (SmatchP) and a symmetric F1 score (SmatchS). We use the optimal ILP implementation of Opitz (2023).

Contextualized matching with WWLK aims at a joint and contextualized assessment of node semantics and node semantics informed by neighborhood structures. Therefore, Opitz et al. (2021a) first iteratively contextualize a vector representation for each node by averaging the embeddings of all nodes in their immediate neighborhood (the iteration count is indicated by K, which we set to 1). The normalized Euclidean distance of the concatenation of these refined vectors defines a cost matrix C, where C_{ij} is the distance of nodes $i \in \mathcal{P}$, $j \in \mathcal{H}$. The AMR similarity score is derived by solving a transportation problem: $WWLK = 1 - \min_F \sum_i \sum_j F_{ij} C_{ij}$ where F_{ij} is the flow between nodes i, j. Opitz et al. constrain $\sum_j F_{*j} = 1/|\mathcal{P}|$ and $\sum_i F_{i*} = 1/|\mathcal{H}|$. We call this symmetric setting **WWLKS**. We additionally propose an asymmetric sub-graph matching score **WWLKP** where we let $\sum_j F_{*j} \leq 1$ instead.

The most reduced version, which deletes all structural information from the graphs, is achieved by setting $k = 0$, which we denote as N(ode)Mover(P|S) score, analogously to the popular word mover's score (Kusner et al., 2015).

3.5 Hybrid model

Our decomposable hybrid model takes the prediction of a text metric, and the prediction of a graph metric, and returns an aggregate score. Such a metric can provide an interesting balance between a score grounded in a linguistic interpretation, and a score obtained from strong language models. If the two scores are both useful *and* complementary, we may even hope for a rise in overall results. To test such a scenario we will combine the best performing $metric_{graph}$ with the best performing $metric_{text}$ via a simple sum ($\alpha = 0.5$):

$$\alpha \cdot metric^{graph} + (1 - \alpha) metric^{text}. \quad (4)$$

4 Evaluation setup

Data sets We employ five standard sentence-level data sets: i) **SICK (test)** by Marelli et al. (2014) and **SNLI (dev & test)** by Bowman et al. (2015), as well as iii) **MNLI (matched & mismatched)** by Williams et al. (2018). Mismatched (henceforth referred to as MNLI-mi) can be understood as a supposedly more challenging data set since it contains entailment problems from a different domain than the training data, allowing a more robust generalization assessment of trained models. By contrast, in MNLI-ma(tched) the domain of the testing data matches that of the training data. For each data set, we map the three NLI labels to a binary TE classification setting, by merging *contradiction* and *neutral* to the *non-entailed* class.[2]

Evaluation metric We expect predictions to correlate with the probability of entailment, i.e.,

$$metric^{\mathcal{D}}_T(x, y) \uparrow \implies P(x \text{ entails } y) \uparrow,$$

[2]Same as in Uhrig et al. (2021), we use the T5-based off-the-shelf parser from amrlib for projecting AMR structures.

where $\uparrow$ means 'higher is better'. The NLI 'gold probability' labels are approximated as binary human majority labels. To circumvent a threshold search and obtain a meaningful evaluation score for comparing our metrics, we follow the advice of Honovich et al. (2022), who evaluate metrics for zero-shot faithfulness evaluation of automatic summarization systems, using mainly the Area Under Curve (AUC) metric. The AUC score is the probability that given randomly drawn instances ($\mathcal{P}, \mathcal{H}$, entailed) and ($\mathcal{P}', \mathcal{H}'$, non-entailed) the entailed instance receive a higher score. To rank metrics, we calculate two averages: AVG^{all} averages the scores over all data sets, while AVG^{nli} excludes SICK.[3]

Trained (upper-bound) We use a BERT trained on 500k SNLI examples.[4] It predicts an entailment probability from a vector representation generated by a transformer model.

5 Results

5.1 Main insights

Main insights can be inferred from Table 1. On all data sets, and overall on average, **asymmetric metrics substantially outperform symmetric metrics**. Sometimes they improve results by up to ten AUC points over their symmetric counterparts (e.g., NMoverS vs. NMoverP, +9.2). Comparing token sets, embedding sets and graphs, we find that both embedding set and graph prove advantageous: NMoverP achieves slightly better results than BertScoP, which has been *pre-trained* on large data. *Fine-tuned* BERT outperforms the tested unsupervised metrics when test data is in-domain (see SNLI results), but falls short at generalization. However, our **simple hybrid model can inform the output with sub-graph overlap and yields a strong boost outperforming all unsupervised and even trained metrics by a large margin (+4.5 points)**.

