

Evaluation Guidelines to Deal with Implicit Phenomena to Assess Factuality in Data-to-Text Generation

Roy Eisenstadt

Ben Gurion University
Computer Science Department
Be'er Sheva, Israel
royes@post.bgu.ac.il

Michael Elhadad

Ben Gurion University
Computer Science Department
Be'er Sheva, Israel
elhadad@cs.bgu.ac.il

Abstract

Data-to-text generation systems are trained on large datasets, such as WebNLG, RotoWire, E2E or DART. Beyond traditional token-overlap evaluation metrics (BLEU or METEOR), a key concern faced by recent generators is to control the factuality of the generated text with respect to the input data specification. We report on our experience when developing an automatic factuality evaluation system for data-to-text generation that we are testing on WebNLG and E2E data. We aim to prepare gold data annotated manually to identify cases where the text communicates more information than is warranted based on the input data (*extra*) or fails to communicate data that is part of the input (*missing*). While analyzing reference (*data*, *text*) samples, we encountered a range of systematic uncertainties that are related to cases on implicit phenomena in text, and the nature of non-linguistic knowledge we expect to be involved when assessing factuality. We derive from our experience a set of evaluation guidelines to reach high inter-annotator agreement on such cases.¹

1 Introduction

We investigate how to deal with implicit phenomena in text when assessing whether generated text is faithful to an input data specification. Recent data-to-text generation systems are trained on large dataset, such as WebNLG (Gardent et al., 2017), E2E (Novikova et al., 2017), WikiBio (Lebret et al., 2016), or RotoWire (Wiseman et al., 2017). Data-to-text systems are usually evaluated by comparing the generated text with reference text with metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) or BertScore (Zhang et al., 2020).

¹The guidelines are publicly available together with our annotated data on <https://github.com/royeis/FactualNLG>.

Yet, recent work has shown that neural generation models risk to create text that is not faithful to the input data specification (Wang, 2019; Chen et al., 2019a), by either introducing content which is not warranted by the input or failing to express some of the content which is part of the input. In fact, studies indicate that the training datasets also suffer from problematic alignment between data and text (e.g., (Rebuffel et al., 2020) and (Dušek et al., 2019)), because large-scale data collection is complex.

In response, new evaluation metrics are being developed to assess the factuality of text with respect to the input data. Goodrich et al. (2019) attempts to measure factuality by extracting data from the generated text using OpenIE techniques and comparing it with the original data. Dušek and Kasner (2020) exploit a Roberta-based NLI system to check bidirectional entailment relation between generated text and input data. Rebuffel et al. (2021) operate without reference text and use a QA-based evaluation method to assess whether the text answers questions that are generated on the basis of the input data.

In this paper, we investigate what it means for text to be faithful to data by revisiting human guidelines, specifically related to implicit phenomena in text. While manually assessing the factuality of text generated by a generator we developed on the WebNLG and E2E datasets, we encountered systematic uncertainties in deciding whether text was missing or unwarranted given the data. We faced such uncertainties in more than half of the cases we analyzed. We provide here a list of such cases that we categorize according to the type of implicit phenomenon which triggers uncertainty.

Our main contribution is a set of guidelines for human annotation of data-to-text datasets in terms of semantic alignment: does the text convey all the data in the input, and does it introduce unwarranted

```

#T1 <S>Ampara_Hospital<P>state<O>Eastern_Province,_Sri_Lanka
#T2 <S>Ampara_District<P>state<O>Eastern_Province,_Sri_Lanka
#T3 <S>Ampara_Hospital<P>region<O>Ampara_District
#T4 <S>Eastern_Province,_Sri_Lanka<P>leaderName<O>Austin_Fernando
#T5 <S>Sri_Lanka<P>leaderName<O>Ranil_Wickremesinghe

```

#Text The leader of Sri Lanka is Ranil Wickremesinghe, but in the Eastern Province it is Austin Fernando. This is where the Ampara Hospital is located in Ampara District.

Figure 1: Pragmatic Inference in WebNLG: the `leaderName` relation between Sri Lanka and Eastern Province conflicts under complex assumptions - leading to the realization with *but*. Is this warranted by the input?

or contradictory content.

