Data Mining and Knowledge Discovery (2022) 36:318–354
https://doi.org/10.1007/s10618-021-00801-4

Controlling hallucinations at word level in data-to-text generation

Clement Rebuffel[1] · **Marco Roberti**[2] · **Laure Soulier**[1] · **Geoffrey Scoutheeten**[3] · **Rossella Cancelliere**[2] · **Patrick Gallinari**[1,4]

Received: 2 February 2021 / Accepted: 17 September 2021 / Published online: 22 October 2021

Abstract

Data-to-Text Generation (DTG) is a subfield of Natural Language Generation aiming at transcribing structured data in natural language descriptions. The field has been recently boosted by the use of neural-based generators which exhibit on one side great syntactic skills without the need of hand-crafted pipelines; on the other side, the quality of the generated text reflects the quality of the training data, which in realistic settings only offer imperfectly aligned structure-text pairs. Consequently, state-of-art neural models include misleading statements –usually called hallucinations—in their outputs. The control of this phenomenon is today a major challenge for DTG, and is the problem addressed in the paper. Previous work deal with this issue at the instance level: using an alignment score for each table-reference pair. In contrast, we propose a finer-grained approach, arguing that hallucinations should rather be treated at the word level. Specifically, we propose a Multi-Branch Decoder which is able to leverage word-level labels to learn the relevant parts of each training instance. These labels are obtained following a simple and efficient scoring procedure based on co-occurrence analysis and dependency parsing. Extensive evaluations, via automated metrics and human judgment on the standard WikiBio benchmark, show the accuracy of our alignment labels and the effectiveness of the proposed Multi-Branch Decoder.

Clement Rebuffel and Marco Roberti have contributed equally to this work.

Responsible Editor: Annalisa Appice, Sergio Escalera, Jose A. Gamez, and Heike Trautman.

✉ Marco Roberti
m.roberti@unito.it

Clement Rebuffel
clement.rebuffel@lip6.fr

[1] CNRS, LIP6, Sorbonne Université, 75005 Paris, France

[2] University of Turin, Turin, Italy

[3] BNP Paribas, Paris, France

[4] Criteo AI Lab, Paris, France

Our model is able to reduce and control hallucinations, while keeping fluency and coherence in generated texts. Further experiments on a degraded version of ToTTo show that our model could be successfully used on very noisy settings.

Keywords Data-to-text generation · Hallucinations · Controlled text generation

1 Introduction

Data-to-Text Generation (DTG) is the subfield of Computational Linguistics and Natural Language Generation (NLG) that is concerned with transcribing structured data into natural language descriptions, or, said otherwise, transcribing machine understandable information into a human understandable description (Gatt and Krahmer 2018). DTG objectives includes *coverage*, i.e. all the required information should be present in the text, and *adequacy*, i.e. the text should not contain information that is not covered by the input data. DTG is a domain distinct from other NLG task (e.g. machine translation (Wiseman et al. 2017), text summarization (Kryscinski et al. 2019)) with its own challenges (Wiseman et al. 2017), starting with the nature of inputs (Reiter and Dale 1997; Narayan and Gardent 2020). Such inputs include and are not limited to: databases of records, spreadsheets, knowledge bases, sensor readings. As an example, Fig. 1 shows an instance of the WikiBio dataset, i.e. a data table containing information about Kian Emadi, paired with its corresponding natural language description found on Wikipedia.

Early approaches to DTG relied on static rules hand-crafted by experts, in which content selection (what to say) and surface realization (how to say it) are typically two separate tasks (Reiter and Dale 1997; Ferreira et al. 2019). In recent years, neural models have blurred this distinction: various approaches showed that both content selection and surface realization can be learned in an end-to-end, data-driven fashion (Mei et al. 2016; Liu et al. 2019a; Roberti et al. 2019; Puduppully et al. 2019a). Based on the now-standard encoder-decoder architecture, with attention and copy mechanisms (Bahdanau et al. 2015; See et al. 2017), neural methods for DTG are able to produce fluent text conditioned on structured data in a number of domains (Lebret et al. 2016; Wiseman et al. 2017; Puduppully et al. 2019b), without relying on heavy manual work from field experts.

Such advances have gone hand in hand with the introduction of larger and more complex benchmarks. In particular, surface-realization abilities have been well studied on hand-crafted datasets such as E2E (Novikova et al. 2017b) and WebNLG (Gardent et al. 2017), while content-selection has been addressed by automatically constructed dataset such as WikiBio (Lebret et al. 2016) or RotoWire (Wiseman et al. 2017). These large corpora are often constructed from internet sources, which, while easy to access and aggregate, do not consist of perfectly aligned source-target pairs (Perez-Beltrachini and Gardent 2017; Dhingra et al. 2019). Consequently, model outputs are often subject to over-generation: misaligned fragments from training instances, namely *divergences*, can induce similarly misaligned outputs during inference, the so-called *hallucinations*.

KEY	VALUE
name	kian emadi
fullname	kian emadi-coffin
currentteam	retired
discipline	track
role	rider
ridertype	sprinter
proyears	2012-present
proteams	sky track cycling

Ref.: kian emadi (born 29 july 1992) is a british track cyclist .

Fig. 1 An example of a WikiBio instance, composed by an input table and its (partially aligned) description

In this paper, we specifically address the issue of hallucinations, which is currently regarded as a major issue in DTG (Narayan and Gardent 2020). Indeed, experimental surveys show that real-life end-users of DTG systems care more about reliability than about readability (Reiter and Belz 2009), as unfaithful texts can potentially mislead decision makers, with dire consequences. Hallucinations-reduction methods such as the one presented here have applications in a broad range of tasks requiring high reliability, like news reports (Leppänen et al. 2017), in which hallucinations may give rise to *fake news*, or summaries of patient information in clinical contexts (Portet et al. 2009; Banaee et al. 2013).

When corpora include a mild amount of noise, as in handcrafted ones (e.g. E2E, WebNLG), dataset regularization techniques (Nie et al. 2019; Dusek et al. 2019) or hand crafted rules (Juraska et al. 2018) can help to reduce hallucinations. Unfortunately, these techniques are not suited to more realistic and noisier datasets, as for instance WikiBio (Lebret et al. 2016) or RotoWire (Wiseman et al. 2017). On these benchmarks, several techniques have been proposed, such as reconstruction loss terms (Wiseman et al. 2017; Wang 2019; Lin et al. 2020) or Reinforcement Learning (RL) based methods (Perez-Beltrachini and Lapata 2018; Liu et al. 2019b; Rebuffel et al. 2020). These approaches suffer however from different issues: (1) the reconstruction loss relies on the hypothesis of one-to-one alignment between source and target which does not fit with content selection in DTG; (2) RL-trained models are based on instance-level rewards (e.g. BLEU (Papineni et al. 2002), PARENT (Dhingra et al. 2019)) which can lead to a loss of signal because divergences occur at the word level. In practice, parts of the target sentence express source attributes (in Fig. 1 name and occupation fields are correctly realized), while others diverge (the birthday and nationality of Kian Emadi are not supported by the source table).

Interestingly, one can view DTG models as Controlled Text Generation (CTG) ones focused on controlling content, as most CTG techniques condition the generation on several key-value pairs of *control factors* (e.g. tone, tense, length) (Dong et al. 2017; Hu et al. 2017; Ficler and Goldberg 2017). Recently, Filippova (2020) explicitly introduced CTG to DTG by leveraging an *hallucination score* simply attached as an additional attribute which reflects the amount of noise in the instance. As an example, the table from Fig. 1 can be augmented with an additional line (`hallucination_score`, `80%`)[1]. However, this approach requires a strict alignment at the instance-level, namely between control factors and target text. A first attempt towards word-level approaches is proposed by Perez-Beltrachini and Lapata (2018) (also *PB&L* in the following). They design word-level alignment labels, denoting the correspondence between the text and the input table, to bootstrap DTG systems. However, they incorporate these labels into a sentence-level RL-reward, which ultimately leads to a loss of this finer-grained signal.

In this paper, we go further in this direction with a DTG model by fully leveraging word-level alignment labels with a CTG perspective. We propose an original approach in which the word-level is integrated at all phases:

- we propose a **word-level labeling procedure** (Sect. 3), based on co-occurrences and sentence structure through dependency parsing. This mitigates the failure of strict word-matching procedure, while still producing relevant labels in complex settings.
- we introduce a **weighted multi-branch neural decoder**(Sect. 4), guided by the proposed alignment labels, acting as word-level control factors. During training, the model is able to distinguish between aligned and unaligned words and learns to generate accurate descriptions without being misled by un-factual reference information. Furthermore, our multi-branch weighting approach enables control at inference time.

We carry out extensive experiments on WikiBio, to evaluate both our labeling procedure and our decoder (Sect. 6). We also test our framework on ToTTo (Parikh et al. 2020), in which models are trained with noisy reference texts, and evaluated on references reviewed and cleaned by human annotators to ensure accuracy. Evaluations are based on a range of automated metrics as well as human judgments, and show increased performances regarding hallucinations reduction, while preserving fluency.

Importantly, our approach makes training neural models on noisy datasets possible, without the need to handcraft instances. This work shows the benefit of word-level techniques, which leverage the entire training set, instead of removing problematic training samples, which may form the great majority of the available data.

2 Related work

Handling hallucinations in noisy datasets The use of Deep Learning based methods to solve DTG tasks has led to sudden improvements in state of the art performances (Lebret et al. 2016; Wiseman et al. 2017; Liu et al. 2018; Puduppully et al. 2019a). As a key aspect in determining a model's performance is the quality of training data, several large corpora have been introduced to train and evaluate models' abilities on diverse tasks. E2E (Novikova et al. 2017b) evaluates surface realization, i.e. the strict transcription of input attributes into natural language; RotoWire (Wiseman et al. 2017) pairs statistics of basketball games with their journalistic descriptions, while WikiBio

[1] The reader may disagree with such a strong hallucination score. Indeed, while the birthdate and nationality are clearly divergences, the rest of the sentence is correct. This illustrates the complexity of handling divergences in complex datasets, where alignment cannot be framed as a simple word-matching task.

(Lebret et al. 2016) maps a Wikipedia info-box with the first paragraph of its associated article. Contrary to E2E, the latter datasets are not limited to surface realization. They were not constructed by human annotators, but rather created from Internet sources, and consist of loosely aligned table-reference pairs: in WikiBio, almost two thirds of the training instances contain divergences (Dhingra et al. 2019), and no instance has a 1-to-1 source-target alignment (Perez-Beltrachini and Gardent 2017).

