Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference

Ondřej Dušek and **Zdeněk Kasner**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia
{odusek,kasner}@ufal.mff.cuni.cz

Abstract

We propose a new metric for evaluating semantic accuracy of data-to-text generation based on a neural model pretrained for natural language inference (NLI). We use NLI to check textual entailment between data and text in both directions, allowing us to reveal omissions or hallucinations. Input data are converted to text for NLI using trivial templates. Our experiments show that our metric can achieve high accuracy in identifying erroneous system outputs (Dušek and Kasner, 2020).

1 Introduction

A major challenge in evaluating data-to-text (D2T) generation is measuring semantic accuracy, i.e. checking if a generated text contains all and only facts from input data. While state-of-the-art neural D2T models produce very natural outputs, they are prone to omitting or hallucinating facts (Gehrmann et al., 2018; Castro Ferreira et al., 2019; Dušek et al., 2020), which restricts their real-world deployment. Recognizing these errors is thus essential for proper evaluation.

Standard word-overlap metrics (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) or trained approaches to NLG evaluation (Zhang et al., 2020a; Sellam et al., 2020) do not cover accuracy explicitly. Handcrafted heuristics (Reed et al., 2018; Mi et al., 2019), which are the go-to method for evaluating D2T accuracy, are not able to capture the variety of possible hallucinations or omissions.

We note that the task of checking if a generated sentence includes/entails a particular fact is very close to the task of natural language inference (NLI). Recently, neural models for NLI (Zhang et al., 2020b; Liu et al., 2019a,b) reached near-human levels of performance and NLI was used for evaluating the output of abstractive summarization systems (Maynez et al., 2020). This leads us to propose a new metric for evaluating the semantic accuracy of D2T generation, which is based on a neural model pretrained for NLI. Using NLI, we check whether the text contains all facts from the input data and vice-versa. We demonstrate that this approach is viable and while not perfect, it can achieve high accuracy.

*This work was supported by the Charles University GAUK grant No. 140320, the SVV project No. 260575, and the Charles University project PRIMUS/19/SCI/10.

2 Our Approach

The input to our metric is a set of facts (the input for a D2T system) and the corresponding verbalization of these facts (the output from a D2T system). In our setup, the facts are RDF-like triples in the *subject-predicate-object* form. Our metric uses an NLI model to check textual entailment between the input data and the output text in both directions. By inferring input facts from the generated text, we can check for *omissions*, while the other direction allows us to check for *hallucinations*.

NLI is a sequence classification task which takes two inputs—a *hypothesis* and a *premise*—and produces one of the possible outputs: the hypothesis is *entailed* by (follows from) the premise, *contradicts* the premise, or their relation is *neutral*. We consider a NLI check as passed if the model predicts *entailment* as the most likely relationship between the premise and hypothesis texts.

The structured D2T input data are converted to text for use with the NLI model by a trivial template for each fact, handcrafted or extracted from NLG systems' training data. We consider two cases:

(1) *Default:* The templates can be handcrafted or extracted from the NLG systems' training data for each predicate.

(2) *Backoff:* We use only a single, universal "backoff" template for all the facts, in the form: *The <predicate> of <subject> is <object>.*

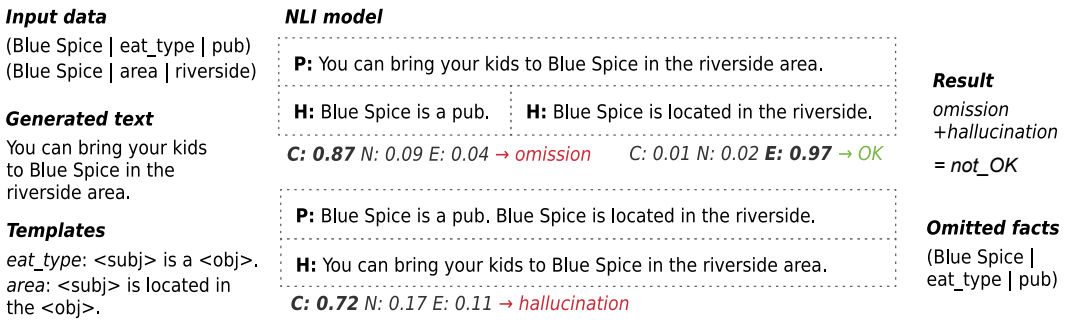

Figure 1: An example of evaluating the output from a D2T system with our metric. The generated text is used as a *premise* (*P*) to check for omissions and as a *hypothesis* (*H*) to check for hallucinations. The NLI model generates probabilities for *contradiction* (*C*), *neutral* (*N*) and *entailment* (*E*).

