Trainable Ranking Models to Evaluate the Semantic Accuracy of Data-to-Text Neural Generator

Nicolas Garneau and Luc Lamontagne
Université Laval, Québec, Canada
Computer Science Department
{first.last}@ift.ulaval.ca

Abstract

In this paper, we introduce a new embedding-based metric relying on trainable ranking models to evaluate the semantic accuracy of neural data-to-text generators. This metric is especially well suited to semantically and factually assess the performance of a text generator when tables can be associated with multiple references and table values contain textual utterances. We first present how one can implement and further *specialize* the metric by training the underlying ranking models on a legal Data-to-Text dataset. We show how it may provide a more robust evaluation than other evaluation schemes in challenging settings using a dataset comprising paraphrases between the table values and their respective references. Finally, we evaluate its generalization capabilities on a well-known dataset, WebNLG, by comparing it with human evaluation and a metric recently introduced based on natural language inference. We then illustrate how it naturally characterizes, both quantitatively and qualitatively, omissions and hallucinations.

1 Introduction

Data-to-Text (D2T) generation (Kukich, 1983; McKeown, 1985; Reiter and Dale, 1997) is a specialized task of natural language generation (NLG) where a model takes as input (semi)-structured data (e.g. a table) and generates a textual utterance that is both syntactical and semantically faithful to the input. Several architectures were proposed to solve this task. They may rely strictly on templates (Gatti et al., 2018; Puzikov and Gurevych, 2018; Wiseman et al., 2018), separate planning (what to say) from generation (how to say it) (Puduppully et al., 2019; Moryossef et al., 2019) or be a fully derivable neural architecture (Lebret et al., 2016; Wiseman et al., 2017; Gehrmann et al., 2018). While achieving interesting performance at natural language generation tasks (Lewis et al., 2020; Gehrmann et al., 2021), pre-trained neural language models

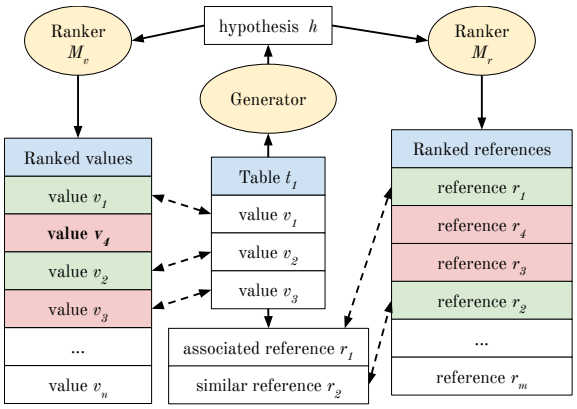

Figure 1: Round-trip evaluation in the table–hypothesis setting (left) and reference–hypothesis setting (right).

(hence neural architectures in general), are prone to hallucinate facts (Dušek et al., 2018) which brings their usability at stake in sensitive domains such as the legal one.

In this paper, we wish to promote the usability of neural architecture by proposing a new trainable automatic evaluation metric well suited to evaluate the semantic accuracy of such D2T generator. This metric is designed as a two factor "round-trip evaluation" in order to assess the accuracy a given generated hypothesis. First, we use the hypothesis to try to recreate the original table by ranking its values amongst all other values in the dataset. Then, we retrieve similar references amongst all other references in the dataset by ranking them still using that same hypothesis. We illustrate both round-trip evaluation scheme (table reconstruction and reference ranking) in Figure 1.

Our approach is well suited to semantically and factually assess a generator's performance in cases where the tables can be associated with multiple references, and the tables' values contain textual utterances. We present how one can further *specialize* the proposed evaluation metric by training the underlying ranking models on the target dataset, hence providing a more robust evaluation. Re-

lying on the mean average precision, we present how it naturally characterizes, both quantitatively and qualitatively, omissions and hallucinations of a given generator. This framework offers great flexibility in how it can be implemented and further improved. We identify two main components that can be tuned to improve the efficiency of our proposed metric when evaluating NLG systems; a similarity function between reference texts and the underlying ranking models.

However, having a metric where the efficiency is highly dependent on how it is implemented is highly problematic for an absolute comparison against other metrics or evaluation methodologies. This is why we propose a way to "fix a metric" i.e. how good, on the gold annotations, one implementation of the metric can be? Having a "fixed metric" then allows us to evaluate NLG systems against each other properly.

