Record-to-Text Generation with Style Imitation

Abstract

Recent neural approaches to record-to-text generation have mostly focused on improving content fidelity while lacking explicit control over writing styles (e.g., sentence structures, word choices). More traditional systems use templates to determine the realization of text. Yet manual or automatic construction of high-quality templates is difficult, and a template acting as hard constraints could harm content fidelity when it does not match the record perfectly. We study a new way of stylistic control by using existing sentences as "soft" templates. That is, a model learns to imitate the writing style of any given exemplar sentence, with automatic adaptions to faithfully describe the record. The problem is challenging due to the lack of parallel data. We develop a neural approach that includes a hybrid attention-copy mechanism, learns with weak supervisions, and is enhanced with a new content coverage constraint. We conduct experiments in restaurants and sports domains. Results show our approach achieves stronger performance than a range of comparison methods. Our approach balances well between content fidelity and style control given exemplars that match the records to varying degrees.

1 Introduction

Recent years have seen remarkable progress in *neural* natural language generation to produce well-formed coherent text (Sutskever et al., 2014; Vaswani et al., 2017). Yet, controllability over various text properties, as an essential demand to ensure the utility of generations in real-world applications, has not attained the same level of advancement. Record-to-text generation is one of such applications with ubiquitous practical use, in which natural language text is generated to describe a given data record such as a box score of a sports player or an infobox table of a restaurant.

Though current record-to-text neural approaches with encoder-decoder models could produce fluent text with high fidelity to content ("what to say"), they largely lack control over the writing style, such as sentence structures and word choices ("how to say"). Many efforts have been made to promote the overall diversity in the record-to-text generation, e.g., through latent variables (Ye et al., 2020) or customized model architectures (Jagfeld et al., 2018; Deriu and Cieliebak, 2018), yet fine-grained style manipulation is not permitted. This contrasts with the traditional text generation systems which separate content planning and surface realization (Reiter and Dale, 1997), and usually determine the realization with explicit templates (Kukich, 1983; McRoy et al., 2000; Kondadadi et al., 2013).

Controlling writing style with "hard" templates could suffer from unscalable template creation and lack of generation flexibility. Though previous work (Wiseman et al., 2018; Dou et al., 2018; Angeli et al., 2010) has enabled automatic template extraction, the templates still act as hard constraints and could harm the content fidelity of generations when the template does not exactly match the content in a record.

In this paper, we study a new way of stylistic control in record-to-text generation by using any existing sentences as "soft" templates. That is, we learn to *imitate* the writing style of a given exemplar sentence. The goal is two-fold: to generate text that not only faithfully describes all content in the record, but also inherits as many of the exemplar's stylistic characteristics as possible (Figure 1). The new approach sidesteps the restrictions with traditional dedicated templates and allows us to use arbitrary exemplar sentences that could be describing distinct content. As shown in Figure 1, the model automatically adopts the soft templates to varying extents based on how well they match the record and precisely express the desired content.

Data Record

Name	Food	Area	Price	Near
Loch Fyne	Italian	Riverside	£20-25	Strada

Exemplar 1 Zizzi is a pub providing fine French dining but with an expensive price, located near Cocum in the city center.

Generation 1 Loch Fyne provides fine Italian dining with a £20-25 price, located near Strada at the riverside.

Exemplar 2 Located near the Blue Spice, there is a highly-rated place, the Mill, as a choice that frugally priced.

Generation 2 Located near Strada by the river, there is a place with Italian foods, Loch Fyne, as a choice that priced £20-25.

Exemplar 3 With a family-friendly atmosphere and a 5-star rating, Aromi is a pub in the city center.

Generation 3 With Italian foods and a moderate price range, Loch Fyne is near Strada at the riverside.

Figure 1: An example of generating sentences that describe the data record and imitate the style of given exemplar sentences (i.e., soft templates). The generations *adaptively* inherit the structural and phrasing characteristics (highlighted with cyan boxes) of the exemplars. For instance, exemplar 2 does not match the record content perfectly (e.g., it does not describe the food). The generation adapts the structure to add "with Italian foods". All such automatic adaptions are highlighted in orange. Note that the word "providing" in exemplar 1 is also adapted to "provides" for grammar correction.

To this end, we develop a neural approach that balances well between content fidelity and style imitation. A key learning challenge is the lack of parallel data, i.e., triples of (record, exemplar sentence, target description). Instead, we usually only have access to abundant record-description pairs[1]. The proposed approach learns with rich weak supervisions derived from the record-description pairs. Architecture-wise, we develop a hybrid attention-copy mechanism that offers differentiated treatments of the content and style sources. Further, based on the structural nature of data records, we devise a new content coverage constraint for the balanced embodiment of both content and style in the generation.