On datasets with a moderate amount of noise, such as E2E, data pre-processing has proven effective for reducing hallucinations. Indeed, rule-based (Dusek et al. 2019) or neural-based methods (Nie et al. 2019) have been proposed, specifically with table regularization techniques, where attributes are added or removed to re-align table and target description. Several successful attempts have also been made in automatically learning alignments between the source tables and reference texts, benefiting from the regularity of the examples (Juraska et al. 2018; Shen et al. 2020; Gehrmann et al. 2018). For instance, Juraska et al. (2018) leverage templating and hand-crafted rules to re-rank the top outputs of a model decoding via beam search; Gehrmann et al. (2018) also leverage the possible templating formats of E2E's reference texts, and train an ensemble of decoders where each decoder is associated to one template; and Kasner and Dusek (2020) produce template-based lexicalizations and improve them via a *sentence fusion* model. The previous techniques are not applicable in more complex, general settings. The work of Dusek et al. (2019) hints at this direction, as authors found that neural models trained on E2E were principally prone to omissions rather than hallucinations. In this direction, Shen et al. (2020) were able to obtain good results at increasing the coverage of neural outputs, by constraining the decoder to focus its attention exclusively on each table cell sequentially until the whole table was realized. On more complex datasets (e.g. WikiBio), a wide range of methods has been explored to deal with factualness such as loss design, either with a reconstruction term (Wiseman et al. 2017; Wang 2019) or with RL-based methods (Perez-Beltrachini and Lapata 2018; Liu et al. 2019b; Rebuffel et al. 2020). Similarly to the coverage constraints, a reconstruction loss has proven only marginally efficient in these settings, as it contradicts the content selection task (Wang 2019), and needs to be well calibrated using expert insight in order to bring improvements. Regarding RL, Perez-Beltrachini and Lapata (2018) build an instance-level reward which sums up word-level scores; Liu et al. (2019b) propose a reward based on document frequency to favor words from the source table more than rare words; and Rebuffel et al. (2020) train a network with a variant of PARENT (Dhingra et al. 2019) using self-critical RL. Note that data regularization techniques have also been proposed (Thomson et al. 2020; Wang 2019), but these methods require heavy manual work and expert insights, and are not readily transposable from one domain to another.

From CTG to controlling hallucinations Controlled Text Generation (CTG) is concerned with constraining a language model's output during inference on a number of desired attributes, or *control factors*, such as the identity of the speaker in a dialog setting (Li et al. 2016), the politeness of the generated text or the text length in machine-translation (Sennrich et al. 2016; Kikuchi et al. 2016), or the tense in generated movie reviews (Hu et al. 2017). Earlier attempts at neural CTG can even be seen as direct instances of DTG as it is currently defined: models are trained to generate text conditioned on attributes of interest, where attributes are key-value pairs. For instance, in the movie review domain, Ficler and Goldberg (2017) proposed an expertly crafted dataset, where sentences are strictly aligned with control factors, being either content or linguistic style aspects (e.g. tone, length).

In the context of dealing with hallucinations in DTG, Filippova (2020) recently proposed a similar framework, by augmenting source tables with an additional attribute that reflects the degree of hallucinated content in the associated target description. During inference, this attribute acts as an *hallucination handle* used to produce more or less factual text. As mentioned in Sect. 1, we argue that a unique value can not accurately represent the correspondence between a table and its description, due to the phrase-based nature of divergences.

Based on the literature review, the lack of model control can be evidenced when loss modification methods are used (Wang 2019; Liu et al. 2019a; Rebuffel et al. 2020), although these approaches can be efficient and transposed from one domain to another. On the other hand, while CTG deals with control and enables choosing the defining features of generated texts (Filippova 2020), standard approaches rely on instance-level control factors that do not fit with hallucinations, which rather appear due to divergences at the word level. Our approach aims at gathering the merits of both trends of models and is guided by previous statements highlighting that word-level is primary in hallucination control. More particularly, our model differs from previous ones in several aspects:

(1) Contrasting with data-driven approaches (i.e. dataset regularization) which are costly in expert time, and loss-driven approaches (i.e. reconstruction or RL losses) which often do not take into account key subtasks of DTG (content-selection, world-level correspondences), we propose a multi-branch modeling procedure which allows the controllability of the hallucination factor in DTG. This multi-branch model can be integrated seamlessly in current approaches, allowing to keep peculiarities of existing DTG models, while deferring hallucination management to a parallel decoding branch.

(2) Unlike previous CTG approaches (Li et al. 2016; Sennrich et al. 2016; Ficler and Goldberg 2017; Filippova 2020) which propose instance-level control factors, the control of the hallucination factor is performed at the word-level to enable finer-grained signal to be sent to the model.

Our model is composed of two main components: (1) a word-level alignment labeling mechanism, which makes the correspondence between the input table and the text explicit, and (2) a multi-branch decoder guided by these alignment labels. The branches separately integrate co-dependent control factors (namely content, hallucination and fluency). We describe these components in Sects. 3 and 4, respectively.

3 Word-level alignment labels

We consider a DTG task, in which the corpus $\mathcal{C}$ is composed of a set of entity-description pairs, (e, y). A *single-entity table* e is a variable-sized set of T_e key-value pairs $x{:=}(k, v)$. A *description* $y{:=}y_{1:T_y}$ is a sequence of T_y tokens representing the

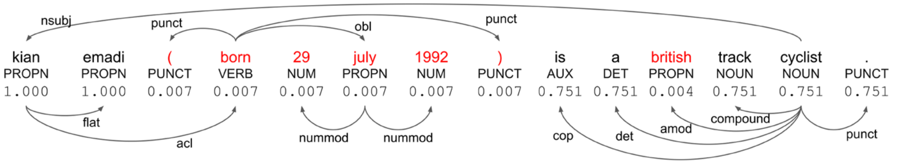


Fig. 2 The reference sentence of the example shown in Fig. 1. Every token is associated to its Part-of-Speech tag and hallucination score s_t. Words in red denote $s_t < \tau$. The dependency parsing is represented by labeled arrows that flow from parents to children. Important words are *kian*, *emadi*, *29*, *july*, *1992*, *british*, *track*, and *cyclist*

natural language description of the entity; we refer to the tokens spanning from indices t to t' of a description y as $y_{t:t'}$. A description is made of *statements*, defined as text spans expressing one single idea ("Appendix A" presents in detail the statement partitioning procedure). We refer to the first index of a statement as t_i, so that $y_{t_i:t_{i+1}-1}$ is the i^{th} statement itself. Figure 1 shows a WikiBio entity made by 8 key-value pairs together with its associated description.

First, we aim at labeling each word from a description, depending on the presence of a correspondence with its associated table. We call such labels *alignment labels*. We drive the word-level labeling procedure on two intuitive constraints: (1) important words (names, adjectives and numbers) should be labeled depending on their alignment with the table, and (2) words from the same statement should have the same label.

With this in mind, the *alignment label* for the t^{th} token y_t is a binary label: $l_t := \mathbb{1}_{\{s_t > \tau\}}$ where s_t refers to the *alignment score* between y_t and the table, and τ is set experimentally (see Sect. 5.3). The *alignment score* s_t acts as a normalized measure of correspondence between a token y_t and the table e:

$$s_t := norm(\max_{x \in e} align(y_t, x), \quad y) \tag{1}$$

where the function *align* estimates the alignment between token y_t and a key-value pair x from the input table e, and *norm* is a normalization function based on the dependency structure of the description y. Figure 2 illustrates our approach: under each word we show its word alignment score, and words are colored in red if this score is lower than τ, denoting an alignment label equal to 0. Below, we describe these functions ("Appendix A" contains reproducibility details).

Co-occurrence-based alignment function (**align**$(\cdot, \mathbf{x})$). This function assigns to important words a score in the interval [0, 1] proportional to their co-occurrence count (a proxy for alignment) with the key-value pair from the input table. If the word y_t appears in the key-value pair $x := (k, v)$, $align(y_t, x)$ outputs 1; otherwise, the output is obtained scaling the number of occurrences $co_{y_t,x}$ between y_t and x through the dataset:

$$align(y_t, x) := \begin{cases} 1 & \text{if } \ y_t \in x \\ a \cdot (co_{y_t,x} - m)^2 & \text{if } \ m \leq co_{y_t,x} \leq M \\ 0 & \text{if } \ 0 \leq co_{y_t,x} \leq m \end{cases} \tag{2}$$

where M is the maximum number of word co-occurrences in the dataset vocabulary and the row x, m is a threshold value, and $a := \frac{1}{(M-m)^2}$.

Score normalization ($\mathbf{norm}(\cdot, \mathbf{y})$). According to the already stated assumption (2)—words inside the same statement should have the same score – , we first split the sentence y into statements $y_{t_i:t_{i+1}-1}$, via dependency parsing and its rule-based conversion to constituency trees (Han et al. 2000; Xia and Palmer 2001; Hwa et al. 2005; Borensztajn et al. 2009). Given a word y_t associated to the score s_t and belonging to statement $y_{t_i:t_{i+1}-1}$, its normalized score corresponds to the average score of all important words in this statement:

$$norm(s_t, y) = \frac{1}{t_{i+1} - t_i} \sum_{j=t_i}^{t_{i+1}-1} s_j \tag{3}$$

This in-statement average depends on both the specific word and its context, leading to coherent hallucination scores which can be thresholded without affecting the syntactical sentence structure, as shown in Fig. 2.

4 Multi-branch architecture

The proposed Multi-Branch Decoder (MBD) architecture aims at separating targeted co-dependent factors during generation. We build upon the standard DTG architecture, an encoder-decoder with attention and copy mechanism, which we modify by duplicating the decoder module into three distinct parallel modules. Each control factor (i.e. content, hallucination or fluency) is modeled via a single decoding module, also called branch, whose output representation can be weighted according to its desired importance. At training time, weights change depending on the word currently being decoded, inducing the desired specialization of each branch. During inference, weights are manually set, according to the desired trade-off between information reliability, sentence diversity and global fluency. Text generation is thus controllable, and consistent with the control factors.

Figure 3 illustrates a training step over the sentence "*Giuseppe Mariani was an Italian art director*", in which *Italian* is a divergent statement (i.e. is not supported by the source table). While decoding factual words, the weight associated to the content (resp. hallucination) branch is set to 0.5 (resp. 0) while during the decoding of *Italian*, the weight associated to the content (resp. hallucination) branch is set to 0 (resp. 0.5). Note that the weight associated to the fluency branch is always set to 0.5, as fluency does not depend on factualness.

The decoding modules' actual architecture may vary, as we framed the MBD model from a high level perspective. Therefore, all types of decoder can be used, such as Recurrent Neural Networks (RNNs) (Rumelhart et al. 1986), Transformers (Vaswani et al. 2017), and Convolutional Neural Networks (Gehring et al. 2017). The framework can be generalized to different merging strategies as well, such as late fusion, in which the final distributions are merged, instead of the presented early fusion, which works at the decoder states level.

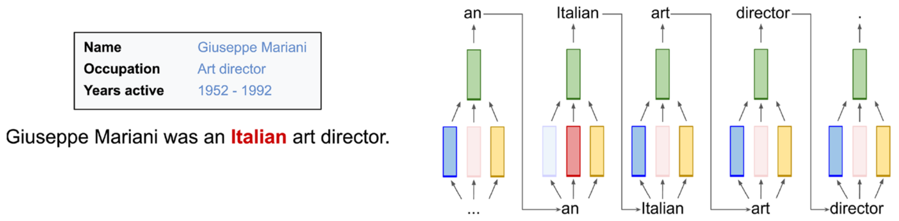


Fig. 3 Our proposed decoder with three branches associated to content (in blue—left), hallucination (in red—middle) and fluency (in yellow—right). Semi-transparent branches are assigned the weight 0

In this paper, experiments are carried out on RNN-based decoders, weighting their hidden states. Sect. 4.1 presents the standard DTG encoder-decoder architecture; Sect. 4.2 shows how it can be extended to MBD, together with its peculiarities and the underlying objectives and assumptions.