5.2 Analysis

Advantage of AMR and AMR metrics: high precision For each metric, we retrieve the p% most probable predictions, and calculate their accuracy. Results, averaged over all data sets, are displayed in Table 2. In high % levels, MR metrics outperform BertScoP by almost 20 points (e.g., BertScoP vs. WWLKP: +17.6 points), and even the fine-tuned BERT is strongly outperformed. Therefore, we can attribute the surprisingly strong performance of the graph metrics (and the hybrid model) to its potential for delivering high scores in which we can trust – if it determines that the semantic graph of $\mathcal{H}$ is (largely) a subgraph of $\mathcal{P}$, true entailment is most likely (in Appendix A, we show two examples).

Advantage of untrained (AMR) metrics: better robustness We check the robustness of our diverse NLI metrics on a controlled substructure of 3,261 SNLI testing examples by Gururangan et al. (2018), who removed examples that show spurious biases and/or annotation artifacts. Results in Table 3 show a catastrophic performance drop by trained BERT (−12.0 points), while untrained metrics such as TokP and WWLKP remain unaffected (+0.4 points) and WWLKP now even outperforms the SNLI-trained BERT model. Lastly, we see that the hybrid model can (partially) mitigate the drop introduced by its trained component (–7.3 points).

Discussion: graph metrics struggle with recall, and other limitations The MR metrics struggle with recall since they have problems to cope with MRs that strongly differ structurally, but not (much) semantically, which is a known issue (Opitz et al., 2021a). An example from our data is the following: In *The man rages*, *man* is the *arg0* of rage, while in the entailed sentence *A person is angry*, *person* is the *arg1* of *angry*, yielding large structural dissimilarity of MR graphs (SmatchP=0.0). In future work we aim to explore and improve this issue, such that we are able to identify that the experiencer of *angry* is strongly related to the *agent* of *rage*.

Potentially unrelated to the recall problem, other issues may hamper AMR usage for NLI, e.g., inconsistent copula modeling (Venant and Lareau, 2023), or parsing errors: even though parsers tend to provide high-quality output structures, they can still suffer from significant flaws (Opitz and Frank, 2022a), and thus their improvement may positively affect AMR4NLI performance.

Weights in hybrid model Recall that we can use α in Eq. 4 to weigh two metrics. We inspect different α in Figure 2 for fusing trainBERT (text) and WWLKP (graph, $\alpha \geq 0.5$: graph metric is weighted higher). While a balance ($\alpha \approx 0.5$) overall seems effective, SNLI profits if the text metric has more influence, and MNLI profits if the graph metric dominates. Finally, again we see more stable

[3]SICK contains entailment labels but not the direction of entailment and thus we do not include it in AVG^{nli}.

[4]https://huggingface.co/textattack/bert-base-uncased-snli

$\mathcal{D}$(omain)	*metric*	SICK	SNLI-dev	SNLI-test	MNLI-ma	MNLI-mi	AVG^{all}	AVG^{nli}
text	TokS	72.1	64.2	64.6	66.7	68.7	67.2	66.0
	TokP	74.7	70.0	70.6	68.2	70.3	70.8	69.8
	BertScoS	79.8	66.7	66.2	68.4	71.6	70.5	68.2
	BertScoP	**82.0**	74.5	74.0	**74.5**	**77.5**	76.5	75.1
AMR graph	GTokS	78.2	63.2	62.6	66.4	68.5	67.8	65.2
	GTokP	81.0	75.1	74.7	71.1	72.6	74.9	73.4
	NMoverS	77.7	65.8	64.9	66.7	68.5	68.7	66.5
	NMoverP	79.4	77.9	77.2	72.9	74.8	76.5	75.7
	SmatchS	76.3	63.3	62.3	65.7	67.6	67.0	64.7
	SmatchP	79.2	72.3	71.6	70.0	71.9	73.0	71.4
	WWLKS	77.2	66.4	65.6	65.7	67.5	68.5	66.3
	WWLKP	79.3	**78.0**	**77.3**	71.9	73.8	76.1	75.3
text	trainBERT	81.0	88.8	88.2	71.5	72.0	80.3	80.1
hybrid	trainBERT + WWLKP	**85.9**	**91.0**	**90.4**	**77.9**	**78.9**	**84.8**	**84.5**

Table 1: Overall AUC results on five data sets. The last two rows involve a trained component.