We conduct empirical studies on corpora from two domains, including restaurant recommendation (Dušek et al., 2019) and NBA reports (Wiseman et al., 2017). Experiments show our models strongly improves over a diverse set of comparison methods in terms of both automatic and human evaluations. In particular, given exemplar sentences that match data records to varying degrees, our approach retains a good content-style balance.

2 Related Work

Record-to-Text Generation Many efforts have been made to improve the fidelity of generated text to the record content, through sophisticated neural architectures (Wiseman et al., 2017; Puduppully et al., 2019; Iso et al., 2019), hybrid retrieval and generation (Hashimoto et al., 2018; Weston et al., 2018; Cao et al., 2018; Pandey et al., 2018; Peng et al., 2019), and others. These approaches do not have the additional goal of style control as ours, and usually perform supervised learning based on record-description pairs. Traditional record-to-text generation systems implement a pipeline architecture consisting of separate components, including content planning, sentence planning, and surface realization (e.g., Reiter and Dale, 1997; Kukich, 1983; McRoy et al., 2000; Kondadadi et al., 2013). Recent work (Wiseman et al., 2018) integrates the template use in a more end-to-end neural model. Differing from the previous work where templates act as hard constraints, we study the new setting of using any existing sentences as exemplars, allowing the model to adaptively imitate the style while ensuring content fidelity.

Text Style Transfer There has been growing interest in text style transfer (Hu et al., 2017; Shen et al., 2017; Prabhumoye et al., 2018; Subramanian et al., 2019; Logeswaran et al., 2018, etc) which assumes an existing sentence of certain content, and modifies single or multiple textual attributes (e.g., sentiment) of the sentence without changing the content. Our problem differs in important ways in that we assume the abstract writing style is encoded in an exemplar sentence and attempts to modify its concrete content to express the new information in a structured record. The different settings can lead to different application scenarios in practice, and

[1]This highlights the difference from the recent retrieval-and-generation work (e.g., Hashimoto et al., 2018; Weston et al., 2018; Cao et al., 2018; Peng et al., 2019) which focuses only on content fidelity and thus is a supervised learning problem given the record-description pairs.

pose varying technical challenges. In particular, though the recent style transfer research (Subramanian et al., 2019; Logeswaran et al., 2018) has controlled multiple categorical attributes which are largely independent or loosely correlated to each other, a data record in our task, in comparison, can contain a varying number of fields, have many possible values, and are structurally coupled. Our empirical studies (sec 5) show the recent models designed for style transfer fail to perform well on the problem under study. We also note recent work of syntactically-controlled paraphrase generation based on either constituency parse (Iyyer et al., 2018) or reference sentences (Chen et al., 2019). The problem nature of record-to-text generation in this work leads to a solution with different architectures and learning approaches.

3 The Task: Record-to-Text Generation with Style Imitation

For clarity, we first formally describe the problem of record-to-text generation with style imitation. We also establish the key notations used in the rest of the paper.

Consider a data record $\boldsymbol{x}$ which consists of a set of *fields* and their values (e.g., field "Food" and its value "Italian" in Figure 1). Note that different records can include different fields. For example, the field "Customer Rating" is included in some records but not the one in Figure 1. Record-to-text generation aims to produce a sentence to describe the content in the record. We are additionally given an exemplar sentence $\boldsymbol{y}_e$ which could be describing distinct content in the same domain. The goal of the task is thus to generate a new sentence $\boldsymbol{y}$ that achieves **(1) content fidelity** by describing the content in $\boldsymbol{x}$ accurately and completely, and **(2) style embodiment** by retaining as much of the writing style (e.g., sentence structure, word choice, etc) of $\boldsymbol{y}_e$ as possible.

A solution to the problem is required to balance well between the two objectives, by adaptively rewriting necessary portions of the reference $\boldsymbol{y}_e$ to express the desired content in a correct and fluent way, while at the same time editing $\boldsymbol{y}_e$ to a minimum extent to inherit its style. The demand for adaptive trade-off necessitates developing learning approaches for flexible imitation and generation.

To the best of our knowledge, there is no large data containing the desired $(\boldsymbol{x}, \boldsymbol{y}_e, \boldsymbol{y})$ *triples* for supervised learning. Instead, we often only have access to *pairs* of record and its description which was originally written without following any designated style. In the next section, we develop a neural approach that learns style imitation given only the paired data.

4 The Approach

Denote the proposed neural model as $p_\theta(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{y}_e)$. The model has a hybrid attention-copy mechanism (sec 4.1) for differentiated treatment of source content and style exemplar. We learn the model by constructing weak supervisions from the available non-parallel data (sec 4.2), and further encourage accurate content description with a content coverage constraint (sec 4.3). Figure 2 presents an overview of the approach.