In our experiments, we first apply our metric on a challenging dataset in the legal domain, *Plum2Text* (Garneau et al., 2021). We show how the specialization of the ranking models can be beneficial or even necessary. More precisely, we illustrate these benefits when paraphrasing between the data (i.e. table) and the reference text highly characterizes the dataset in hand. Then, we illustrate its generalization capabilities even in simpler settings on a well-known D2T dataset, WebNLG (Gardent et al., 2017). We show how it is able to discriminate a set of generators, and correlates positively with human judgment. Our contribution is thus a new trainable automatic D2T evaluation metric that naturally characterizes both omissions and hallucinations of neural architectures.

2 Evaluating Data-to-Text Generation

Evaluating natural language generated text is a very hard task. Reiter and Belz (2009) and Reiter (2018) question the validity of widely used metrics. Sai et al. (2020) provide an extensive survey of the field, and more precisely separate the D2T evaluation metrics along 2 dimensions: either they use the table t or not (i.e. Table-Free), and either they are trained or untrained metrics. For instance, automatic evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) or METEOR (Banerjee and Lavie, 2005) are not trained and neither use the table t. They only partially account for the faithfulness of a given generated hypothesis w.r.t its associated references. Even though these metrics are widely used, they fall short to capture factual aspects in a D2T setting, and correlate poorly with human judgment (Liu et al., 2016; Novikova et al., 2017).

BLEURT (Sellam et al., 2020) is a metric trained on English texts that is designed to better model human judgment on generated texts. However, it does not take into account the input table. Sun and Zhou (2012) proposed iBLEU for paraphrase generation, an adaptation of the BLEU score that takes into account the context (the original phrase), the generated hypothesis and the reference. Its variant, BLEU-T, rewards an hypothesis h that overlaps with the content of the input table t as follows;

BLEU-T $= \alpha\text{BLEU}(h,t) + (1-\alpha)\text{BLEU}(h,r)$ where α is a parameter that balances faithfulness between the table t and the reference r.

	Table-Free	**Table**
Untrained	BLEU, ROUGE, METEOR	PARENT, BLEU-T/iBLEU, NLI, **Ours**
Trained	BLEURT	**Ours**

Table 1: Different metrics and their position in the evaluation spectrum, including the two variants of our proposed metric which can be trained or not.

Wiseman et al. (2017) proposed an extractive evaluation scheme where a model tries to identify relations in h between a pair of entities in order to recreate the table t that was used for the generation. Matching between the extracted entities and the table values is simply done via string-to-string comparison since the values in their dataset are short textual utterance of up to a few tokens. Dhingra et al. (2019) extended this metric (PARENT) by considering overlapping n-grams in the generated hypothesis h with both the table t and the reference text r. More recently, Dušek and Kasner (2020) proposed a metric that relies strictly on a pre-trained version of a natural language inference model (which will be referenced as "NLI" from now on) that verifies if a given hypothesis is entailed or not by the input table. They framed the evaluation as a categorical result given an hypothesis h (e.g. as being "Correct" or "Incorrect") but one can also use the underlying NLI model's confidence score for a softer evaluation. Using ranking models supplement their approach by identifying hallucinations *and* omissions (to some extent) and

quantitatively characterizes both phenomena.

Trained metrics using the context (in our case t) have been proposed for dialogue generation tasks (Lowe et al., 2017; Tao et al., 2018). However none of them is suitable for a D2T setting where the characterization of hallucinations (and to some extent omissions) are required in order to use a neural D2T generator in production. We thus wish to fill in this gap by presenting in the following section a new evaluation scheme. This scheme offers the advantage to exploit both the table t and the reference r, to be based on ranking models, and to be either trained or not. We illustrate in Table 1 where the metrics discussed in this section lie in the table–table-free and trained–untrained spectrum.

3 Data-to-Text Evaluation through Ranking

To assess the accuracy of a generated hypothesis h, it can be useful to consider both the input table t and the target reference r, i.e. validate the correctness of h according to its table and its reference. We thus propose a way to assess the fidelity of h by reconstructing the table t ($h \nrightarrow t$) and by retrieving its corresponding reference(s) r ($h \nrightarrow r$) using ranking models. This premise is highly motivated by the fact that similar text descriptions should be associated to semantically similar table contents. In both settings, the ranking models are evaluated using the Average Precision (AP), which we describe next.