4.1 Standard DTG architecture

Neural DTG approaches typically use an encoder-decoder architecture (Wiseman et al. 2017) in which (1) the encoder relies on a RNN to encode each element of the source table into a fixed-size latent representation h_j (elements of the input table are first embedded into T_e N-dimensional vectors, and then fed sequentially to the RNN (Wiseman et al. 2017)), and (2) the decoder generates a textual description y using a RNN augmented with attention and copy mechanisms (See et al. 2017). Words are generated in an auto-regressive way. The decoder's RNN updates its hidden state d_t as:

$$d_t := \mathrm{RNN}(d_{t-1}, [y_{t-1}, c_t]) \tag{4}$$

where y_{t-1} is the previous word and c_t is the context vector obtained through the attention mechanism. Finally, a word is drawn from the distribution computed via a copy mechanism (See et al. 2017).

4.2 Controlling hallucinations via a multi-branch model

Our objective is to enrich the decoder in order to be able to tune the content/hallucination ratio during generation, aiming at enabling generation of hallucination-free text when needed. Our key assumption is that the decoder's generation is conditioned by three co-dependent factors:

- *Content factor* constrains the generation to realize only the information included in the input;
- *Hallucinating factor* favors lexically richer and more diverse text, but may lead to hallucinations not grounded by the input;
- *Fluency factor*[2] conditions the generated sentences toward global syntactic correctness, regardless of the relevance.

[2] Wiseman et al. (2018) showed that the explicit modeling of a fluency latent factor improves performance.

Based on this assumption, we propose a multi-branch encoder-decoder network, whose branches are constrained on the above factors at word-level, as illustrated in Fig. 3. Our network has a single encoder and $F = 3$ distinct decoding RNNs, noted RNN^f respectively, one for each factor. During each decoding step, the previously decoded word y_{t-1} is fed to all RNNs, and a final decoder state d_t is computed using a weighted sum of all the corresponding hidden states,

$$d_t^f := \text{RNN}^f(d_{t-1}^f, [y_{t-1}, c_t]) \tag{5}$$

$$d_t := \sum_{f=1}^{F} \omega_t^f d_t^f \tag{6}$$

where d_t^f and ω_t^f are respectively the hidden state and the weight of the f^{th} RNN at time t.

Weights are used to constrain the decoder branches to the desired control factors ($\omega_t^0, \omega_t^1, \omega_t^2$ for the content, hallucination and fluency factors respectively) and sum to one.

During training, their values are dynamically set depending on the *alignment label* $l_t \in \{0, 1\}$ of the target token y_t (see Sect. 5.3). While a number of mappings can be used to set the weights given the alignment label, early experiments have shown that better results were achieved when using a binary switch for each factor, i.e. activating/deactivating each branch, as shown in Fig. 3 (note that fluency should not depend on content and therefore its associated branch is always active).

During inference, the weights of the decoder's branches are set manually by a user, according to the desired trade-off between information reliability, sentence diversity and global fluency. Text generation is then controllable and consistent with the control factors.

5 Experimental setup

5.1 Datasets

We evaluated the model on two representative large size datasets. Both have been collected automatically and present a significant amount of table-text divergences for training. Both datasets involve content selection and surface realization, and represent a relatively realistic setting.

WikiBio (Lebret et al. 2016) contains 728, 321 tables, automatically paired with the first sentence of the corresponding Wikipedia English article. Reference text's average length is 26 words, and tables have on average 12 key-value pairs. We use the original data partition: 80% for the train set, and 10% for validation and test sets. This dataset has been automatically built from the Internet; concerning divergences, 62% of the references mention extra information not grounded by the table (Dhingra et al. 2019).

ToTTo (Parikh et al. 2020) contains 120, 761 training examples, and 7, 700 validation and test examples. For a given Wikipedia page, an example is built up by pairing its summary table and a candidate sentence, selected across the whole page via simple similarity heuristics. Such a sentence may accordingly realize whichever table cells, making content selection arbitrary; furthermore, its lexical form may strongly depend on the original context, because of pronouns or anaphoras. Divergences are of course present as well. Those issues have been addressed by Parikh et al. (2020) by (1) *highlighting* the input cells realized by the output, and (2) removing divergences and making the sentence self-contained (e.g. replacing pronouns with their invoked noun or noun phrase). Figure 6 exemplifies the difference between noisy and clean ToTTo sentences. In our experiments, we limit the input to the highlighted cells and use the original, noisy sentence as output. Noisy texts' average length is 17.4 words, and 3.55 table cells are highlighted, on average.

5.2 Baselines

We assess the accuracy and relevance of our alignment labels against the ones proposed by Perez-Beltrachini and Lapata (2018), which is, to the best of our knowledge, the only work proposing such a fine-grained alignment labeling.

To evaluate our Multi-Branch Decoder (*MBD*), we consider five baselines:

- *stnd* (See et al. 2017), a LSTM-based encoder-decoder model with attention and copy mechanisms. This is the standard sequence-to-sequence recurrent architecture.
- *stnd_filtered*, the previous model trained on a filtered version of the training set: tokens deemed hallucinated according to their hallucination scores, are removed from target sentences.
- *hsmm* (Wiseman et al. 2018), an encoder-decoder model with a multi-branch decoder. The branches are not constrained by explicit control factors. This is used as a baseline to show that the multi-branch architecture by itself does not guarantee the absence of hallucinations.
- *hier* (Liu et al. 2019a), a hierarchical sequence-to-sequence model, with a coarse-to-fine attention mechanism to better fit the *attribute-value* structure of the tables. This model is trained with three auxiliary tasks to capture more accurate semantic representations of the tables.
- hal_{WO} (Filippova 2020), a *stnd*-like model trained by augmenting each source table with an additional attribute (*hallucination ratio*, *value*).

We ran our own implementations of *stnd*, *stnd_filtered* and hal_{WO}. Authors of *hier* and *hsmm* models kindly provided us their WikiBio's test set outputs. The metrics described in Sect. 5.4 were directly applied on them.

5.3 Implementation details

During training of our multi-branch decoder the fluency branch is always active ($\omega_t^2 = 0.5$) while the content and hallucination branches are alternatively activated, depending on the alignment label l_t: $\omega_t^0 = 0.5$ (content factor) and $\omega_t^1 = 0$ (hallucination factor) when $l_t = 1$, and conversely. The threshold τ used to obtain l_t is set to 0.4 using human tuning to optimize for highest accuracy.[3] All hyperparameters were tuned in order to optimize the validation PARENT F-measure (Dhingra et al. 2019). In particular, we use the [0.4 0.1 0.5] weight combination during inference. See Sect. 6.2 for a discussion about weight combinations and "Appendix B" for other implementation details.[4]

5.4 Metrics

To evaluate our model, we carried out (1) an automatic analysis and (2) a human evaluation for a qualitative analysis of generated sentences.

For the automatic analysis, we use five metrics:

- BLEU (Papineni et al. 2002) is a length-penalized precision score over n-grams, $n \in [\![1, 4]\!]$, optionally improved with a smoothing technique (Chen and Cherry 2014). Despite being the standard choice, recent findings show that it correlates poorly with human evaluation, especially on the sentence level (Novikova et al. 2017a; Reiter 2018), and that it is a proxy for sentence grammar and fluency aspects rather than semantics (Dhingra et al. 2019).
- PARENT (Dhingra et al. 2019) computes smoothed n-gram precision and recall over both the reference and the input table. It is explicitly designed for DTG tasks, and its F-measure shows "the highest correlation with humans across a range of settings with divergent references in WikiBio." (Dhingra et al. 2019)
- The *hallucination rate* computes the percentage of tokens labeled as hallucinations (Sect. 3).
- The average generated sentence length in number of words.
- The classic readability Flesch index (Flesch 1962), which is based on words per sentence and syllables per word, and is still used as a standard benchmark (Kosmajac and Keselj 2019; Smeuninx et al. 2020; Stajner and Hulpus 2020; Stajner et al. 2020).

Finally, we perform qualitative evaluations of the results obtained on WikiBIO and ToTTo, following the best practices outlined by van der Lee et al. (2019). Our human annotators are from several countries across Europe, between 20 and 55 years old and proficient in English. They have been assigned two different tasks: (i) hallucination labeling, i.e. the selection of sentence pieces which include incorrect information, and (ii) sentence analysis, i.e. evaluating different realizations of the same table according to their fluency, factualness and coverage. Scores are presented as a 3-level Likert scale for Fluency (*Fluent*, *Mostly fluent*, or *Not fluent*) and Factualness (likewise), while coverage is the number of cells from the table that have been realized in the description.

To avoid all bias, annotators are shown a randomly selected table at a time, together with its corresponding descriptions, both from the dataset and the models that are being evaluated. Sentences are presented each time in a different order. Following Tian et al. (2019), we first tasked three expert annotators to annotate a pilot batch

[3] Note that accuracy is not heavily impacted by different choices of τ. We report in "Appendix B" the respective accuracy scores of our proposed automated labels for different values of τ.

[4] Code is given to reviewers and will be available upon acceptance.

Table 1 Performances of hallucination scores on WikiBio test set, w.r.t. human-designated labels (upper table) and *MBD* trained with different labeling procedures (lower table). Our model always significantly overpasses *PB&L* (T-test with $p < 0.005$)

Labels	Accuracy	Precision	Recall	F-measure
PB&L	46.9%	21.3%	49.2%	29.7%
Ours	**87.5%**	**80.6%**	**59.8%**	**68.7%**

Labels	BLEU	PARENT		
		Precision	Recall	F-measure
PB&L	32.15%	76.91%	39.28%	48.75%
Ours	**40.51%**	**77.71%**	**45.01%**	**54.57%**

of 50 sentences. Once confirmed that Inter-Annotator Agreement was approx. 75% (a similar finding to Tian et al. (2019)), we asked 16 annotators to annotate a bigger sample of 300 instances (where each instance consists of one table and four associated outputs), as Liu et al. (2019a).[5]