		AVG Accuracy scores									
$\mathcal{D}$(omain)	*metric*	1%	2%	3%	4%	5%	7%	10%	15%	AVG^{all}	AVG^{nli}
text	TokP	88.4	87.1	81.0	74.4	72.8	71.4	68.3	64.2	76.0	77.3
	BertScoP	74.5	74.0	73.3	73.9	73.9	73.0	72.0	69.4	73.0	73.8
AMR graph	GTokP	86.5	86.5	87.1	88.0	87.7	86.1	80.4	73.6	84.5	88.4
	NMoverP	85.3	84.5	85.0	85.2	86.2	84.7	82.4	74.2	83.4	89.6
	SmatchP	90.0	89.1	88.4	85.2	81.9	77.9	74.2	68.3	81.9	83.8
	WWLKP	**97.3**	**96.8**	**96.1**	**95.0**	**93.8**	88.4	82.4	74.8	90.6	90.7
text	trainBERT	84.5	84.0	82.9	81.5	80.6	79.0	76.8	73.2	80.3	81.9
hybrid	trainBERT + WWLKP	96.7	95.7	94.3	93.4	92.5	**90.2**	**86.7**	**82.2**	**91.5**	**92.9**

Table 2: Precision assessment. We select p% of a metric's highest predictions and check the ratio of true entailment.

training	no		yes	no	no/yes
domain	text	text embedding		AMR	hybrid
metric	TokP	BScoP	BERT	WWLKP	+BERT
AUC	71.0	71.4	76.2	77.7	83.1
AUC Δ	**+0.4**	**-3.6**	**-12.0**	**+0.4**	**-7.3**

Table 3: Evaluation on 3,261 *hard* SNLI-test examples. AUC Δ: observed change in performance (cf. Table 1).

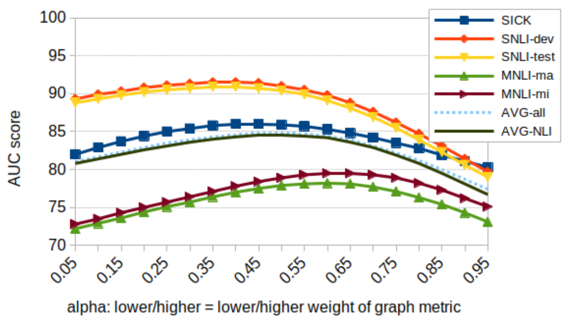

Figure 2: Balancing the hybrid text-graph metric.

performance of graph metrics overall (converging AUC with high α vs. diverging AUC with low α).

6 Conclusion

We find that metrics defined on advanced semantic representations are useful predictors of entailment. This is especially true for metrics performing asymmetric measurements on graph-structured meaning representations and sets of contextualized embeddings. Interestingly, meaning representation-based metrics offer advantages over strong embedding-based metrics beyond just interpretability: while showing similar performance as BERTscore, they are more robust than fine-tuned BERT *and* offer high-precision predictions. With this, we show that linguistic and neural representations can complement each other in a hybrid model, leading to substantial improvement over both untrained and trained neural approaches.

Acknowledgments

We thank anonymous reviewers for their feedback. This work is partially supported by a Clare Boothe Luce Scholarship.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

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.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Claire Bonial, Stephanie M. Lukin, David Doughty, Steven Hill, and Clare Voss. 2020. InfoForager: Leveraging semantic search with AMR for COVID-19 research. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 67–77, Barcelona Spain (online). Association for Computational Linguistics.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Aljoscha Burchardt and Anette Frank. 2006. Approaching textual entailment with lfg and framenet frames. In *Proc. of the Second PASCAL RTE Challenge Workshop.[-]*.

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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the Tip of the Iceberg: A Data Set for Idiom Translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating bert for natural language inference: A case study on the commitmentbank. 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 6086–6091.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 1109–1121, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. 2004. The significance of recall in automatic metrics for mt evaluation. In *Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004. Proceedings 6*, pages 134–143. Springer.

Wai Ching Leung, Shira Wein, and Nathan Schneider. 2022. Semantic similarity as a window into vector- and graph-based metrics. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 106–115, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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-2014)*, pages 216–223, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Almuth Müller and Achim Kuwertz. 2022. Evaluation of a semantic search approach based on amr for information retrieval in image exploitation. In *2022 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–6. IEEE.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.

Juri Opitz, Angel Daza, and Anette Frank. 2021a. 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 and Anette Frank. 2022a. Better Smatch = better parser? AMR evaluation is not so simple anymore. In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 32–43, Online. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2022b. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 625–638, Online only. Association for Computational Linguistics.

Juri Opitz, Philipp Heinisch, Philipp Wiesenbach, Philipp Cimiano, and Anette Frank. 2021b. Explainable unsupervised argument similarity rating with Abstract Meaning Representation and conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juri Opitz, Philipp Meier, and Anette Frank. 2022. Smaragd: Synthesized smatch for accurate and rapid amr graph distance. *arXiv preprint arXiv:2203.13226*.