3.1 Table Reconstruction

Borrowing from information retrieval terminology, in the context of $h \nrightarrow t$, we treat every value v in the table t as a document and h as a query. Different from the information extraction method proposed by Wiseman et al. (2017)[1], we wish to recreate t by finding the corresponding set of values $\mathbf{v}_i$ amongst the set of possible table values V, given a ranker $\mathcal{M}_v$ and the query h_i. To do so, we retrieve a ranked list of all the possible values $\hat{V} = \mathcal{M}_v(h_i)$ and compute the Precision at k (P@k) of the query h_i in the following way;

$$\text{P@}k_i = \frac{|\mathbf{v}_i@k \cap \hat{V}@k|}{|\hat{V}@k|} \tag{1}$$

We can then compute the Average Precision of the table reconstruction ($\text{AP}^{h \nrightarrow t}$) given the following formula;

$$\text{AP}^{h \nrightarrow t} = \frac{1}{|\mathbf{v}_i|} \sum_{k=1}^{|\hat{V}|} \text{P@}k_i, \tag{2}$$

giving us a sense of how well the ranker $\mathcal{M}_v$ is able to retrieve the set of table values $\mathbf{v}_i$ corresponding to the hypothesis h_i.

3.2 Reference Ranking

In the context of $h \nrightarrow r$, we still treat h as the query, but r as the document. In a case where multiple references can be deduced by the same table (or similar ones), it makes sense to take into consideration other references that share some similarities with the real reference. More precisely, a table t could refer to multiple references with a certain degree of correspondence, hence these references can be seen as similar documents. We can even push this further by assuming that if two tables t_i and t_j share similarities amongst their values, their corresponding references r_i and r_j will be semantically similar.

Given t_i, r_i and t_j, r_j, we define the following similarity function;

$$f : (t_i, t_j) \rightarrow d_{i,j} \tag{3}$$

such that $d_{i,j}$ is the degree of similarity between between two tables t_i and t_j. This similarity function is a proxy to the semantic similarity between r_i and r_j. For instance, by using the intersection over union of t_i and t_j table values, we use the following function; $f = (t_i \cap t_j)/(t_i \cup t_j)$. We thus consider, for a given r_i and its associated table t_i, the set of references where $d_{i,*} > \delta$ as being relevant references[2]. Given an hypothesis h_i, we can then query the set of references R in order to get a ranked list of references $\hat{R}_i = \mathcal{M}_r(h_i)$ where $\mathcal{M}_r$ is a ranking model for the references. We define R_i^* as being the ordered gold set of references according to f.

Let the Cumulative Relevance Score (CRS) of h_i be $\sum_{j=1}^{k} d_{i,j}$. We define the *estimated* and

[1]The information extraction scheme is relevant where the table values can be framed as triplets, where a model tries to put the different extracted entities into relation (e.g. the Rotowire dataset). Our method generalizes the table reconstruction step whereas one can freely design its own ranker model.

[2]We use $\delta > 0$ in our experiments.

true CRS at k being CRS applied on $\hat{R}_i$ and R_i^*, yielding $\hat{R}_i$-CRS@k and R_i^*-CRS@k respectively. Formally, $\hat{R}_i$-CRS@$k = \sum_{j=1}^{k} f(h_i, \hat{r}_j)$ and R_i^*-CRS@$k = \sum_{j=1}^{k} f(h_i, r_j^*)$. We thus compute Precision at k in the following way;

$$\text{P@}k_i = \frac{\hat{R}_i\text{-CRS@}k_i}{R_i^*\text{-CRS@}k_i} \tag{4}$$

obtaining the $\text{AP}^{h \to r}$ of h_i with the following formula;

$$\text{AP}^{h \to r} = \frac{1}{R_i^*\text{-CRS}} \sum_{k=1}^{|R|} \text{P@}k_i \times d_{i,k} \tag{5}$$

where $d_{i,k}$, properly scales P@k_i so that $\text{AP}^{h \to r}$ is between 0 and 1.

Finally, in both settings, we respectively compute the mean Average Precision (mAP) over the set of Hypotheses H;

$$\text{mAP}^{h \to t} = \frac{1}{|H|} \sum_{i=1}^{|H|} \text{AP}_i^{h \to t} \tag{6}$$

$$\text{mAP}^{h \to r} = \frac{1}{|H|} \sum_{i=1}^{|H|} \text{AP}_i^{h \to r} \tag{7}$$

where $\text{mAP}^{h \to t}$ illustrate the capacity of $\mathcal{M}_v$ to rank the hypotheses H accordingly to their respective table values, and $\text{mAP}^{h \to r}$ the capacity of $\mathcal{M}_r$ to rank the hypotheses H according to their similar references (and implicitly their respective tables).