6 Results

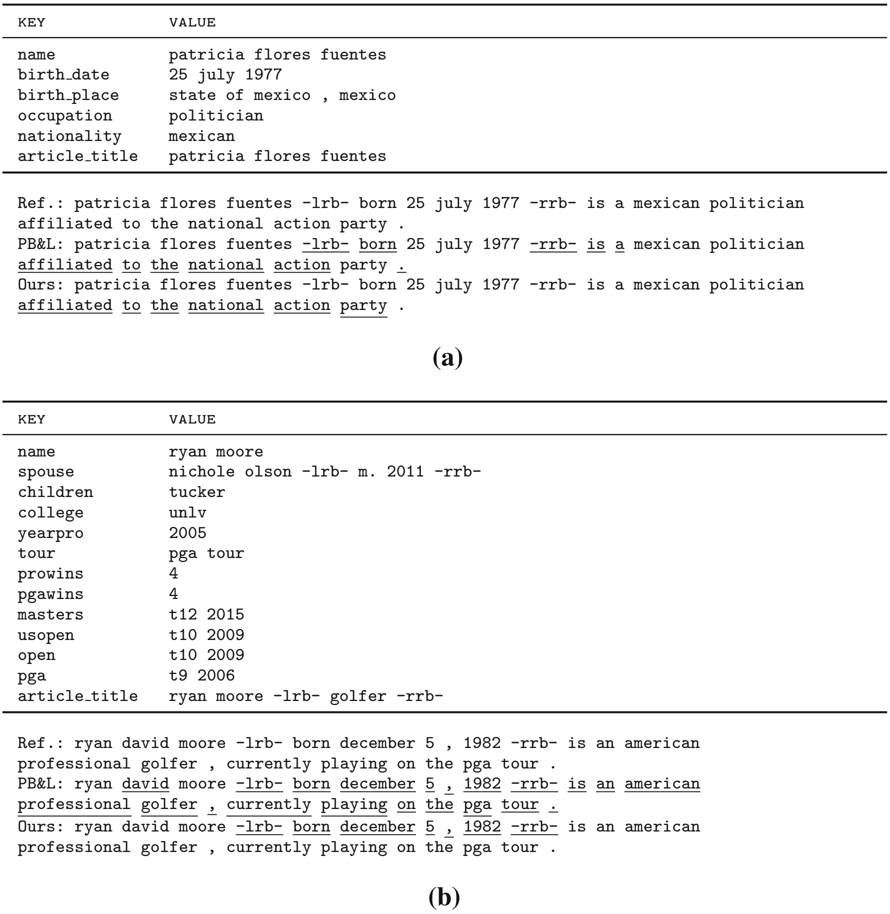
We perform an extensive evaluation of our scoring procedure and multi-branch architecture on the WikiBio dataset: we evaluate—the quality of the proposed alignment labels, both intrinsically using human judgment and extrinsically by means of the DTG downstream task and—the performance of our model with respect to the baselines. Additionally, we assess the applicability of our framework on the more noisy ToTTo benchmark, which represents a harder challenge for today's DTG models.

6.1 Validation of alignment labels

To assess the effectiveness of our alignment labels (Sect. 3), we first compare the alignment labels against human judgment, and then explore their impact on a DTG task. As a baseline for comparison we report performances of *PB&L*.
Intrinsic performance Table 1 (top) compares the labeling performance of our method and *PB&L* against human judgment. Our scoring procedure significantly improves over *PB&L*: the latter only achieves 46.9% accuracy and 29.7% F-measure, against 87.5% and 68.7% respectively for our proposed procedure. Perez-Beltrachini and Lapata (2018) report a F-measure of 36%, a discrepancy that can be explained by the difference between the evaluation procedures: *PB&L* evaluate on 132 sentences, several of which can be tied to the same table, whereas we explicitly chose to evaluate on 300 sentences all from different tables in order to minimize correlation.

We remark that beyond F-measure, the precision of *PB&L*'s scoring procedure is at 21.3% compared to 80.6% for ours, and recall stands at 49.2% against 59.8%. We argue that selecting a negative instance at random for training their classifier leads the network to incoherently label words, without apparent justification. See Fig. 4 for two examples of this phenomenon; and "Appendix D" for other comparisons. In

[5] An eyesight of our platform is available in "Appendix C".



KEY	VALUE
name	patricia flores fuentes
birth_date	25 july 1977
birth_place	state of mexico , mexico
occupation	politician
nationality	mexican
article_title	patricia flores fuentes

Ref.: patricia flores fuentes -lrb- born 25 july 1977 -rrb- is a mexican politician affiliated to the national action party .
PB&L: patricia flores fuentes -lrb- born 25 july 1977 -rrb- is a mexican politician affiliated to the national action party .
Ours: patricia flores fuentes -lrb- born 25 july 1977 -rrb- is a mexican politician affiliated to the national action party .

(a)

KEY	VALUE
name	ryan moore
spouse	nichole olson -lrb- m. 2011 -rrb-
children	tucker
college	unlv
yearpro	2005
tour	pga tour
prowins	4
pgawins	4
masters	t12 2015
usopen	t10 2009
open	t10 2009
pga	t9 2006
article_title	ryan moore -lrb- golfer -rrb-

Ref.: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .
PB&L: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .
Ours: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .

(b)

Fig. 4 WikiBio instances' hallucinated words according either to our scoring procedure or to the method proposed by Perez-Beltrachini and Lapata (2018). *PB&L* labels word incoherently (**a**), and sometimes the whole reference text (**b**). In comparison, our approach leads to a fluent breakdown of the sentences in hallucinated/factual statements

contrast, our method is able to detect hallucinated statements inside a sentence, without incorrectly labeling the whole sentence as hallucinated.

Impact on a DTG downstream task Additionally, we assess the difference of both scoring procedures using their impact on the WikiBio DTG task. Specifically, Table 1 (bottom) shows the results of training *MBD* using either *PB&L*'s or our labels. We observe significant improvements, especially in BLEU and PARENT-recall (40.5% vs 32.2% and 45% vs 39.3%), showing that our labeling procedure is more helpful at retaining information from training instances (the system better picks up what humans picked-up, ultimately resulting in better BLEU and recall).

Table 2 Comparison results on WikiBio. ↑ (resp. ↓) means higher (resp. lower) is better. “Gold” refers to the gold standard, i.e. the reference texts included in the dataset. Best values are bolded

Model	BLEU↑	PARENT↑			Halluc. rate↓	Mean sent. length	Flesch↓
		Precision	Recall	F-measure			
Gold	–	–	–	–	23.82%	19.20	**53.80%**
stnd	41.77%	79.75%	45.02%	55.28%	4.20%	13.80	58.90%
stnd_filtered	34.66%	**80.90%**	42.48%	53.27%	**0.74%**	12.00	62.10%
hsmm	35.17%	71.72%	39.84%	48.32%	7.98%	14.80	58.60%
hier	**45.14%**	75.09%	46.02%	54.65%	10.10%	16.80	56.20%
$\mathtt{hal}_{\mathtt{WO}}$	36.50%	79.50%	40.50%	51.70%	–	–	–
MBD	41.56%	79.00%	**46.40%**	**56.16%**	1.43%	14.60	58.80%

6.2 Automatic system evaluation

Comparison with SOTA systems Table 2 shows the performances of our model and all baselines according to the metrics of Sect. 5.4. Two qualitative examples are presented in Fig. 5 and more are available in "Appendix D".

First of all, reducing hallucinations is reached with success, as highlighted by the hallucination rate (1.43% vs. 4.20% for a standard encoder-decoder and 10.10% for the best SOTA model on BLEU). The only model which gets a lower hallucination rate (0.74%, corroborated by its PARENT-precision of 80.9%), *stnd_filtered*, achieves such a result at a high cost. As can be seen in Fig. 5 where its output is factual but cut short, its sentences are the shortest and the most naive in terms of the Flesch readability index, which is also reflected by a lower BLEU score. The high PARENT precision—mostly due to the shortness of the outputs—is counterbalanced by a low recall: the F-measure indicates the overall lack of competitiveness of this trade-off. This shows that the naive approach of simply filtering training instances is not the appropriate solution for hallucination reduction. This echoes (Filippova 2020) who trained a vanilla network on the cleanest 20% of the data and found that predictions are more precise than those of a model trained on 100% but that PARENT-recall and BLEU scores are low.

At the other extreme, the best model in terms of BLEU, *hier*, falls short regarding precision, suggesting that often the generated text is not matched in the input table; this issue is also reflected by the highest hallucination rate of all models (10.10%). A reason could be the introduction of their auxiliary training tasks which often drive the decoder to excess in mimicking human behavior. While BLEU score improves, overall factualness of outputs decreases, showing that the model picks up domain lingo (how to formulate ideas) but not domain insight (which ideas to formulate) (see Fig. 5). This is in line with (Reiter 2018; Filippova 2020) who argue that BLEU is an inappropriate metric for generation tasks other than machine translation.

The analysis of *hsmm*, and especially of its relatively weak performance both in terms of BLEU and PARENT, highlights the insufficiency of the multi-branch architecture by itself. This reinforces the need of the additional hallucinations supervision provided by our labeling procedure.

Finally, in the comparisons with hal_{WO}, we can see that while it achieves one of the highest performances in term of precision (79.5%), this comes at the cost of the lowest recall (40.5%) of all models and thus poor F-measure. This confirms our hypothesis that, while effective at producing mostly factual content, modeling hallucination only as a fixed value for a whole instance is detrimental to the content generation procedure. Finer-grain annotations are required, as shown by our model recall (46.4%), coupled with a robust precision (79.0%).

Weight impact on decoding As we deal with a CTG system, we can guide our network at inference to generate sentences following desired attributes. The impact of different weight combinations is explored in Table 3. In particular, we can see that changing weights in favor of the hallucination factor (top five lines) leads to decreases in both precision and recall (from 80.37% to 57.88% and 44.96% 4.82% respectively). We also observe that strongly relying on the hallucinating branch dramatically impacts

name	zack lee
birth_name	zack lee jowono
nationality	indonesian
occupation	actor , boxer , model
birth_date	15 august 1984
birth_place	liverpool , merseyside , england , uk
years_active	2003 -- present
parents	hendra and ayu jowono
spouse	nafa urbach (2007 -- present)
article_title	zack lee

Gold	zack lee (born 15 august 1984) is an indonesian actor , model and boxer of british descent .
stnd	zack lee jowono (born 15 august 1984) is an indonesian actor and model .
stnd_filtered	zack lee (born zack lee jowono ; 15 august 1984) is an indonesian actor .
hsmm	zack lee jowono (born 15 august 1984) is an indonesian actor who has appeared in tamil films .
hier	zack lee jowono (born 15 august 1984) , better known by his stage name zack lee , is an indonesian actor , model and model .
MBD[.4, .1, .5]	zack lee (born zack lee jowono ; 15 august 1984) is an indonesian actor , boxer and model .

(a)

name	wayne r. dynes
birth_date	23 august 1934
occupation	professor , historian , and encyclopedist
article_title	wayne r. dynes

Gold	wayne r. dynes (born august 23 , 1934) is an american art historian , encyclopedist , and bibliographer .
stnd	wayne r. dynes (born august 23 , 1934) is an american historian and encyclopedist .
stnd_filtered	wayne r. dynes is a professor .
hsmm	wayne r. dynes (born august 23 , 1934) is an american historian , historian and encyclopedist .
hier	wayne r. dynes (born august 23 , 1934) is an american professor of history at the university of texas at austin .
MBD[.4, .1, .5]	wayne r. dynes (born august 23 , 1934) is an american professor , historian , and encyclopedist .

(b)

Fig. 5 Qualitative examples of our model and baselines on the WikiBio test set. Note that: (1) *gold* references may contain divergences; (2) *stnd* and *hsmm* seem to perform well superficially, but often hallucinate; (3) *stnd_filtered* doesn't hallucinate but struggles with fluency; (4) *hier* overgenerate "human-sounding" statements, that lacks factualness; (5) *MBD* sticks to the fact contained by the table, in concise and fluent sentences

performances ([0.0 0.5 0.5] obtains near 0 BLEU and F-measure), as it is never fed with complete, coherent sentences during training. However, some performance can still be restored via the fluency branch: [0.0 0.1 0.9] performs at 15.51% BLEU and 36.88% F-measure.

