Evaluating AMR-to-English NLG Evaluation

Emma Manning **Shira Wein** **Nathan Schneider**
Georgetown University
{esm76, sw1158, nathan.schneider}@georgetown.edu

Introduction Abstract Meaning Representation, or AMR (Banarescu et al., 2013), is a representation of the meaning of a sentence as a rooted, labeled, directed acyclic graph. For example,

```
(l / label-01
  :ARG0 (c / country :wiki
"Georgia_(country)"
    :name (n / name :op1 "Georgia"))
  :ARG1 (s / support-01
    :ARG0 (c2 / country :wiki "Russia"
      :name (n2 / name :op1 "Russia")))
  :ARG2 (a / act-02
    :mod (a2 / annex-01)))
```

represents the sentence "Georgia labeled Russia's support an act of annexation." AMR does not represent some morphological and syntactic details such as tense, number, definiteness, and word order; thus, this same AMR could also represent alternate phrasings such as "Russia's support is being labeled an act of annexation by Georgia."

AMR generation is the task of generating a sentence in natural language (in this case, English) from an AMR graph. Like other Natural Language Generation (NLG) tasks, this is difficult to evaluate due to the range of possible valid sentences corresponding to any single AMR.

AMR generation systems are often evaluated only with automatic metrics such as BLEU (Papineni et al., 2002) that compare a generated sentence to a single human-authored reference; for AMR, this is the sentence from which the AMR graph was created. However, there is evidence that these metrics may not be a good representation of human judgments for AMR generation (May and Priyadarshi, 2017) and NLG in general. Thus, we present a new human evaluation of several recent AMR generation systems, most of which had not previously been manually evaluated.

System	$F_{\uparrow}$		$A_{\uparrow}$		$INC_{\downarrow}$	$MI_{\downarrow}$	$AI_{\downarrow}$
Konstas	**78.14**	1	**81.46**	1	**10.0**	34.5	12.0
Zhu	71.61	2	74.13	2	15.5	36.0	25.5
Ribeiro	67.05	3	64.37	4	19.5	47.0	31.5
Guo	62.13	4	68.52	3	22.0	41.0	21.5
Manning	36.89	5	54.10	5	57.5	**17.5**	**9.0**
Reference	87.56		93.68		5.0	4.5	10.0

Table 1: For each system, average fluency and adequacy scores and percentage where each adequacy error type was selected.

Methodology We conduct a human evaluation of several AMR generation systems: Konstas et al. (2017), Guo et al. (2019), Manning (2019), Ribeiro et al. (2019), and Zhu et al. (2019).

We sample 100 AMRs from the LDC2017T10 AMR test set; for each of these, we collect judgments on 6 sentences: the reference, and the output produced by each of the 5 generation systems. Each sentence is double-annotated by two annotators.

Annotators give separate scalar scores for fluency and adequacy via sliders representing an underlying 0-100 scale. They also give binary judgments of where certain types of errors apply:

- They cannot understand the meaning of the utterance (i.e. it is disfluent enough to be incomprehensible, making it difficult to meaningfully judge adequacy)
- Info in the AMR is missing from the utterance
- Info not in the AMR is added in the utterance

Annotators assess the fluency of each sentence based on the sentence alone; when assessing adequacy and error types, they are shown the AMR alongside the generated sentence.

Quality of Systems Table 1 shows the average score given for each system for fluency and adequacy, and how often each was marked as having each adequacy error type. We find that on both fluency and adequacy scores, Konstas performs best, followed by Zhu, and Manning performs the worst. Guo and Ribeiro are in between and within 5 points

System	BLEU$_\uparrow$	MET$_\uparrow$	TER$_\downarrow$	CF$_\uparrow$	BERT$_\uparrow$
Konstas	**38.1**	**39.2**	45.1	**64.3**	**95.0**
Zhu	**38.1**	38.7	**44.2**	56.3	92.7
Ribeiro	31.9	35.8	53.8	52.1	92.4
Guo	28.1	35.0	56.7	50.2	92.4
Manning	10.6	28.1	67.6	48.5	89.8

Table 2: Each system's scores on automatic metrics for the 100 sentences used in the human evaluation. MET = METEOR; CF = CHRF++; BERT = BERTScore.