2 Content Conveyed Implicitly by Text vs. Data

We report on observations gathered during a larger effort: to study the factuality of generated text vs. a data input specification, we sample (data, text) pairs from existing datasets, and synthesize new noisy pairs where we either add a predicate to the data side, remove one, or alter an existing predicate (e.g., transform a triplet *region(Ampara_Hospital, Ampara_District)* into *region(Ampara_Hospital, Northern_Province)*). Given these pairs (both original and synthesized), we manually annotate the pairs as either: **reliable** (text faithfully matches the data), **missing** (text fails to cover part of the data), **extra** (text hallucinates content which is not part of the data) and **perturbation** (a combination of missing and extra, meaning some of the content conveyed by the text was altered with respect to the input data).

While annotating this data, we identified systematic cases of uncertainty: We categorize six cases where the relation between generated text and the input data is uncertain, either because the text conveys content in an implicit way or because the input data entails additional facts. We give examples taken from reference text in the WebNLG dataset. We then provide statistics on the prevalence of these cases of vague semantic relation.

2.1 Non-arbitrary Labels in the Data

In WebNLG and WikiBio, entities are represented with strings derived from WikiData, which are often complex. For example, the label `Fall_Creek_Township, Madison_County_Indiana` refers to a specific township. The label is not transparent, in the sense that one can infer from the label itself a set of relations: `Fall_Creek` is a township, this township is located in the `Madison_County`, which is in

turn located in the Indiana state.

The relations expressed by the label itself are implicit: the fact that `Fall_Creek` is a `Township` is expressed by an underscore, but one cannot infer that `Fall` is a `Creek`. Similarly, location is expressed by commas in the label, but for different entity types, a different semantic relation would be expressed by the same mechanisms.

The annotation question that arises is whether semantic relations conveyed implicitly by non-arbitrary labels should be considered a part of the input to be conveyed. In other words, if a relation expressed in the label is not conveyed in the text, do we deem the text to be *missing*, and conversely, if the relation is expressed explicitly in the text, is it warranted by the input?

2.2 Bridging Anaphora

Bridging anaphora are effective at conveying a relation between parts of the text in a cohesive and succinct manner. In general, the resolution of bridging anaphora relies on non-linguistic knowledge. Consider the example (data, text) pair in Fig.1. The bridging reference in *the Eastern Province* (meant as a Province in Sri Lanka) is based on the non-arbitrary label of the Province. The fact that this Province is part of Sri Lanka is otherwise not stated in the input as an explicit relation.

If we consider that the label structure provides information in the input to be covered in the text, does the fact that a bridging anaphora is used cover this data? If conversely we consider that labels do not convey data to be covered, does the fact that the bridging anaphora requires the knowledge that the Province is located in the Country convey unwarranted extra information?

Similarly, in an example where the data states `location(Palace, London)`, `builtBy(Palace, Smith)`, `builtBy(OperaHouse, Smith)`, the text includes: *The Palace is located in London. The*

```

#T1 <S>Andrew_Rayel<P>associatedBand/associatedMusicalArtist<O>Armin_van_Buuren
#T2 <S>Andrew_Rayel<P>associatedBand/associatedMusicalArtist<O>Bobina
#T3 <S>Andrew_Rayel<P>associatedBand/associatedMusicalArtist<O>"Armin Van Buuren,
    Bobina, Mark Sixma"
#T4 <S>Andrew_Rayel<P>genre<O>Trance_music
#T5 <S>Trance_music<P>stylisticOrigin<O>Pop_music

#Text
Andrew Rayel has performed the genre of Trance music which has its stylistic origins in pop music.
He has been associated with the following musical artists: Bobina, Armin Van Buuren, Bobina, and Mark Sixma.

```

Figure 2: Collective properties in WebNLG

architect John Smith also built the Opera House. The relation between the palace and the architect is conveyed through a bridging reference, and is entailed by the usage of *also*. Do we annotate in such a case that the relation is covered by the text?

2.3 Conjunctions

In Fig.2, the same entity is associated through the same property to multiple values (T1, T2, T3). The name of the relation indicates that it is collective (i.e., when $r(s, o1)$ and $r(s, o2)$ we infer $r(s, (o1, o2))$), and hence, the realization can flatten the relation into a single conjunction. In this particular case, the input includes repetition (*Bobina* appears both in T2 and in T3), and the relation refers both to objects of types *Band* and *Artist*. The realization entails all the values in the conjunction are *Artists* (and not *Bands*). The fact that *Bobina* and *Sixma* are independent *Artists* is not stated in the input.

In other cases, though, repeated attributes are not to be understood as collective, but as successive events. For example, when describing the professional positions people took over their career. A sentence stating *Mr. X is president, a businessman and the host of a TV show* would introduce an unwarranted entailment (that the positions are filled simultaneously). Should such an implied conclusion be considered extra content unwarranted by the input?