In an ideal world, evaluating on gold annotations, both ranking models $\mathcal{M}_v$ and $\mathcal{M}_r$ should obtain a mAP of 1. In practice, however, we can only hope that each model will be as close as possible to 1, mostly due to noise in the data, annotation errors, or to the distribution of the data itself. In the next section, we introduce a robust ranking model based on sentence embeddings applied in both $h \to t$ and $h \to r$ settings. We also introduce how this model can be trained on the dataset in hand.

3.3 Training Ranking Models

We consider the embedding-based ranking models using the information retrieval version of Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), based on the BERT architecture (Devlin et al., 2019)[3]. More concretely, we use SBERT to encode both the set of possible table values V and the set of references R into a list of vector representations (i.e. matrices) $\mathbf{V}$ and $\mathbf{R}$. We then encode the hypothesis h into its respective vector representation $\mathbf{h}$. The models $\mathcal{M}_v$ and $\mathcal{M}_r$ re-order the vectors in $\mathbf{V}$ and $\mathbf{R}$ according to the cosine similarity with $\mathbf{h}$, yielding $\hat{V}$ and $\hat{R}$ needed for the computation of our metric.

Reimers and Gurevych (2019) showed that fine-tuning SBERT on the downstream task's dataset can lead to substantial improvements. To this end, we propose to fine-tune SBERT in both settings, $h \to t$ and $h \to r$, by creating their very specific training datasets. In our experiments, we used the multilingual version of BERT (Devlin et al., 2019) as the base model, XLM-R (Conneau et al., 2020; Reimers and Gurevych, 2020).

In the $h \to t$ setting, for every value $v_{i,j}$ in table t_i with its corresponding reference r_i, we create a positive pair $(v_{i,j}, r_i, 1)$. We then randomly sample negative examples such that $(v_{i,j}, r_m)$ is not within the original dataset and fine-tune SBERT to discriminate positive from negative pairs, $(v_{i,j}, r_m, 0)$.

In the $h \to r$ setting, we use Equation 3 to determine the similarity between two given references r_i and r_j. Using the cartesian product of $R \times R$, we thus generate every possible reference pairs with their respective similarity value, $(r_i, r_j, d_{i,j})$ and fine-tune SBERT to maximize the similarity between similar pairs, and minimize it between dissimilar pairs. In practice, we down sample pairs where $d_{i,j} = 0$ since it corresponds to 80% of the generated pairs. While this training procedure is very generic an applicable to most D2T datasets, it can be modified to suit one's specifics.

In both settings, we split the training data in a train and validation sets of 80% and 20% respectively. We fine-tuned SBERT for 4 epochs, keeping only the best performing model on the validation set. Regardless of the downstream dataset used, using a GeForce 2080 graphic card, this process took 4 hours for each setting.

3.4 Characterizing omissions and hallucinations

An insightful way of qualitatively analyzing the capacity of a given neural D2T generator is by characterizing omissions and hallucinations. More precisely, we want to know which element from the table may have been forgotten by the generator and which element in the generated text may be consid-

[3] In our experiments, we also tested a word co-occurrence ranking, Elasticsearch (https://www.elastic.co/), as explained in Section 4.1.

ered as hallucinations (Dušek et al., 2018; Dušek and Kasner, 2020). We acknowledge the fact that characterizing omissions may not be relevant in cases we would only like to describe highlights of a basketball game (Wiseman et al., 2017), especially when there is a seperate planning step (Puduppully et al., 2019; Moryossef et al., 2019). However, in the legal domain, describing a semi-structured document in its whole (omissions), and solely this document (hallucinations), is of high importance to foster a truthful view of a legal system (Beauchemin et al., 2020).

Using a ranking-based metric, we can potentially identify which elements from the table are considered as omissions and what has been hallucinated in the hypothesis. Intuitively, ranking models basically offer this characterization for free. Indeed, on the table reconstruction side, considering the gold set of n values v and the set of top n returned values $\hat{v}$ from $\mathcal{M}_t$, we obtain the omissions by computing the set difference of $o_t = v - \hat{v}$. As per the implicit definition of hallucinations and assuming a neural generator has been trained on a training set, we define hallucinations as being values from the training set that have been highly ranked. Therefore, we compute hallucinations as $a_t = \hat{v} - v$ such that $a_t \in V_{train}$. On the reference retrieval side, we consider the first retrieved reference and its associated table (or set of values), $\hat{v}$. We can similarly compute omissions and hallucinations as on the table reconstruction side, obtaining o_r and a_r. A given value $v_i \in v$ will considered omitted if it is present in o_t and o_r (i.e. 1.0), partially omitted if it is present in either o_t or o_r (i.e. 0.5) and not omitted if it is not present in any of the sets (i.e. 0.0). The same logic applies to hallucinations.