4.1 Hybrid Attention-Copy Architecture

The overall architecture of the neural model consists of two encoders and one decoder. The two encoders extract the representation of the data record $\boldsymbol{x}$ and exemplar $\boldsymbol{y}_e$, respectively. Concretely, for each field in $\boldsymbol{x}$, we concatenate the embedding vectors of the field and its value, and feed the sequence of field-value embeddings to the encoder.

The decoder generates the output sentence with a hybrid attention-copy mechanism. In particular, the decoder applies joint attention over both $\boldsymbol{y}_e$ and $\boldsymbol{x}$, and uses a copy mechanism (Gu et al., 2016) *only* on the field values in the record $\boldsymbol{x}$. More concretely, at each step t, the decoder first attends jointly to the hidden states of both encoders, and obtains a decoding hidden state $\boldsymbol{h}_t$. The final output distribution is the weighted-sum of two distributions:

$$\boldsymbol{P}_{out}^{(t)} = g_t \cdot \boldsymbol{P}_V^{(t)} + (1 - g_t) \cdot \boldsymbol{P}_{\boldsymbol{x}}^{(t)} \quad (1)$$

where g_t is the probability of generating a token from the vocabulary; $\boldsymbol{P}_V^{(t)}$ is the generation distribution over the whole vocabulary; $\boldsymbol{P}_{\boldsymbol{x}}^{(t)}$ is the copy distribution over the field values in the record.

4.2 Learning with Weak Supervisions

The two problem goals, namely content fidelity and style embodiment, are complementary and to some extent competitive. We derive weak forms of supervisions for each of them, respectively, based on the corpus of record-description pairs available.

Exemplar Retrieval First, for each record $\boldsymbol{x}$, we automatically construct the exemplar $\boldsymbol{y}_e$ through retrieval. Specifically, we use $\boldsymbol{x}$ to retrieve another record $\boldsymbol{x}_e$ based on their *distance*, and use

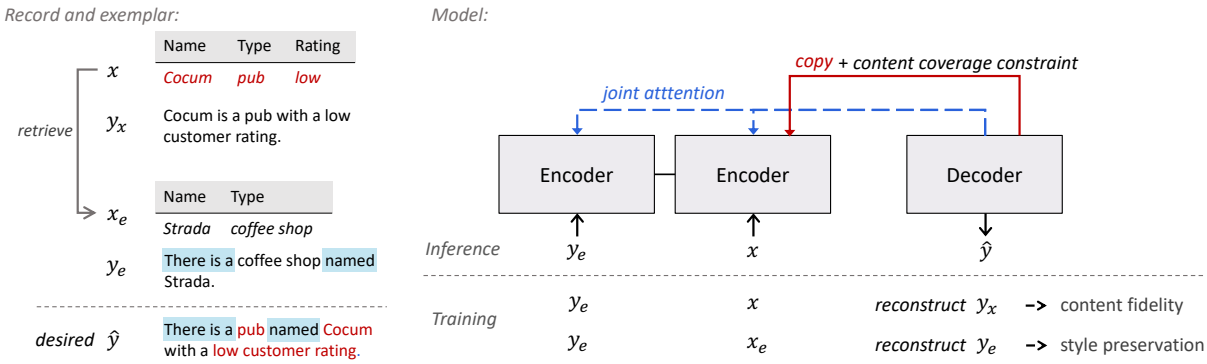

Figure 2: A (simplified) data example and retrieval **(left)** and the model overview **(right)**. The proposed approach uses a hybrid attention-copy mechanism, and is learned with weak supervisions and a content coverage constraint.

the description associated with $\boldsymbol{x}_e$ as the exemplar sentence $\boldsymbol{y}_e$ in training. The distance between two records is defined as the total number of different fields included in the records (i.e., fields that occur in one record but not the other). We present more details of the retrieval in the Appendix A. Figure 2 gives an illustration of retrieved exemplar (with distance $= 1$). We study the effect of training with exemplars of varying distances in the experiments.

Content Objective Given the retrieved results, we next tackle content fidelity. Consider the description associated with $\boldsymbol{x}$, which, though not following the desired style of $\boldsymbol{y}_e$, has accurately presented the content in $\boldsymbol{x}$. Denote the description as $\boldsymbol{y}_x$. We thus devise the first learning objective that reconstructs $\boldsymbol{y}_x$ given $(\boldsymbol{x}, \boldsymbol{y}_e)$, in order to provide the model with the hints on how the $\boldsymbol{x}$ content can be presented in natural language:

$$\mathcal{L}_{content}(\boldsymbol{\theta}) = \log p_\theta(\boldsymbol{y}_x|\boldsymbol{x}, \boldsymbol{y}_e). \quad (2)$$

Style Objective For the second goal of style embodiment, we want to encourage the model to generate sentences in a similar form of $\boldsymbol{y}_e$. To this end, we notice that, if we feed the model with the exemplar sentence $\boldsymbol{y}_e$ and its corresponding record $\boldsymbol{x}_e$, then by definition the desired output would be $\boldsymbol{y}_e$ itself. We thus devise the second learning objective that reconstructs $\boldsymbol{y}_e$ given $(\boldsymbol{x}_e, \boldsymbol{y}_e)$:

$$\mathcal{L}_{style}(\boldsymbol{\theta}) = \log p_\theta(\boldsymbol{y}_e|\boldsymbol{x}_e, \boldsymbol{y}_e). \quad (3)$$

The objective essentially treats the exemplar sentence encoder and the decoder together as an auto-encoding module, which effectively drives the decoder to reproduce the exemplar's characteristics.