2.4 Pragmatic Inference

In contrast to the monotonic relation seen in Fig.2, the example in Fig.1 uses the relation *leaderName*. The reference text relies on multiple phenomena to realize the following sentence:

The leader of Sri Lanka is Ranil Wickremesinghe, but in the Eastern Province it is Austin Fernando.

The usage of the *but* connective relies on multiple assumptions: First, the fact that the Province is part of Sri Lanka as discussed above; Second, the fact that a Province in a country has a differ-

ent leader than the country would be surprising (meaning, the province is separated, the leader of the province does not report to the leader of the country, there are not two leaders for one region).

A similar example appears in a reference text, where the facts *nationality* (*Anders, US*) and *birthPlace* (*Anders, Hong Kong*) are realized as: *William Anders, a US national (although born in British Hong Kong)*. The fact implied by the usage of *although* is the common sense assumption that being born outside of the US entails not being a US national. One more instance of this category is related to presuppositions. If the input data includes *languageSpoken* (*Philippines, PhilippinesSpanish*), can we infer that this is the only language spoken in the country? This determines whether the realization *The language spoken in the Philippines is Philippines Spanish* is faithful.

The annotation uncertainty is whether such semantic facts pragmatically inferred from the usage of connectives such as *but* or *although* or from presuppositions are warranted by the input.

2.5 Measurements: Units and Rounding

WebNLG covers domains such as description of astronomical entities and airports. In these domains, many facts are provided as measurements. Units are not encoded in a systematic manner in the data formalism. Generators tend to complete these units based on commonsense or world knowledge inferred from the domain (either as part of pre-trained language models or from the data-to-text training data).

For example, in Fig.3, the *mass* property has a unit explicitly specified in the input. In contrast, the reference text assumes the units for the *periapsis* and *orbitalPeriod* properties are kilometers. This turns out to be incorrect for *orbitalPeriod* (which should be measured in days or years).

```

#T1 <S> (19255)_1994_VK8<P>mass<O>5.6 (kilograms)
#T2 <S> (19255)_1994_VK8<P>periapsis<O>615591000000.0
#T3 <S> (19255)_1994_VK8<P>epoch<O>2006-12-31
#T4 <S> (19255)_1994_VK8<P>orbitalPeriod<O>8788850000.0
#T5 <S> (19255)_1994_VK8<P>apoapsis<O>6603633000.0 (kilometres)

#Text
The epoch of 19255 1994 VK8, which has a mass of 5.6 kilograms is December 31st, 2006.
Its orbital period is 8,788,850,000 kilometres, with a periapsis of 6,155,910,000,000 kilometres and an apoapsis of 6,603,633,000 kilometres.

```

Figure 3: Usage of units in WebNLG: the inferred units for `epoch` is incorrect.

The annotation uncertainty is whether text that leaves the units unspecified when the data has it specified is considered missing. Conversely, is the specification of units warranted by data input that does not specify units?

An additional uncertainty related to measurements is whether rounding in the text is acceptable: in the same example, would the text be acceptable with an approximate realization such as *an apoapsis of over 6 billion kms*.

2.6 Implicit World Knowledge, Implied Data, Redundant Data

Chen et al. (2019b) noted that data to text systems benefit from the introduction of additional background knowledge at training time, beyond the data observed in the dataset. Reliance on implicit world knowledge has become prevalent with the usage of large pre-trained language models which encapsulate such knowledge, such as RoBERTa or T5.

In many examples, the reference text refers to the type of an entity, even if the type is not part of the input. For example, in Fig.4, the fact that Turner is a musician is not stated in the input, yet it is mentioned explicitly in the reference text. This fact is entailed by the type of the properties in which the entity participates, but it can be left under-specified.

In other cases, the input data includes facts which can be considered redundant: either they can be inferred on the basis of other facts, or they are covered by the interpretation of non-arbitrary complex labels. Consider the example in Fig.5, the fact T1 is implied by T4 and the structure of the label `Spaceport_Launch_Pad_0` which indicates the Launch Pad is located in the Spaceport. The text does not cover explicitly the fact T1 (that the launch site of the rocket is the spaceport), but this is recoverable from the fact that the launch pad is mentioned in relation to the spaceport. Should this text be labeled as *missing* part of the input?