Roughly speaking, the mean average precision is a quantitative proxy loosely characterizing the omissions and hallucinations. Indeed, the average precision will consider not only the top n, but up to the last value that should have been retrieve in the list, thus making it an optimistic approximation of omissions and hallucinations. On the reference side, then again our metric is an optimistic approximation of omissions and hallucination in the sense that we consider every other references having a similarity score > 0. One could design a more exact quantitative approximation of omissions and hallucinations by only considering the first returned reference by the ranking model and analyzing its table with the corresponding true table t, as previously proposed.

	$h \to t$	$h \to r$	**Avg.**
Elasticsearch	0.588	0.596	0.592
SBERT Untrained	0.274	0.584	0.429
SBERT Trained	**0.831**	**0.871**	**0.851**

Table 2: Results of "fixing the metric" on *Plum2Text* using Elasticsearch, SBERT Untrained and SBERT Trained as different ranking models.

4 Experiments

In this section, we first illustrate the benefits of fine-tuning our proposed metric on the *Plum2Text* target dataset (Garneau et al., 2021). We then show how our metric can discriminate generators, and analyze omission and hallucination rates using WebNLG (Gardent et al., 2017).

4.1 Experiments with *Plum2Text*

In this section, we apply our metric on a challenging French dataset in the legal field, *Plum2Text* (Garneau et al., 2021), comprising references being a paraphrase of the table's values. It is composed of pairs of *plumitif*–description. A plumitif is a structured document containing all the key steps of a judicial case. The purpose of this dataset is to make the *plumitifs* more understandable to the population by generating a description from the input data.

There is an interesting exercise when comes the time to evaluate the effectiveness of a metric, especially when it is embedding-based. We dub this exercise as "fixing the metric", whereas we apply it on the gold annotations, i.e. $h = r$. This exercise tells us, to some extent, "how far we can go" with a given generator w.r.t the evaluation metric in the case we would know the answer. Using metrics based on word overlap would obviously yield a perfect score in the reference ranking setting. In these experiments, we thus only consider "fixing the metric" on the *Plum2Text* dataset since we do not have access to human evaluation over systems' outputs. This illustrates the challenge posed by the *Plum2Text* dataset (paraphrasing) as well as the benefits of fine-tuning our metric. As a baseline, we use Elasticsearch's ranking model (ES) based on word co-occurrence.

We can see from Table 2 that, on the Gold annotations, ES has decent performance on both $h \to t$ and $h \to r$. Supporting the findings of Reimers and

h	$\nrightarrow$	t	Rank of v	
Reference r_i		**A Table t_i's Value v**	**ES**	**SBERT**
the accused is charged with **sexual touching** of his stepdaughter when she was between 10 and 15 years old.		*Section 151* – **Sexual interference**; Any person who, for a sexual purpose, directly or indirectly touches, with any part of the body or with any object, any part of the body of a child under the age of sixteen years	20	1
the accused pleaded guilty to the following charges : to have, on #DATE, in its possession 0,61 gram of cannabis		*Section 4* – **Possession of substance**. Except as authorized by the regulations, the **possession** of any **substance** listed in Appendix I, II or III is prohibited.	27	1

Table 3: Qualitative results of ES and SBERT fine-tuned model in the $h \nrightarrow t$ setting. We illustrate the ability of a fine-tuned ranking model to properly rank a particular value v in the table t_i given the hypothesis h_i (in the case of Gold annotations, r_i).

h	$\nrightarrow$	r	Rank of r_j	
Reference r_i		**Paraphrased Reference $r_j \vert d_{i,j} = 1.0$**	**ES**	**SBERT**
he **denies** any sexual act committed.		the accused states that **he has nothing to reproach himself for**.	N/A	3
PER **pleaded guilty** to **computer luring** of six teenage girls **between the ages of 13 and 17**.		a **guilty plea** [...] to **communicating by means of a computer** with x, a person **under the age of sixteen** [...]	N/A	4

Table 4: Qualitative results of both ES and SBERT fine-tuned in the $h \nrightarrow r$ setting. We show the ranking of a paraphrased reference r_j according to the hypothesis h_i (in the case of Gold annotations, r_i).