The generated text is considered correct if it mentions *all* and *only* the input facts. We thus check if the text contains any omissions or hallucinations in two steps (see Figure 1 for an example):

(1) To check for omissions, we use the whole generated text as a premise and sequentially feed each fact as a hypothesis to the NLI model. Any failed NLI check is considered an omission.

(2) To check for hallucinations, we use a concatenation of all facts as a premise and feed the generated text as a hypothesis to the NLI model. If this NLI check fails, the text is considered to contain hallucination.

Our metric's output is either 4-way: *OK* (all NLI checks passed), *omission*, *hallucination* or *omission+hallucination* (based on failed checks), or 2-way where the latter three collapse into *not_OK*.

3 Experiments and Results

We use pretrained RoBERTa (Liu et al., 2019b), which was finetuned on the MultiNLI data (Williams et al., 2018), as our NLI model. We use the model *as is*, without any further training.

We experiment with two recent English data-to-text datasets with a triple-like format: WebNLG (Gardent et al., 2017) and E2E (Novikova et al., 2017).[1] For our *Default* setup, we automatically extracted templates from single-triple examples in the WebNLG training data and handcrafted 8 templates for E2E. For WebNLG, we compare our metric with crowdsourced human ratings of semantic adequacy (Shimorina et al., 2019). For the E2E dataset, we compare against the handcrafted automatic script that was used to check the challenge results for semantic accuracy (Dušek et al., 2020). We use the 2-way outputs of our metric in both cases. We additionally performed a manual error analysis of 100 error examples for each dataset.

We show that even without any human references or in-domain training and with minimal handcrafting, our approach achieves high accuracy (*Default*: 93.3%, *Backoff*: 87.4%) on the E2E Challenge data (Dušek et al., 2020) when compared against the handcrafted evaluation script. Our manual error analysis identified several issues: (1) problems with interpreting some values (such as "less than £20" as "cheap"), (2) errors in the handcrafted automatic evaluation script, (3) edge cases (should "high restaurant" be considered OK for "high price range"?), (4) hallucinations that do not correspond to slots and cannot be detected by the handcrafted script. We consider 45 out of the 100 error examples as correctly classified by our metric.

On the more challenging WebNLG dataset, our metric performs worse but still produces useful results (*Default*: 77.5%, *Backoff*: 76.8% accuracy). Our manual error analysis indicates several reasons for differences from human judgments: (1) while our metric is binary, the human judgments are on a scale and it is hard to pick a threshold, (2) imprecise templates can confuse the NLI (cf. the less than 1% difference in the accuracy of the *Default* and *Backoff* setups; this could be mitigated by a better template selection), (3) human annotators tend to give lower score to accurate but ungrammatical or poorly organized texts. Again, our re-examination shows that 42 out of the 100 error examples were in fact correctly classified by our metric.

In sum, we believe that our approach is competitive with crowdsourced human ratings or handcrafted scripts customized for each domain while requiring much less manual effort.

[1] E2E data use attribute-value pairs relating to a restaurant; we convert them to triples where the restaurant is the subject.

References

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.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. 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 552–562, Hong Kong.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language*, 59:123–156.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG 2020)*, Online.

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

Sebastian Gehrmann, Falcon Z. Dai, Henry Elder, and Alexander M. Rush. 2018. End-to-End Content and Plan Selection for Data-to-Text Generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, Tilburg, The Netherlands.

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

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3151–3157.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany.

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.

Lena Reed, Shereen Oraby, and Marilyn Walker. 2018. Can neural generators for dialogue learn sentence planning and discourse structuring? In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 284–295, Tilburg University, The Netherlands.

Thibault Sellam, Dipanjan Das, and Ankur P. 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.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2019. WebNLG challenge: Human evaluation results. Technical report, LORIA.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating Text Generation with BERT. In *ICLR*, Online.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020b. Semantics-aware BERT for language understanding. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*.