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

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

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

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2021. Evaluating gender bias in natural language inference. *arXiv preprint arXiv:2105.05541*.

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

Antoine Venant and François Lareau. 2023. Predicates and entities in Abstract Meaning Representation. In *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*, pages 32–41, Washington, D.C. Association for Computational Linguistics.

Shira Wein, Wai Ching Leung, Yifu Mu, and Nathan Schneider. 2022. Effect of source language on AMR structure. In *Proceedings of The 16th Linguistic Annotation Workshop (LAW)*, Marseille, France. European Language Resources Association (ELRA).

Shira Wein and Nathan Schneider. 2021. Classifying divergences in cross-lingual AMR pairs. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 56–65, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2022. Accounting for language effect in the evaluation of cross-lingual AMR parsers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3824–3834, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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*.

Yitao Zhang and Jon Patrick. 2005. Paraphrase identification by text canonicalization. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 160–166.

A Appendix

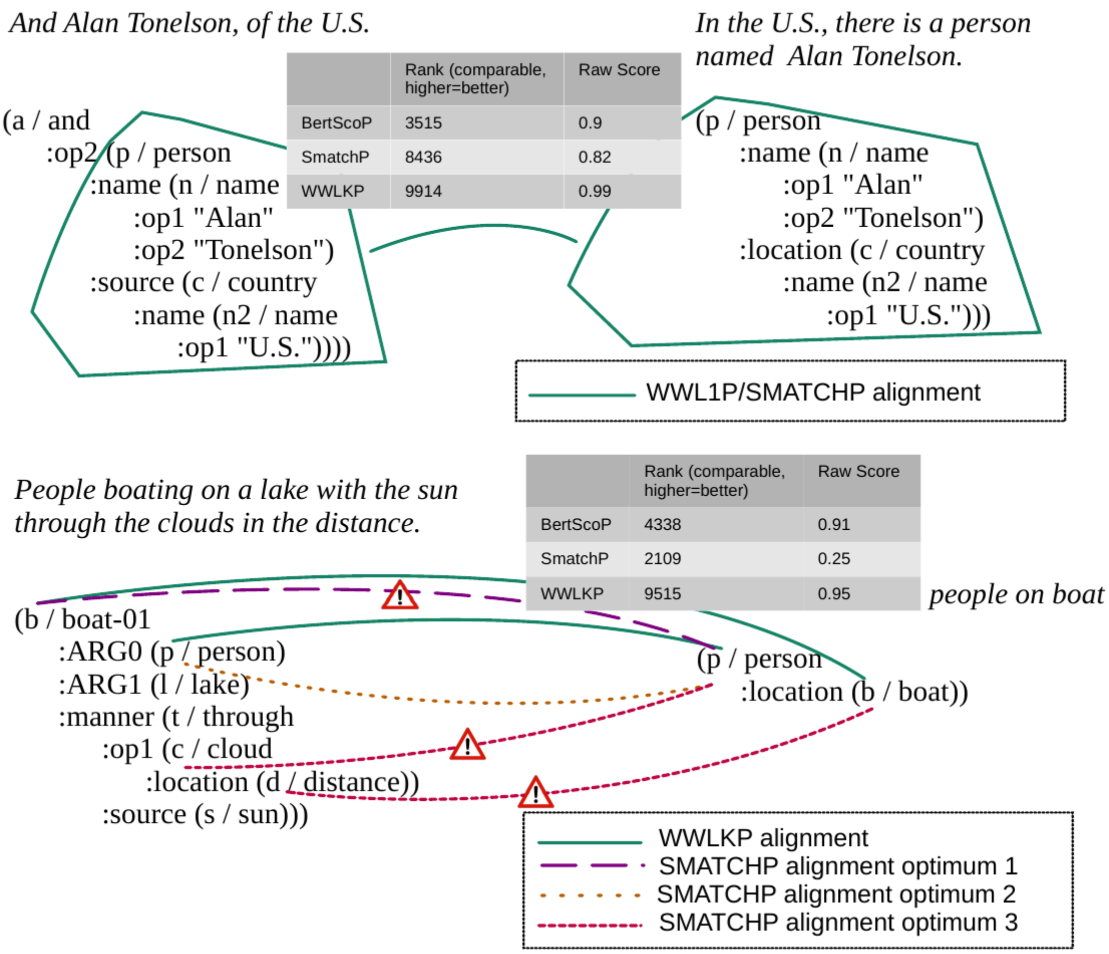

Figure 3: Two example ratings assessing true entailment: The first shows how MR can define a useful semantic set, the second shows that sometimes embedding-based graph metrics, such as WWLKP, are needed to assess the subgraph properly (in this example, SmatchP provides semantically meaningless alignments and a score that is too low.)