It is interesting to note that the relaxation of the strict constraint on the content factor in favor of the hallucination factor, ([0.4 0.1 0.5] → [0.5 0.0 0.5]) obtains better performances (56.16% vs 55.29% F-measure). This highlights that strictly constraining on content yields sensibly more factual outputs (79% vs 80.37% precision), at the cost of constraining the model's generation creativity (46.40% vs 44.96% recall). The [0.4 0.1 0.5] variant has more "freedom of speech" and sticks more faithfully to domain lingo (recall and BLEU), without compromising too much in terms of content.

Table 3 Performances of *MBD* on WikiBio validation set, with various weight settings. Weights' order is (*content*, *hallucination*, *fluency*)

Weights	BLEU$^{\uparrow}$	PARENT$^{\uparrow}$		
		Precision	Recall	F-measure
0.5 0.0 0.5	38.90%	80.37%	44.96%	55.29%
0.4 0.1 0.5	41.56%	79.00%	46.40%	56.16%
0.3 0.2 0.5	42.68%	72.99%	45.81%	53.74%
0.2 0.3 0.5	22.64%	53.92%	32.96%	36.55%
0.1 0.4 0.5	2.03%	57.88%	4.82%	6.79%
0.0 0.5 0.5	0.32%	85.01%	1.02%	1.78%
0.0 0.4 0.6	1.07%	62.71%	2.47%	3.66%
0.0 0.3 0.7	2.81%	42.86%	6.15%	7.94%
0.0 0.2 0.8	7.30%	41.78%	16.58%	18.68%
0.0 0.1 0.9	15.51%	56.93%	32.85%	36.88%

Table 4 Results of the human evaluation on WikiBio. Best values are bolded

Model	Fluency	Factualness	Coverage
Gold	98.7%	32.0%	**4.47**
`stnd_filtered`	93.5%	**86.1%**	4.07
`hier`	97.4%	55.0%	4.45
`MBD`	**99.6%**	76.6%	4.46

Fluency reports the sum of "fluent" and "mostly fluent", as "mostly fluent" often comes from misplaced punctuation and doesn't really impact readability. However, Factualness reports only the count of "factual", as "mostly factual" sentences contain hallucinations and cannot be considered "factual"

6.3 Human evaluation

To measure subtleties which are not captured by automatic metrics, we report in Table 4 human ratings of our model, two baselines and the gold. These baselines have been selected because they showcase interesting behaviors on automatic metrics: *hier* obtains the best BLEU score but a poor precision, and *stnd_filtered* gets the best precision but poor BLEU, length and Flesch index.

First, coherently with (Dhingra et al. 2019), we found that around two thirds of gold references contain divergences from their associated tables. Such data also confirm our analysis on the *stnd_filtered* baseline: it's training on truncated sentences lead to an unquestionable ability to avoid hallucinations, while dramatically impacting both its fluency and coverage, leading to less desired outputs overall, despite the high PARENT-precision score.

The comparison between *hier* and *MBD* shows that both approaches lead to similar coverage, with *MBD* obtaining significantly better performances in terms of factualness. We also highlight that *MBD* is evaluated as being the most fluent one, even better than the reference (which can be explained by the imperfect pre-processing done by Lebret et al. (2016)).

Table 5 Comparison results on ToTTo. $^{\uparrow}$ (resp. $^{\downarrow}$) means higher (resp. lower) is better. In human evaluation for Fluency, reported are for "Fluent" and "Mostly Fluent", with only "Fluent" in parentheses. Same for Factualness. Best values are bolded

Model	BLEU$^{\uparrow}$	PARENT$^{\uparrow}$			Human evaluation		
		Precision	Recall	F-measure	Fluency$^{\uparrow}$	Factualness$^{\uparrow}$	Coverage
Gold(noisy)	–	–	–	–	97.1% (97.1)	91.2% (79.4)	3.618
stnd	**21.27%**	56.60%	25.16%	**29.71%**	55.9% (26.5)	53.0% (20.6)	2.824
stnd_filtered	19.48%	56.69%	22.31%	27.18%	29.4% (8.8)	70.6% (50.0)	2.706
hal_{WO}	17.06%	**77.64%**	22.65%	29.38%	61.7% (38.2)	61.8% (32.4)	2.725
MBD	18.35%	50.44%	**25.25%**	28.25%	**91.2%** (50.0)	**85.3%** (55.9)	**3.613**

6.4 ToTTo: a considerably noisy setting

The ToTTo dataset is used in the following experiments to explore models' robustness to the impact of extreme noise during training. As stated in Sect. 5.1, we use as inputs only the *highlighted* cells, as content selection is arbitrary (i.e. the cells were chosen depending on the target sentence, and not vice versa). On the other hand, we use as targets the noisy references, which may contain both divergences and lexical issues. This setting is particularly challenging and is more effective in recreating a representational, hallucination-prone real-life context than WikiBio. Other datasets (Novikova et al. 2017b; Gardent et al. 2017; Wen et al. 2015) available in literature are too similar to WikiBio concerning their goals and challenges, and are therefore less interesting in this context.

Table 5 reports the performances of *stnd*, *stnd_filtered*, hal_{WO} and *MBD* with regards to automatic metrics and human evaluation. Compared to their respective performances on WikiBio, all models show significantly decreased scores. They struggle at generating syntactically correct sentences but, at the same time, they have still learned to leverage their copy mechanism and to stick to the input. This behavior is illustrated in both examples of Fig. 6. In particular, hal_{WO}'s high PARENT-precision score (77.64%) seems to be due to its tendency to blindly copy input data without framing them in a sentence structure, as its low BLEU and PARENT-recall scores suggests (17.06% and 22.65%). These lower scores are good indicators that the ToTTo task, as framed in this paper, is difficult. Following the same evaluation protocol than for WikiBio, we report human ratings of different models, also included in Table 5.

MBD's factualness is judged favorably, with 55.9% hallucination-free texts, and up to 85.3% texts with a single error at most. In contrast, hal_{WO} stands at 32.4% and 61.8% for error-free texts and single-error texts respectively. Interestingly, *stnd_filtered* obtains the second best performance (70.6% texts with a single error).

Fluency scores are also meaningful: hal_{WO} and *MBD* respectively obtain 61.7% and 91.2%. Word-based filtering is not suitable for noisy datasets, as shown by *stnd_filtered*'s worse fluency score, 29.4%.

As for coverage performances, our model *MBD* obtains the maximum coverage score 3.613, surpassing all baselines by at least 0.789 slots (the second best coverage score is obtained by *stnd* at 2.824), and getting very close to the Gold value (which

page_title	Huge (TV series)
section_title	Episodes
Original_air_date	June 28 2010
U.S._viewers_(millions)	2.53

Gold (clean)	The TV series , Huge , premiered on June 28 , 2010 with 2.53 million viewers.
Gold (noisy)	The series premiered on June 28 , 2010 at 9 p.m. with 2.53 million viewers .
stnd	On June 28 , 2010 , it was watched by 2.53 million viewers .
stnd_filtered	was watched by 2.53 on June 28 , 2010 .
hal$_{\text{WO}}$	June 28 , 2010 : Huge million viewers .
MBD[.4, .1, .5]	Huge 's first episode , aired on June 28 , 2010 , was watched by 2.53 million .

(a)

page_title	LM317
section_title	Specification
Parameter	Output voltage range
Value	1.25 - 37

Gold (clean)	LM317 produces a voltage of 1.25 V .
Gold (noisy)	Internally the device has a bandgap voltage reference which produces a stable reference voltage of Vref= 1.25 V followed by a feedback-stabilized amplifier with a relatively high output current capacity .
stnd	The Output is a Output range of 1.25 – 37 .
stnd_filtered	range from 1.25 to 37 .
hal$_{\text{WO}}$	Output voltage range 1.25 – 37 – 37 .
MBD[.4, .1, .5]	The Output 's range is approximately 1.25 .

(b)

Fig. 6 Qualitative examples of *MBD* and hal_{WO} on ToTTo. hal_{WO}'s poor generation quality is not detected by discrete metrics. In contrast, *MBD* generates fluent and naively factual sentences. Note that *stnd* and *stnd_filtered* have the same behavior as on WikiBio: the former produces fluent but nonsensical text; the latter generates very un-fluent, but factual, text

stands at 3.618). These performances, and qualitative examples of Fig. 6, suggest that *stnd_filtered* and hal_{WO} try to reduce hallucinations at the cost of missing some input slot, while *MBD* effectively balances both goals.

The analysis of Factualness, Fluency and Coverage can be enhanced using qualitative error analysis on randomly sampled generated texts (we report two such examples in Fig. 6). In particular, we want to highlight the following considerations:

- As most training examples are very noisy, sentence-level models fail at learning from them. *stnd_filtered* has been trained on factual statements only, at the cost of using mostly incomplete sentences during training. On both examples of Fig. 6, it generated truncated sentences, missing their subjects. Its relatively high Factualness and low Fluency scores indicate that it did not learn to produce diverging outputs, nor complete sentences. Differently, hal_{WO} generates incorrectly ordered sequences of words extracted from the table (Fig. 6a), or repetitions (Fig. 6b). The low number of training instances containing the input pair *(hallucination ratio, 0)* does not allow to learn what a non-hallucinated sentence actually consists in.
- In contrast, our proposed finer-grained approach proves helpful in this setting, as shown by the human evaluation: sentences generated by *MBD* are more fluent and more factual. The multi-branch design enables the model to leverage the most of each training instance, leading to better performances overall.

- Finally, we acknowledge that despite over-performing other models, *MBD* obtains only 55.9% of *factual* sentences. For instance, in Fig. 6b, our model does not understand that a range consists of two numbers. The difficulty of current models to learn on very noisy and diverse datasets shows that there is still room for improvement in hallucination reduction in DTG.

7 Conclusion

We proposed a Multi-Branch decoder, able to leverage word-level alignment labels in order to produce factual and coherent outputs. Our proposed labeling procedure is more accurate than previous work, and outputs from our model are estimated, by automatic metrics and human judgment alike, more fluent, factual, and relevant. We obtain state-of-the-art performances on WikiBio for PARENT F-measure, and show that our approach is promising in the context of a noisier setting.

We designed our alignment procedure to be general and easily reproducible on any DTG dataset. One strength of our approach is that co-occurrences and dependency parsing can be used intuitively to extract more information from the tables than a naive word matching procedure. However, in the context of tables mainly including numbers (e.g., RotoWire), the effectiveness of the co-occurrence analysis is not guaranteed. A future work will be to improve upon the co-occurrence analysis to generalize to tables which contain less semantic inputs. For instance, the labeling procedure of Perez-Beltrachini and Lapata (2018) might be revised so that adverse instances are not selected randomly, which we hypothesize would result in more relevant labels.

Finally, experiments on ToTTo outline the narrow exposure to language of current models when used on very noisy datasets. Our model has shown interesting properties through the human evaluation but is still perfectible. Recently introduced large pretrained language models, which have seen significantly more varied texts, may attenuate this problem. In this direction, adapting the work of (Chen et al. 2020; Kale and Rastogi 2020) to our model could bring improvements to the results presented in this paper.