Gurevych (2019), a pre-trained version of SBERT applied without fine-tuning on *Plum2Text* struggles at ranking, especially in the $h \nrightarrow t$ setting (comprising a lot of paraphrasing). However a fine-tuned version of SBERT on *Plum2Text* yields strong performance, achieving a score of 0.851. There is inevitably a trade-off between using an untrained or trained version of our metric. From the results shown in Table 2 and in the context of paraphrases and synonyms, an embedding-based ranking model is definitely improving the evaluation.

4.1.1 Rankers' Behavior on Paraphrases and Synonyms

Supporting our claim that an embedding-based ranking model would be better at evaluating the generated hypotheses, especially when paraphrases characterize the dataset in hand, we qualitatively compare the ranking capabilities of the ES ranker and SBERT fine-tuned model.

In the $h \nrightarrow t$ setting, we extract references r that are a paraphrased version of a given table value v. As illustrated in Table 3[4], SBERT learned synonyms such as "sexual *contacts*" and "sexual *touching*". SBERT also learned that "possession of *substance*" is related to drugs like *cannabis* or *cocaïne*, and is thus able to properly rank the associated reference even though there are different types of "possessions" (e.g. child pornography, illegal firearms). As illustrated in the results, simply relying on word co-occurrences yields poor ranking in cases where synonyms are used and shows that an embedding-based ranking model is clearly providing a better performance.

In the $h \nrightarrow r$ setting, we analyze again the ranking behaviour of both rankers on paraphrased references. We can see in Table 4 that SBERT learned the various ways "pleading guilty" can be expressed (rows 1 and 2). It also learned that "communicating with people of 16 years and under using a computer" is similar to "computer luring of people between 13 and 17 years old". In most cases, we did not see the paraphrased reference among the top 200 results the ES ranker returned (N/A).

4.1.2 Motivation over a metric based on word overlap

To motivate the need to fine-tune rankers (hence the metric) for specific types of D2T datasets, we illustrate the performance of the recently introduced metric PARENT (Dhingra et al., 2019) on the gold annotations of *Plum2Text*, referred to as "Original" in Table 5. Without any surprises, the precision is 1.0 when $h = r$. The lower performance on the recall (and thus F1-Score) is due to paraphrasing, an inherent problem Dhingra et al. (2019) raised

[4]Note that all examples have been translated from French to facilitate the comprehension of the reader.

when they introduced their metric. To better illustrate this problem, we evaluate PARENT on a augmented version of *Plum2Text* i.e. for every pair r_i, r_j where $d_{i,j} = 1.0$ and $r_i \neq r_j$, we create a paraphrased example (t_i, r_j, r_i) where t_i is the table, r_i the hypothesis and r_j its associated reference. We can see in Table 5 that the results drop significantly due to the word overlap evaluation behavior, even though they should be similar to the Original Dataset.

Dataset	Precision	Recall	F1-Score
Original	1.0	0.565	0.673
Augmented	**0.355**	**0.24**	**0.191**

Table 5: Evaluation of PARENT on the Original and a Augmented version of *Plum2Text* (Garneau et al., 2021)

4.2 Experiments with WebNLG

In this section, we show that, even if our metric is well-suited for D2T dataset comprising textual utterance as table values, it generalizes to more common D2T settings as in WebNLG (Gardent et al., 2017). Shimorina et al. (2018) provide human evaluation on the different systems' outputs (listed in Figure 2, as well as on the gold annotations (`webnlg`). They considered three evaluation dimensions; Fluency, Grammar, and Semantic. In our case, we consider only the Semantic dimension.

First, we simply consider "fixing the metric", i.e. we apply our metric on the `webnlg` team's outputs (i.e. the gold annotations) and compare its values against the human evaluation[5] and the recently introduced metric by Dušek and Kasner (2020) (NLI). Results can be found in Figure 2. Intuitively, human evaluation should be very close to one (0.92)[6], and our trained metric achieves 0.88, which is close to human evaluation. The NLI metric average score is 0.73.