Finally, we observe many cases where content explicitly expressed in the text is in-

duced from predicates in the input. For example, in many cases in WebNLG, a configuration such as: (`City_X is_in County_Y`, `City_X is_in State_Z`) and the text conveys the induced fact (`County_Y is_in State_Z`) in a realization such as *city in county, state*. In this realization, implicit world knowledge indicates a transitive inclusion (*city in county in state*) but this chain is not explicitly present in the input.

3 Discussion

The review of the examples above illustrates the complexity of determining whether text conveys data in a faithful manner. In the same way as text conveys implicit content, we observe that the small data snippets currently used as input to data to text systems do not have precise semantics: are the relations collective, transitive, symmetric, time is not specified, entities are referenced with non-arbitrary labels which are interpreted in vague manner. As a consequence, we suggest that we should consider the task of aligning text with data as a text to text alignment, which demands the annotator to exploit world knowledge and common sense. We follow in this the approach of Dušek and Kasner (2020) who cast the task of factuality checking as bidirectional textual entailment and Rebuffel et al. (2021) who view it as question-answering. Our contribution is to translate this approach into more precise guidelines for human evaluation, taking into account aspects of implicit communication in language.

We have prepared a set of guidelines answering the uncertainties listed above on the basis of this general approach. Based on these guidelines, we have manually annotated 200 samples from WebNLG with two annotators. We found a high rate of samples in the reference data which suffer from poor alignment, as was reported in previous work for a variety of datasets (e.g., (Dušek et al., 2019)). We also find low alignment between our manual annotation and the automatic assessment

```

#T1 <S>Aaron Turner<P>ASSOCIATED_MUSICAL_ARTIST<O>Old Man Gloom
#T2 <S>Aaron Turner<P>ASSOCIATED_BAND_ASSOCIATED_MUSICAL_ARTIST<O>Lotus Eaters (band)
#T3 <S>Aaron Turner<P>GENRE<O>Black metal
#T4 <S>Aaron Turner<P>ORIGIN<O>United States
#T5 <S>Aaron Turner<P>ACTIVE_YEARS_START_YEAR<O>1995

#Text
Aaron Turner came from the U.S.
He is a Black metal musician who started performing in 1995.
He plays in the Lotus Eaters band having previously performed with Old Man Gloom.

```

Figure 4: Expression of implicit knowledge in WebNLG: the fact that Turner is a musician is not explicitly stated in the input

```

#T1 <S>Antares_(rocket)<P>launchSite<O>Mid-Atlantic_Regional_Spaceport
#T2 <S>Antares_(rocket)<P>comparable<O>Delta_II
#T3 <S>Delta_II<P>countryOrigin<O>United_States
#T4 <S>Antares_(rocket)<P>launchSite<O>Mid-Atlantic_Regional_Spaceport_Launch_Pad_0
#T5 <S>Mid-Atlantic_Regional_Spaceport_Launch_Pad_0<P>associatedRocket<O>Minotaur_IV

#Text
The Antares rocket is comparable to the Delta II, which originates from the United States.
The launch site of the Antares was the Mid Atlantic Regional Spaceport Launch Pad 0, which is also associated with the rocket Minotaur IV.

```

Figure 5: Redundant data in WebNLG input: T1 is implied by T4 and the form of the Launch0 label

tool provided by (Rebuffel et al., 2021). This indicates the task of assessing the semantic faithfulness of generated text in data to text remains challenging, both manually and automatically.

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. Association for Computational Linguistics.
- Shuang Chen, Jinpeng Wang, Xiaocheng Feng, Feng Jiang, Bing Qin, and Chin-Yew Lin. 2019a. [Enhancing neural data-to-text generation models with external background knowledge](#). 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 3022–3032, Hong Kong, China. Association for Computational Linguistics.
- Shuang Chen, Jinpeng Wang, Xiaocheng Feng, Feng Jiang, Bing Qin, and Chin-Yew Lin. 2019b. [Enhancing Neural Data-To-Text Generation Models with External Background Knowledge](#). 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 3022–3032, Hong Kong, China. Association for Computational Linguistics.
- Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. [Semantic noise matters for neural natural language generation](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.
- 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*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Ben Goodrich, Vinay Rao, Mohammad Saleh, and Peter J. Liu. 2019. [Assessing the factual accuracy of generated text](#). *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- Rémi Lebret, David Grangier, and Michael Auli. 2016. [Neural text generation from structured data with application to the biography domain](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.
- 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. 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.

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scuttheeten, and Patrick Gallinari. 2021. [Data-QuestEval: A Referenceless Metric for Data to Text Semantic Evaluation](#). *CoRR*, abs/2104.07555.