Fundings Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. We would like to thank the H2020 project AI4EU (825619) and the ANR JCJC SESAMS (ANR-18-CE23-0001) for supporting this work. This research has been partially carried on in the context of the Visiting Professor Program of the Italian Istituto Nazionale di Alta Matematica (INdAM).

Declarations

Conflict of interest No conflict of interest.

Code availability Code is available at https://github.com/KaijuML/dtt-multi-branch.

A Alignment labels reproducibility

We consider as *important words*, i.e. nouns, adjectives or numbers, those which are Part-of-Speech tagged as `NUM`, `ADJ`, `NOUN` and `PROPN`.

In order to apply the score normalization function $norm(\cdot, y)$, we separate sentences y into statements $y_{t_i:t_{i+1}-1}$. To do so, we identify the set of introductory dependency relation labels[6], following previous work on rule-based systems for the conversion of dependency relations trees to constituency trees (Han et al. 2000; Xia and Palmer 2001; Hwa et al. 2005; Borensztajn et al. 2009).

Our segmentation algorithm considers every leaf token in the dependency tree, and seeks its nearest ancestor which is the root of a statement.

Two heuristics enforce the score normalization: (i) conjunctions and commas next to hallucinated tokens acquires these lasts' hallucination scores, and (ii) paired parentheses and quotes acquire the minimum inner tokens' hallucination score.

Part-of-Speech tagging has been done using the HuggingFace's Transformers library (Wolf et al. 2019) to fine-tune a BERT model (Devlin et al. 2019) on the UD English ParTUT dataset (Sanguinetti and Bosco 2015); Stanza (Qi et al. 2020) has been exploited for dependency parsing.

B Implementation details

Our system is implemented in Python 3.8[7] and PyTorch 1.4.0.[8] In particular, our multi-branch architecture is developed, trained and tested as an OpenNMT (Klein et al. 2017) model. Sentence lengths and Flesch index (Flesch 1962) are computed using the standard `style` Unix command.

Differently to Perez-Beltrachini and Lapata (2018), we did not adapt the original WikiBio dataset[9] in any manner: as we work on the model side, we fairly preserve the dataset's noisiness.

Word-level alignment labels are computed setting $m = 5$, following Mikolov et al. (2013). As stated in Sect. 5.3, the threshold τ's value is optimized for highest accuracy via human tuning: Table 6 shows accuracy scores of our proposed automated labels for different values of τ.

We share the vocabulary between input and output, limiting its size to 20000 tokens. Hyperparameters were tuned using performances on the development set: Table 7

[6] `acl`, `advcl`, `amod`, `appos`, `ccomp`, `conj`, `csubj`, `iobj`, `list`, `nmod`, `nsubj`, `obj`, `orphan`, `parataxis`, `reparandum`, `vocative`, `xcomp`; every dependency relation is documented in the https://universaldependencies.org/u/dep/index.htmlUniversal Dependencies website.

[7] http://www.python.org.

[8] http://www.pytorch.org.

[9] https://github.com/DavidGrangier/wikipedia-biography-dataset.

Table 6 Accuracy scores of our proposed word-level automated labels for different values of the threshold τ

Threshold	Accuracy	F-measure	Precision	Recall
0.0	70.2%	56.8%	42.2%	86.7%
0.4	**86.0%**	**70.6%**	67.0%	74.6%
0.8	85.8%	62.8%	77.3%	52.9%

Table 7 The performances of our model on the WikiBio validation set

Model	BLEU	PARENT		
		Precision	Recall	F-measure
`MBD[.4, .1, .5]`	42.50%	79.26%	46.09%	55.95%

reports the performances of our best performing *MBD* on the development set. Our encoder consist of a 600-dimensional embedding layer followed by a 2-layered bidirectional LSTM network with hidden states sized 600. We use the *general* attention mechanism with input feeding (Luong et al. 2015) and the same copy mechanism as See et al. (2017). Each branch of the multi-branch decoder is a 2-layered LSTM network with hidden states sized 600 as well.

Training is performed using the Adam algorithm (Kingma and Ba 2015) with learning rate $\eta = 10^{-3}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is decayed with a factor of 0.5 every 10000 steps, starting from the 5000th one. We used minibatches of size 64 and regularized via clipping the gradient norm to 5 and using a dropout rate of 0.3. We used beam search during inference, with a beam size of 10.

All experiments were performed on a single NVIDIA Titan XP GPU. Number of parameters and training times are shown in Table 8. Same model's differences between WikiBio and ToTTo are justified by the different datasets' number of instances and input vocabulary sizes.

C Annotation interface

The human annotation procedure is done via a web application specifically developed for this research. Figure 7a shows how the hallucination tagging user interface looked like in practice, while in Fig. 7b a typical sentence analysis page is shown.

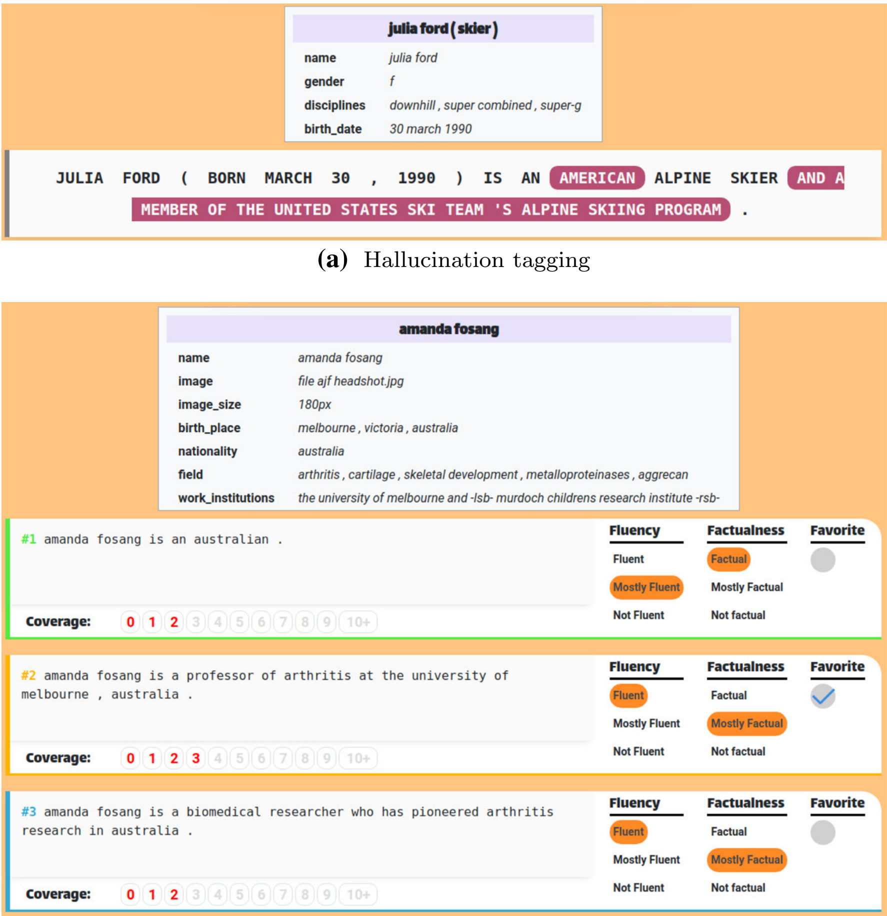
D Qualitative examples

Tables 9 and 10 show word-level labeling of WikiBio training examples. Underlined, red words are hallucinated according either to our scoring procedure or to the method proposed by Perez-Beltrachini and Lapata (2018).

In the subsequent tables, some WikiBio (Tables 11, 12, 13, 14, 15, 16, 17, 18, 19 20) and ToTTo (Tables 21, 22, 23) inputs are shown, coupled with the corresponding sentences, either as found in the dataset, or as generated by our models and baselines.

Table 8 Sizes and training times of the implemented models

Dataset	Model	Size [M]	Training time [h]
WikiBio	`stnd`	41	5
	`stnd_filtered`	41	5
	$\mathtt{hal}_{\mathtt{WO}}$	41	5
	`MBD`	55	10
ToTTo	`stnd`	62	4
	`stnd_filtered`	62	4
	$\mathtt{hal}_{\mathtt{WO}}$	62	4
	`MBD`	76	8



julia ford (skier)	
name	*julia ford*
gender	*f*
disciplines	*downhill , super combined , super-g*
birth_date	*30 march 1990*

JULIA FORD (BORN MARCH 30 , 1990) IS AN **AMERICAN** ALPINE SKIER **AND A MEMBER OF THE UNITED STATES SKI TEAM 'S ALPINE SKIING PROGRAM** .

(a) Hallucination tagging

amanda fosang	
name	*amanda fosang*
image	*file ajf headshot.jpg*
image_size	*180px*
birth_place	*melbourne , victoria , australia*
nationality	*australia*
field	*arthritis , cartilage , skeletal development , metalloproteinases , aggrecan*
work_institutions	*the university of melbourne and -lsb- murdoch childrens research institute -rsb-*

#1 amanda fosang is an australian .

Coverage: [✓ 0] [✓ 1] [✓ 2] 3 4 5 6 7 8 9 10+

Fluency: Fluent [✓ Mostly Fluent] Not Fluent
Factualness: [✓ Factual] Mostly Factual Not factual
Favorite: ○

#2 amanda fosang is a professor of arthritis at the university of melbourne , australia .

Coverage: [✓ 0] [✓ 1] [✓ 2] [✓ 3] 4 5 6 7 8 9 10+

Fluency: [✓ Fluent] Mostly Fluent Not Fluent
Factualness: Factual [✓ Mostly Factual] Not factual
Favorite: ✓

#3 amanda fosang is a biomedical researcher who has pioneered arthritis research in australia .

Coverage: [✓ 0] [✓ 1] [✓ 2] 3 4 5 6 7 8 9 10+

Fluency: [✓ Fluent] Mostly Fluent Not Fluent
Factualness: Factual [✓ Mostly Factual] Not factual
Favorite: ○

(b) Sentence analysis

Fig. 7 The human annotation tasks, as presented to the annotators

Table 9 Hallucinated words according either to our scoring procedure or to the method proposed by Perez-Beltrachini and Lapata (2018)

KEY	VALUE
name	susan blu
birth_name	susan maria blupka
birth_date	12 july 1948
birth_place	st paul , minnesota , u.s.
occupation	actress , director , casting director
yearsactive	1968 -- present
article_title	susan blu

Ref.: susan maria blu -lrb- born july 12 , 1948 -rrb- , sometimes credited as sue blu , is an american voice actress , voice director and casting director in american and canadian cinema and television .

PB&L: susan maria blu -lrb- born july 12 , 1948 -rrb- , sometimes credited as sue blu , is an american voice actress , voice director and casting director in american and canadian cinema and television .

Ours: susan maria blu -lrb- born july 12 , 1948 -rrb- , sometimes credited as sue blu , is an american voice actress , voice director and casting director in american and canadian cinema and television .

KEY	VALUE
name	patricia flores fuentes
birth_date	25 july 1977
birth_place	state of mexico , mexico
occupation	politician
nationality	mexican
article_title	patricia flores fuentes

Ref.: patricia flores fuentes -lrb- born 25 july 1977 -rrb- is a mexican politician affiliated to the national action party .