Metric	Fluency	Adequacy
BLEU$_\uparrow$	0.40	0.52
METEOR$_\uparrow$	0.41	0.57
TER$_\downarrow$	-0.33	-0.43
CHRF++$_\uparrow$	0.32	0.47
BERTScore$_\uparrow$	**0.47**	**0.60**

Table 3: Sentence-level correlations of each metric with average human judgments.

of each other on each measure, with Ribeiro performing better on fluency and Guo on adequacy.

Comparison to Automatic Metrics To investigate how well automatic metrics align with human judgments of the relative quality of these systems, we compute BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), CHRF++ (Popović, 2017), and BERTScore (Zhang et al., 2020) for each system. Results are shown in table 2.

All these metrics at least agree with humans that the Konstas and Zhu systems are the best, followed by Ribeiro and Guo, and that Manning is the worst. However, there is some variation:

- Humans liked Konstas best; BLEU has it tied with Zhu, while TER finds Zhu slightly better.
- Humans prefer Ribeiro on fluency but prefer Guo on adequacy. All metrics except BERTScore prefer Ribeiro, while BERTScore has them tied.

Overall, these metrics mostly capture human rankings of these systems on this dataset. However, the results also highlight the limitations of metrics that produce single scores—while the metrics can only capture that Ribeiro and Guo are similar, our human evaluation found more nuance by identifying criteria on which each one outperforms the other.

Since all metrics give similar results on system-level rankings, we also calculate each metric's sentence-level correlation with human judgments for adequacy and fluency for more insight into the relative abilities of the metrics to capture human judgments. Results are shown in table 3. We find that each metric correlates more strongly with adequacy than with fluency, and that BERTScore has the strongest correlation with human judgments of both. Our results indicate that BERTScore is currently the strongest automatic metric for evaluating AMR generation, and that METEOR also appears slightly more reliable than BLEU.

Error Analysis To examine what factors contributed to particularly low scores, we identify and analyze sentences for which both annotators gave low fluency or adequacy ratings.

Added information is perhaps the most troubling form of error; AMR generation systems will have severely limited potential for use in practical applications as long as they hallucinate meaning. In one example, a reference to prostitution is inserted:

REF: A high-security Russian laboratory complex storing anthrax, plague and other deadly bacteria faces loosing electricity for lack of payment to the mosenergo electric utility.

RIBEIRO: the russian laboratory complex as a high - security complex will be faced with anthrax , prostitution , and and other killing bacterium losing electricity as it is lack of paying for mosenergo .

For the neural systems (all but Manning), common sources of low fluency scores included anonymization and repetition of words. For example, for the AMR in the introduction, Guo loses the word 'annexation' to anonymization:

GUO: georgia labels russia 's support for the <unk> act .

Several low-fluency sentences have unhumanlike repetition of words or phrases, for example:

REF: and I happen to LIKE large lot development .

RIBEIRO: and i happen to like a large lot of a lot .

Conclusion Our analysis points toward directions for researchers developing NLG systems, especially for AMR, to improve their output. We recommend seeking solutions to common issues that led to low scores, such as anonymization, repetition, and hallucination.

While this study found that popular automatic metrics were mostly successful in ranking these systems in the same order humans did, we also found that human evaluation could identify strengths and weaknesses of systems with more nuance than a single number can convey. We suggest that researchers in AMR generation and other NLG tasks continue to supplement automatic metrics with human evaluation as much as possible.

References

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. Association for Computational Linguistics.

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

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.

Emma Manning. 2019. A partially rule-based approach to AMR generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 61–70, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. 2019. Enhancing AMR-to-text generation with dual graph representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3183–3194, Hong Kong, China. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

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 (ICLR 2020)*, Online.

Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. 2019. Modeling graph structure in transformer for better AMR-to-text generation. 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 5459–5468, Hong Kong, China. Association for Computational Linguistics.