Clement Rebuffel, Laure Soulier, Geoffrey Scuttheeten, and Patrick Gallinari. 2020. [PAR-ENTing via model-agnostic reinforcement learning to correct pathological behaviors in data-to-text generation](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 120–130, Dublin, Ireland. Association for Computational Linguistics.

Hongmin Wang. 2019. [Revisiting challenges in data-to-text generation with fact grounding](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 311–322, Tokyo, Japan. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. [Challenges in data-to-document generation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. 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*.

A Data Description

We sampled 100 pairs (data, text) from the original WebNLG dataset, and expanded it with 100 additional pairs of synthetic perturbation of the data side (addition or retraction of a triplet, or transformation of an argument of an existing triplet).

We manually annotated each pair with the following labels (as shown in Fig.6):

1. **Factuality: OK**
2. **Factuality: missing** - in this case, we annotate which of the input triplets in the data is missing (`#Missing-pred`).
3. **Factuality: extra** - in this case, we annotate a span of text which is not warranted by the input data (`#Extra-content-in-text`).

4. **Factuality: perturbation** - in this case, we annotate both `#Missing-pred` and `#Extra-content-in-text`.

Finally, we manually identify which of the uncertainties which make the annotation difficult. Each case is labeled with one of the six categories identified in this paper:

1. **Complex Label**: a label conveying additional or redundant data with triplets in the data is present in the data.
2. **Bridging anaphora**: it is necessary to exploit world knowledge which may not be part of the input data to interpret a bridging anaphora.
3. **Aggregation**: data is aggregated in the text relying on the semantics of a relation in the data (collective, distributive).
4. **Pragmatic inference**: data in the input is implicated by the text through complex pragmatic inference (through presupposition, scalar implicature, marked by connectives).
5. **Units and rounding**: measurement is conveyed with unit that is inferred from the data (but not specified explicitly) or without unit; measurement is realized in an approximate manner.
6. **World Knowledge, Redundant Data, Implied Data**: input data is implied from content conveyed explicitly in the text, but it is not explicitly realized. Conversely, input data is logically redundant, and a redundant part of the data is not repeated in the text. Final case: content which can be inferred based on the type of the relations in the data is made explicit in the text (e.g., specify that a person is a Musician or an Architect even though this is not explicitly stated in the input data).

Each pair (data, text) can be annotated by multiple "uncertainties".

B Data Statistics

The prevalence of the uncertainty labels over the 200 manually annotated samples is shown in Table 1. We found similar frequency of the uncertainties over the original WebNLG sample and the synthetic noisy samples. We observe that these uncertainties are systematic: we found them on more than half of the pairs that we annotated.

```

#T1 <S>107_Camilla<P>discoverer<O>N._R._Pogson
#T2 <S>N._R._Pogson<P>deathPlace<O>Chennai
#T3 <S>107_Camilla<P>absoluteMagnitude<O>7.08
#T4 <S>N._R._Pogson<P>birthPlace<O>Nottingham
#T5 <S>N._R._Pogson<P>nationality<O>England

#Text
N. R. Pogson was born in Nottingham in the U.K. and died in Chennai.
He discovered 107 Camilla which has an absolute magnitude of 7.08.

#Factuality perturbation
#Missing-pred T5
#Extra-content-in-text "Nottingham in the U.K."

#Uncertainty
+ born in the UK implies nationality - 4 (Pragmatic inference)
+ absolute magnitude has no unit - 5 (Units and approximation)

```

Figure 6: Perturbed data in WebNLG: T5 is missing in the text which also conveys data not specified in the input

Label	Bridging	Aggregation	Pragmatic Inf.	Units	World Knowledge
52	12	35	63	12	81

Table 1: Number of occurrences of uncertainty labels over the 200 annotated samples

The distribution of the labels of factuality on the original WebNLG sample is shown in Table 2. We found that 20 of the 100 instances of the original WebNLG data were annotated with a non-reliable factuality label (missing, extra or perturbation). On the synthetic data, 95 of the 100 noisy label were annotated as non-reliable.

Factuality	#Occ.	Label	Bridging	Aggregation	Pragmatic Inf.	Units	World Knowledge
OK	80	23	3	20	22	2	29
Missing	11	6	0	1	5	0	12
Extra	4	0	0	0	2	0	2
Perturbation	5	1	0	1	1	0	3
Total	100	30	3	22	30	2	46

Table 2: Frequency of uncertainty labels per factuality on the original 100 WebNLG samples