We then run the same experiment on a set of generators' outputs (Shimorina et al., 2018) in order to assess the capabilities of our proposed metric to discriminate amongst systems (i.e. teams). By looking at Figure 2, we can see that there is an agreement on the macro level between the human evaluation, our metric, and NLI in order to discriminate between the teams that performed well from the teams that did not. While the sample size is rather small (10 teams), the Pearson correlation score between human evaluation and our metric, human evaluation and NLI, are both 0.92 ($\rho < 0.005$). The average difference between human evaluation and our metric is 0.09 while human evaluation and NLI is 0.26. On the micro level, correlation scores show another story; human evaluation and our metric yield a Pearson correlation score of 0.47, human evaluation and NLI 0.59, our metric and NLI 0.43 ($\rho < 0.005$ in every cases). While there is a slight correlation between the different evaluation scheme, it seems that they do not always agree at the utterance level, contradicting one another in some cases. This point has already been raised by Dušek and Kasner (2020), suggesting that in some cases, the human evaluation is not accurate. However we decide to leave this specific analysis for future work.

4.2.1 Analysis of omissions and hallucinations

We further analyze the capacity of our metric to characterize omissions and hallucinations on the systems' outputs. To this end, we follow the methodology introduced in Section 3.4 and compute the *estimated* omission and hallucination rates w.r.t to the capacity of underlying rankers. Omission rate is the number of times an input value v was considered omitted by our metric, over the set of n input values. Hallucinations rate is the number of times a value v from the training set has been improperly ranked at the top n expected values[7]. We thus average the rates per system overall 223 examples for the WebNLG's test set.

In this experiment, we are interested in comparing Neural vs Non-Neural architectures, and see if our metric captures the implicit omission/hallucination behavior of neural generators. Results are displayed in Table 6. On the gold annotations, we obtained 0.41 and 0.37 omission and hallucination rates respectively. This is expected mostly because the underlying rankers are not perfect. While achieving 0.88 average precision on the gold annotations, this is an optimistic estimation of the ranking capabilities (see Section 3.4

[5]Human annotators used a three-point Likert scale (1 = Incorrect, 2 = Medium, 3 = Correct) and answers are averaged over multiple annotators. We normalized the scores between 0–1 for an easier comparison

[6]Out of the 224 human evaluation on the gold annotations, 38 have a score below or equal to 0.77.

[7]Rates are computed w.r.t the hypotheses produced by a given system. For example, Vietnam only produced 55 hypotheses given 223 input tables. We thus considered the input values of the 55 input tables for the calculation.

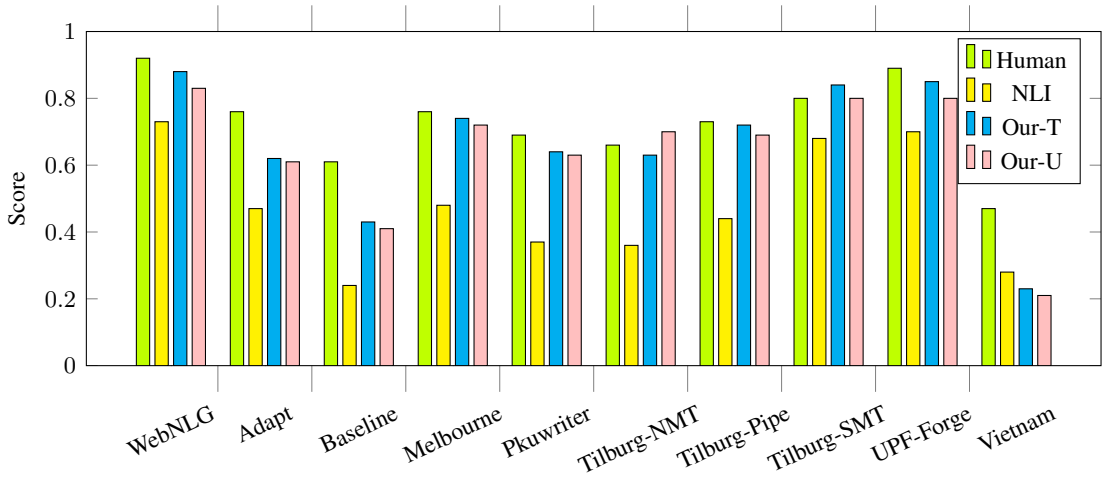

Figure 2: A comparison across human evaluation (semantic) on systems' outputs of Shimorina et al. (2018), the NLI metric proposed by Dušek and Kasner (2020), our metric trained and untrained. Our proposed metric, in every cases except one, is closer to human evaluation than the NLI metric.

for a discussion on this topic). Also, according to Dušek and Kasner (2020), there is some noise in the human evaluation.