PB&L: patricia flores fuentes -lrb- born 25 july 1977 -rrb- is a mexican politician affiliated to the national action party .

Ours: patricia flores fuentes -lrb- born 25 july 1977 -rrb- is a mexican politician affiliated to the national action party .

KEY	VALUE
name	ate faber
birth_date	19 march 1894
birth_place	leeuwarden , netherlands
death_date	19 march 1962
death_place	zutphen , netherlands
sport	fencing
article_title	ate faber

Ref.: ate faber -lrb- 19 march 1894 -- 19 march 1962 -rrb- was a dutch fencer .

PB&L: ate faber -lrb- 19 march 1894 -- 19 march 1962 -rrb- was a dutch fencer .

Ours: ate faber -lrb- 19 march 1894 -- 19 march 1962 -rrb- was a dutch fencer .

Table 10 Hallucinated words according either to our scoring procedure or to the method proposed by Perez-Beltrachini and Lapata (2018)

KEY	VALUE
name	alex wilmot sitwell
birth_date	16 march 1961
birth_place	uk
occupation	president , europe and emerging markets -lrb- ex-asia -rrb- of bank of america merrill lynch
article_title	alex wilmot-sitwell

Ref.: alex wilmot-sitwell heads bank of america merrill lynch 's businesses across europe and emerging markets excluding asia .

PB&L: alex wilmot-sitwell heads bank of america merrill lynch 's businesses across europe and emerging markets excluding asia .

Ours: alex wilmot-sitwell heads bank of america merrill lynch 's businesses across europe and emerging markets excluding asia .

KEY	VALUE
name	ryan moore
spouse	nichole olson -lrb- m. 2011 -rrb-
children	tucker
college	unlv
yearpro	2005
tour	pga tour
prowins	4
pgawins	4
masters	t12 2015
usopen	t10 2009
open	t10 2009
pga	t9 2006
article_title	ryan moore -lrb- golfer -rrb-

Ref.: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .

PB&L: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .

Ours: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .

Table 11 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

title	prince of noër
name	prince frederick
image	prinsen af noer.jpg
image_size	200px
spouse	countess henriette of danneskjold-samsøe mary esther lee
issue	prince frederick, count of noer prince christian louise, princess michael vlangali-handjeri princess marie
house	house ofschleswig-holstein-sonderburg-augustenburg
father	frederick christian ii, duke of schleswig-holstein-sonderburg-augustenburg
mother	princess louise auguste of denmark
birth_date	23 august 1800
birth_place	kiel
death_date	2 july 1865
death_place	beirut
article_title	prince frederick of schleswig-holstein-sonderburg-augustenburg
Gold	prince frederick emil august of schleswig-holstein-sonderburg-augustenburg (kiel, 23 august 1800 – beirut, 2 july 1865), usually simply known by just his first name, frederick , “ prince of noër ” , was a prince of the house of schleswig-holstein-sonderburg-augustenburg and a cadet-line descendant of the danish royal house
stnd	prince frederick of schleswig-holstein-sonderburg-augustenburg (23 august 1800 – 2 july 1865) was a member of the house of schleswig-holstein-sonderburg-augustenburg
stnd_filtered	prince frederick of schleswig-holstein-sonderburg-augustenburg (23 august 1800 – 2 july 1865) was a german
hsmm	prince frederick of schleswig-holstein-sonderburg-augustenburg (23 august 1800 – 2 july 1865) was a danish noblewoman
hier	prince frederick of schleswig-holstein-sonderburg-augustenburg ( ) (23 august 1800 – 2 july 1865) was a german prince of the house of schleswig-holstein-sonderburg-augustenburg
MBD[.4, .1, .5]	prince frederick of schleswig-holstein-sonderburg-augustenburg (; 23 august 1800 – 2 july 1865) was the son of frederick christian ii, duke of schleswig-holstein-sonderburg-augustenburg and princess louise auguste of denmark

Table 12 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	godgory
background	group_or_band
origin	karlstad , sweden
genre	melodic death metal doom metal gothic metal
years_active	1992 – 2004
label	nuclear blast
current_members	matte andersson erik andersson
past_members	mikael dahlqvist stefan grundel henrik lindström fredric danielsson thomas heder
article_title	godgory
Gold	godgory was a swedish melodic death metal or death doom metal band that was formed august 1992 by erik andersson (drums) and matte andersson (vocals) of no relation
stnd	godgory was a melodic death metal band from karlstad, sweden
stnd_filtered	godgory was a swedish metal band
hsmm	godgory (10 august 2004 – 4 january 2010) was a melodic death metal band from karlstad, sweden
hier	godgory was a melodic death metal band from sweden, sweden
MBD[.4, .1, .5]	godgory was a swedish melodic death metal band from karlstad

Table 13 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	christy mihos
image	christy2006wp.jpeg
px|office	massachusetts turnpike commission member
term_start	july 1, 1999
term_end	july 1, 2004
successor	daniel grabauskas
birth_date	13 june 1949
birth_place	brockton, massachusetts
nationality	american
occupation	convenience store franchise owner
spouse	andrea mihos (1974–2013; divorce)
religion	greek orthodox
alma_mater	stonehill college
party	republican
article_title	christy mihos
Gold	christy peter mihos is an american politician and businessman from massachusetts
stnd	christy mihos (born june 13, 1949 in brockton, massachusetts) is a former republican member of the massachusetts house of representatives

Table 13 continued

stnd_filtered	christy mihos (born june 13, 1949 in brockton, massachusetts) is an american politician
hsmm	christy mihos (born june 13, 1949) is an american politician from the state of massachusetts
hier	christy mihos (born june 13, 1949) is an american politician who served as the massachusetts state representative from 1999 to 2004
MBD[.4, .1, .5]	christy mihos (born june 13, 1949, in brockton, massachusetts) is a former member of the massachusetts turnpike commission

Table 14 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	duncan eagleson
nationality	american
area	painter, writer, penciller, colorist, sculptor, designer
article_title	duncan eagleson
Gold	duncan eagleson is an american self-trained painter and former graffiti artist
stnd	duncan eagleson is an american painter, writer, and penciller
stnd_filtered	duncan eagleson is an american painter
hsmm	duncan eagleson (born 1975) is an american comic book painter and writer
hier	duncan eagleson is an american painter, illustrator, and designer
MBD[.4, .1, .5]	duncan eagleson is an american painter, writer, and sculptor

Table 15 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	gerald warner brace
imagesize	208px
birth_date	24 september 1901
birth_place	islip , long island , suffolk county , new york
death_date	20 july 1978
death_place	blue hill , maine
occupation	novelist , writer , educator , sailor , boat builder
nationality	american
genre	fiction , non-fiction
article_title	gerald warner brace
Gold	gerald warner brace (september 24, 1901 – july 20, 1978) was an american novelist , writer , educator , sailor and boat builder
stnd	gerald warner brace (september 24, 1901 – july 20, 1978) was an american novelist , writer , and boat builder
stnd_filtered	gerald warner brace (september 24 , 1901 – july 20, 1978) was an american novelist
hsmm	gerald warner brace (september 24, 1901 – july 20, 1978) was an american novelist and writer

Table 15 continued

`hier`	gerald warner brace (september 24, 1901 – july 20, 1978) was an american novelist , short story writer , educator , and sailor
`MBD[.4, .1, .5]`	gerald warner brace (september 24, 1901 – july 20, 1978) was an american author , educator , sailor , and boat builder

Table 16 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

`name`	`robert b. murrett`
`image`	`robertbmurrett.jpg`

`office`	`4th director of the national geospatial-intelligence agency`
	`director of the office of naval intelligence`

`president`	`george w. bush barack obama george w. bush`
`term_start`	`2006 2005`
`term_end`	`2010 2006`
`predecessor`	`james r. clapper richard b. porterfield`
`successor`	`letitia long tony l. cothron`
`alma_mater`	`university at buffalo georgetown university joint military intelligence college`
`branch`	`united states navy`
`rank`	`vice admiral 20px`
`article_title`	`robert b. murrett`
Gold	vice admiral robert b. murrett was the fourth director of the national geospatial-intelligence agency , from 7 july 2006 through july 2010
`stnd`	robert b. murrett is a retired vice admiral of the united states navy
`stnd_filtered`	robert b. murrett is the director of the national geospatial-intelligence agency
`hsmm`	robert b. “ bob ” murrett (born 1956) is an american naval officer and the director
`hier`	robert b. murrett is a retired vice admiral in the united states navy
`MBD[.4, .1, .5]`	robert b. murrett is a vice admiral in the united states navy

Table 17 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	rosane ferreira
image	deputada federal rosane ferreira.jpg
office	federal deputy for state of parná
term_start	1 february 2011
term_end	actual
president	dilma rousseff
order	federal deputy for the state of roraima
birth_date	31 july 1963
birth_place	clevelândia, parná, brazil
dead	alive
nationality	brazilian
party	green party (brazil)
article_title	rosane ferreira
Gold	rosane ferreira (cleusa rosane ribas ferreira , born clevelândia , paraná, july 31 , 1963) , is a nurse and a brazilian politician
stnd	rosane ferreira (born 31 july 1963 in clevelândia, parná) is a brazilian politician
stnd_filtered	rosane ferreira (born 31 july 1963) is a brazilian politician
hsmm	rosane ferreira (born july 31, 1963) is a brazilian politician and the federal deputy
hier	rosane ferreira (born 31 july 1963) is a brazilian politician and the current federal deputy for the state of roraima
MBD[.4, .1, .5]	rosane ferreira (born 31 july 1963 in clevelândia, parná, brazil) is a brazilian politician

Table 18 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	polina miller
gender	f
birth_date	23 november 1988
birth_place	saint petersburg, russian sfsr, soviet union
discipline	wag
article_title	polina miller
Gold	polina miller (born november 23, 1988 in saint petersburg) is a russian gymnast
stnd	polina miller (born november 23, 1988) is a russian artistic gymnast
stnd_filtered	polina miller (born november 23, 1988) is a
hsmm	polina miller (born 23 november 1988 in saint petersburg) is a russian artistic gymnast
hier	polina miller (born 23 november 1988) is a russian rhythmic gymnast
MBD[.4, .1, .5]	polina miller (born 23 november 1988 in saint petersburg, russian sfsr, soviet union) is a russian gymnast

Table 19 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	pat burke
irish	pádraig de búrca
sport	gaelic football
code	football
county	dublin
province	leinster
clposition	corner forward
club	kilmacud crokes
clubs	kilmacud crokes
counties	dublin
icprovince	1
birth_place	dublin , ireland
article_title	pat burke (gaelic footballer)
Gold	pat burke is an irish gaelic footballer who plays for dublin and kilmacud crokes
stnd	pat burke is a gaelic footballer from dublin , ireland
stnd_filtered	pat burke is a gaelic footballer for dublin
hsmm	pat burke (born in dublin) is a former irish gaelic footballer who played as a gaelic footballer
hier	pat burke is a former gaelic footballer for dublin
MBD[.4, .1, .5]	pat burke is a gaelic footballer from county dublin