Regarding the teams' statistics, we denote higher omission and hallucination rates for 3 out 5 neural systems. Interestingly, the estimated omission and hallucination rates of Adapt are quite high, while having a high semantic score from the human evaluation. On the contrary, Tilburg-NMT has low omission and hallucination rates while having a low human evaluation score. Melbourne has low omission and hallucination rates, which corroborates the fact that it was a good system. Non-Neural systems tend to be more stable on omissions and hallucinations, which is expected. In future experiments, we would like to compute the exact omission and hallucination rates of each team. While being a very time-consuming task, this evaluation will enable an in-depth analysis of omissions and hallucinations per system. This leads to the conclusion that, while being a first step at characterizing/quantifying omissions and hallucinations, more work has to be done towards this direction since it is a crucial evaluation aspect in D2T evaluation.

4.3 Is Fine-Tuning Worth the Shot?

In an era where deep learning models seem to be the norm, it is nonetheless legitimate to ask ourselves if training such a metric is worth the shot. In the case of WebNLG, where the data is extracted from Wikidata (Vrandečić and Krötzsch, 2014), a proxy of Wikipedia, and the underlying transformer models of the metric have been pre-trained on Wikipedia, we see a negligible gain from fine-tuning the metric. The lexical field is pretty much the same, and the reference text does not show many signs of paraphrasing.

	Team	Omission	Hallucination
	WebNLG	0.41	0.37
Neural	Adapt	**0.55**	**0.47**
	Baseline	**0.65**	**0.61**
	Melbourne	0.44	0.38
	Pkuwriter	**0.53**	**0.50**
	Tilburg-NMT	0.43	0.36
Non-Neural	Tilburg-pipe	0.45	0.38
	Tilburg-SMT	0.40	0.35
	UPF-Forge	0.41	0.35
	Vietnam	0.33	0.33

Table 6: Omission and Hallucination rates per team. We compare the omission and hallucination rates between WebNLG (the gold standard), Neural and Non-Neural architectures.

5 Conclusion

In this paper, we introduced a new trainable automation evaluation metric relying on ranking models which is specific to the D2T setting. To the best of our knowledge, this is the first metric that naturally quantifies omissions and hallucinations of neural textual generators that can also handle paraphrases. Characterizing omissions and hallucinations are of important matter in a sensitive area such as the legal domain. This lack of characterization is often a

blocker to the use of recent neural generator models in the legal field. We hope that this metric will also promote the use of recent advances in neural textual generation in sensitive domains such as the medical field.

In our future works, we would like to use our metric to guide the decoding steps of neural D2T generators in order to produce faithful textual descriptions. As previously mentioned, the metric that we proposed is well suited to semantically and factually assess a generator's performance in cases where the tables can be associated with multiple references, and the tables' values contain textual utterances. We wish to generalize the way our method ranks table values through relevance matching (Guo et al., 2016). This would be highly desirable in cases where table values are only up to a few tokens.

Acknowledgments

We thank the reviewers for their thoughtful comments and suggestions. Nicolas is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

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.

David Beauchemin, Nicolas Garneau, Eve Gaumond, Pierre-Luc Déziel, Richard Khoury, and Luc Lamontagne. 2020. Generating intelligible plumitifs descriptions: Use case application with ethical considerations. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 15–21, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. 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.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2021. Plum2text: A french plumitifs–descriptions data-to-text dataset for natural language generation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, Sao Paulo, Brazil. International Association for Artificial Intelligence and Law.

Lorenzo Gatti, Chris van der Lee, and Mariët Theune. 2018. Template-based multilingual football reports generation using Wikidata as a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 183–188, Tilburg University, The Netherlands. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman

Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander 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*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 55–64, New York, NY, USA. Association for Computing Machinery.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 145–150, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Kathleen R. McKeown. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, USA.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. 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.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6908–6915.

Yevgeniy Puzikov and Iryna Gurevych. 2018. E2E NLG challenge: Neural models vs. templates. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 463–471, Tilburg University, The Netherlands. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A survey of evaluation metrics used for nlg systems. *ArXiv*, abs/2008.12009.

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.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2018. WebNLG Challenge: Human Evaluation Results. Technical report, Loria & Inria Grand Est.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. *AAAI Conference on Artificial Intelligence*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

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.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.