Table 20 A WikiBio input table, coupled with the corresponding sentence and the models-generated outputs

name	odiakes
background	non_vocal_instrumentalist
birth_date	march 22
origin	tokyo , japan
instrument	keyboard , synthesizer
genre	j-pop , techno
occupation	composer
years_active	1998 - present
article_title	odiakes
Gold	odiakes (born march 22) is a japanese composer from tokyo, japan who has worked for a variety of visual novel companies
stnd	, better known by his stage name odiakes, is a japanese composer
stnd_filtered	odiakes is a japanese composer
hsmm	odiakes " odiakes " (born march 22) is a japanese composer
hier	composer (born march 22) is a japanese j-pop player
MBD[.4, .1, .5]	odiakes (born march 22 in tokyo, japan) is a japanese composer

Table 21 A ToTTo input table, coupled with the corresponding sentence and the models-generated outputs

Club	Istiklol
Season	2015
League	Tajik League
page_title	Parvizdzhon Umarbayev
section_title	Club
section_text	As of match played 29 July 2018
Gold	In 2015 , Umarbayev signed for Tajik League FC Istiklol
stnd	Umarbayev joined Tajik League side Istiklol in 2015
stnd_filtered	2015, he joined Istiklol in the Tajik League
hal_WO	Parvizdzhon joined Tajik League club Istiklol in 2015
MBD[.4, .1, .5]	Umarbayev signed with Istiklol ahead of the 2015 Tajik League season

Table 22 A ToTTo input table, coupled with the corresponding sentence and the models-generated outputs

Rank	5
Island	Hulhumeedhoo
page_title	List of islands of the Maldives
section_title	Islands by area size
section_text	This list ranks the top 10 islands of the Maldives by area Some islands in the Maldives, although geographically one island , are divided into two administrative islands (for example, Gan and Maandhoo in Laamu Atoll)
Gold	Hulhumeedhoo is the 5th largest island in Maldives
stnd	It has a area of Hulhumeedhoo km^2 (Islands sq mi)
stnd_filtered	is the fourth of the Maldives in Maldives
hal_WO	Hulhumeedhoo is the largest islands of the Maldives by area size
MBD[.4, .1, .5]	Hulhumeedhoo is the fifth largest island by area size

Table 23 A ToTTo input table, coupled with the corresponding sentence and the models-generated outputs

Single	24.7 (Twenty-Four Seven)
page_title	Singular (band)
section_title	2010
Gold	In 2010, Singular released its first single , " 24.7 (Twenty-Four Seven) "
stnd	The first single, 24.7 (Twenty-Four Seven) , was released in 2010
stnd_filtered	The band won the 24.7 (Twenty-Four Seven)
hal_WO	24.7 (Twenty-Four Seven)
MBD[.4, .1, .5]	Singular released their first album, 24.7 (Twenty-Four Seven)

References

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: ICLR

Banaee H, Ahmed MU, Loutfi A (2013) Towards NLG for physiological data monitoring with body area networks. In: ENLG

Borensztajn G, Zuidema WH, Bod R (2009) Children's grammars grow more abstract with age - evidence from an automatic procedure for identifying the productive units of language. TopiCS, 1:175-188

Chen B, Cherry C (2014) A systematic comparison of smoothing techniques for sentence-level BLEU. In: WMT@ACL

Chen Z, Eavani H, Chen W, Liu Y, Wang WY (2020) Few-shot NLG with pre-trained language model. In: ACL

Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL

Dhingra B, Faruqui M, Parikh A, Chang MW, Das D, Cohen W (2019) Handling divergent reference texts when evaluating table-to-text generation. In: ACL

Dong L, Huang S, Wei F, Lapata M, Zhou M, Xu K (2017) Learning to generate product reviews from attributes. In: EACL

Dusek O, Howcroft DM, Rieser V (2019) Semantic noise matters for neural natural language generation. In: INLG

Ferreira TC, van der Lee C, van Miltenburg E, Krahmer E (2019) Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In: EMNLP-IJCNLP

Ficler J, Goldberg Y (2017) Controlling linguistic style aspects in neural language generation. In: Workshop on Stylistic Variation @ ACL

Filippova K (2020) Controlled hallucinations: Learning to generate faithfully from noisy data. In: Findings of EMNLP

Flesch R (1962) The Art of Readable Writing

Gardent C, Shimorina A, Narayan S, Perez-Beltrachini L (2017) Creating training corpora for NLG micro-planners. In: ACL

Gatt A, Krahmer E (2018) Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. J Artif Intell Res 61:65–170

Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: ICML

Gehrmann S, Dai F, Elder H, Rush A (2018) End-to-end content and plan selection for data-to-text generation. In: INLG

Han C, Lavoie B, Palmer MS, Rambow O, Kittredge RI, Korelsky T, Kim N, Kim M (2000) Handling stuctural divergences and recovering dropped arguments in a korean/english machine translation system. In: AMTA

Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. In: ICML

Hwa R, Resnik P, Weinberg A, Cabezas CI, Kolak O (2005) Bootstrapping parsers via syntactic projection across parallel texts. Nat Lang Eng 11:311–325

Juraska J, Karagiannis P, Bowden KK, Walker MA (2018) A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In: NAACL-HLT

Kale M, Rastogi A (2020) Text-to-text pre-training for data-to-text tasks. In: INLG

Kasner Z, Dusek O (2020) Data-to-text generation with iterative text editing. In: INLG

Kikuchi Y, Neubig G, Sasano R, Takamura H, Okumura M (2016) Controlling output length in neural encoder-decoders. In: EMNLP

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: ICLR

Klein G, Kim Y, Deng Y, Senellart J, Rush AM (2017) OpenNMT: Open-source toolkit for neural machine translation. In: Proc. ACL

Kosmajac D, Keselj V (2019) Twitter user profiling: Bot and gender identification. In: CLEF

Kryscinski W, McCann B, Xiong C, Socher R (2019) Evaluating the factual consistency of abstractive text summarization, http://arxiv.org/abs/1910.12840

Lebret R, Grangier D, Auli M (2016) Neural text generation from structured data with application to the biography domain. In: EMNLP

Leppänen L, Munezero M, Granroth-Wilding M, Toivonen H (2017) Data-driven news generation for automated journalism. In: INLG

Li J, Galley M, Brockett C, Spithourakis GP, Gao J, Dolan B (2016) A persona-based neural conversation model. In: ACL

Lin S, Wang W, Yang Z, Liang X, Xu FF, Xing EP, Hu Z (2020) Record-to-text generation with style imitation. In: EMNLP

Liu T, Luo F, Xia Q, Ma S, Chang B, Sui Z (2019a) Hierarchical encoder with auxiliary supervision for neural table-to-text generation: Learning better representation for tables. In: AAAI

Liu T, Luo F, Yang P, Wu W, Chang B, Sui Z (2019b) Towards comprehensive description generation from factual attribute-value tables. In: ACLs

Liu T, Wang K, Sha L, Chang B, Sui Z (2018) Table-to-text generation by structure-aware seq2seq learning. In: AAAI

Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: EMNLP

Mei H, Bansal M, Walter MR (2016) What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In: NAACL-HLT

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: NIPS

Narayan S, Gardent C (2020) Deep learning approaches to text production. Synth Lect Human Lang Technol 13:1–199

Nie F, Yao JG, Wang J, Pan R, Lin CY A (2019) simple recipe towards reducing hallucination in neural surface realisation. In: ACL

Novikova J, Dusek O, Curry AC, Rieser V (2017a) Why we need new evaluation metrics for NLG. In: EMNLP

Novikova J, Dusek O, Rieser V (2017b) The E2E dataset: New challenges for end-to-end generation. In: SIGdial Meeting on Discourse and Dialogue

Papineni K, Roukos S, Ward T, Zhu W (2002) Bleu: a method for automatic evaluation of machine translation. In: ACL

Parikh AP, Wang X, Gehrmann S, Faruqui M, Dhingra B, Yang D, Das D (2020) ToTTo: A Controlled Table-To-Text Generation Dataset. In: EMNLP

Perez-Beltrachini L, Gardent C (2017) Analysing data-to-text generation benchmarks. INLG

Perez-Beltrachini L, Lapata M (2018) Bootstrapping generators from noisy data. In: NAACL-HLT

Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, Sykes C (2009) Automatic generation of textual summaries from neonatal intensive care data. Artif Intell 173:789–816

Puduppully R, Dong L, Lapata M (2019a) Data-to-text generation with content selection and planning. In: AAAI

Puduppully R, Dong L, Lapata M (2019b) Data-to-text generation with entity modeling. In: ACL

Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) Stanza: A Python natural language processing toolkit for many human languages. In: System Demonstrations @ ACL

Rebuffel C, Soulier L, Scoutheeten G, Gallinari P (2020) Parenting via model-agnostic reinforcement learning to correct pathological behaviors in data-to-text generation. In: INLG

Reiter E (2018) A structured review of the validity of BLEU. Comput Linguist 44:393–401

Reiter E, Belz A (2009) An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Comput Linguist 35:529–558

Reiter E, Dale R (1997) Building applied natural language generation systems. Nat Lang Eng 3:57–87

Roberti M, Bonetta G, Cancelliere R, Gallinari P (2019) Copy mechanism and tailored training for character-based data-to-text generation. In: ECML-PKDD

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536

Sanguinetti M, Bosco C (2015) Parttut: The turin university parallel treebank. In: Basili R, Bosco C, Delmonte R, Moschitti A, Simi M (eds) Parli. Springer, Cham

See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. In: ACL

Sennrich R, Haddow B, Birch A (2016) Controlling politeness in neural machine translation via side constraints. In: NAACL-HLT

Shen X, Chang E, Su H, Zhou J, Klakow D (2020) Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence. In: ACL

Smeuninx N, Clerck BD, Aerts W (2020) Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and nlp. Int J Bus Commun 57(1):52–85

Stajner S, Hulpus I (2020) When shallow is good enough: Automatic assessment of conceptual text complexity using shallow semantic features. In: LREC

Stajner S, Nisioi S, Hulpus I (2020) Coco: A tool for automatically assessing conceptual complexity of texts. In: LREC

Thomson C, Zhao Z, Sripada S (2020) Studying the Impact of Filling Information Gaps on the Output Quality of Neural Data-to-Text. In: INLG

Tian R, Narayan S, Sellam T, Parikh AP (2019) Sticking to the facts: Confident decoding for faithful data-to-text generation http://arxiv.org/abs/1910.08684

van der Lee C, Gatt A, van Miltenburg E, Wubben S, Krahmer E (2019) Best practices for the human evaluation of automatically generated text. In: INLG

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on

Neural Information Processing Systems 2017, 4–9 December 2017. Long Beach, CA, USA, pp 5998–6008

Wang H (2019) Revisiting challenges in data-to-text generation with fact grounding. In: INLG

Wen T, Gasic M, Mrksic N, Su P, Vandyke D, Young SJ (2015) Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In: Empirical Methods in Natural Language Processing

Wiseman S, Shieber SM, Rush, AM (2017) Challenges in data-to-document generation. In: Empirical Methods in Natural Language Processing

Wiseman S, Shieber SM, Rush, AM (2018) Learning neural templates for text generation. In: Empirical Methods in Natural Language Processing

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019) Huggingface's transformers: State-of-the-art natural language processing. http://arxiv.org/abs/1910.03771

Xia F, Palmer M (2001) Converting dependency structures to phrase structures. In: HLT