The above two learning objectives are competitive with each other such that, by combining them and optimizing jointly, the model is encouraged to learn to balance between content fidelity and style embodiment. A similar learning strategy of dividing a learning problem into multiple competitive objectives has also been used in previous work such as text style transfer (Shen et al., 2017; Hu et al., 2017). More formally, the above two objectives are coupled together to train the model as follows:

$$\mathcal{L}_{joint}(\boldsymbol{\theta}) = \lambda\mathcal{L}_{content}(\boldsymbol{\theta}) + (1-\lambda)\mathcal{L}_{style}(\boldsymbol{\theta}), \quad (4)$$

where $\lambda \in (0, 1)$ is the balancing weight.

4.3 Content Coverage Constraint

As shown in the empirical study (section 5), the above learning performs well in general yet sometimes still fall short of expressing the record accurately. We thus devise an additional learning constraint to enhance content fidelity. The intuition is that, given the copy mechanism over the record $\boldsymbol{x}$, each field value in $\boldsymbol{x}$ should be copied exactly once. We thus minimize the following L2 constraint that encourages the temporally aggregated copy probability of each field value in $\boldsymbol{x}$ to be 1:

$$\mathcal{C}(\boldsymbol{\theta}) = \left\| \sum_t \boldsymbol{P}_{\boldsymbol{x}}^{(t)} - \mathbf{1} \right\|^2 \quad (5)$$

where $\boldsymbol{P}_{\boldsymbol{x}}^{(t)}$, as defined in Eq.(1), denotes the copy distribution over all field values at decoding step t; and $\mathbf{1}$ is a vector with all ones.

The full model training objective with the constraint is thus written as:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{joint}(\boldsymbol{\theta}) - \eta \cdot \mathcal{C}(\boldsymbol{\theta}) \quad (6)$$

where $\eta \geq 0$ is the weight of constraint.

5 Experiments

We study on two datasets in the restaurant recommendations and NBA reports domains, respectively. We conduct both automatic and human evaluations to assess model performance. Experiment results validate the proposed approach in learning an effective, balanced control of content and style.

	Restaurant Recommend.			NBA Reports		
	Train	Dev	Test	Train	Dev	Test
#Instances	29,486	6,299	6,273	31,444	6,765	6,930
#Tokens	0.54M	0.12M	0.12M	7.88M	1.69M	1.75M
Avg Text Length	18.36	18.34	18.35	25.07	25.10	25.32
#Unique Fields	8	8	8	34	34	34
Avg #Fields	5.38	5.38	5.35	4.32	4.31	4.35

Table 1: Statistics of the two datasets.

5.1 Datasets

We derived and processed the two existing popular corpora as below. As defined in section 3, each resulting dataset contains record-description pairs. We provide more details of data processing in Appendix B. Table 1 shows the data statistics.

Restaurant Recommendations The dataset is extracted from the E2E NLG challenge (Dušek et al., 2019). A restaurant record can contain a subset of 8 fields, such as *Eat Type*, *Price Range*, and others. See Figure 1 for an example record and the different possible ways of description.

NBA Reports We extract the dataset from the NBA game corpus developed in (Wiseman et al., 2017). The original corpus consists of box-score tables of NBA matches and the corresponding full-length match reports. We first split each report into individual sentences and extract the associated information from the box-score table as the data record. The data contains 34 unique fields, such as *Points*, *Rebounds* , *Field-Goal Percentage*, etc. Though the recorded fields look regular, the natural language descriptions are rich with variation. For example, for a field value *Points: 14*, one could say *"contributed 18 points"*, *"reached double figures"*, or, fusing with other fields, *"scored an amazingly efficient 18 points on 7-of-8 shooting"*, etc.

5.2 Setup

Comparison Approaches

We compare with diverse approaches for a comprehensive analysis of the task and proposed approach:

- **Reference for Content Fidelity: AttnCopy-S2S.** We first consider a conventional record-to-text model designed for only expressing the content. As style imitation is omitted, the method is expected to excel on content fidelity but fail on style control. Specifically, we use a sequence-to-sequence model (Sutskever et al., 2014) augmented with the proposed attention-copy mechanism (Section 4.1), which is trained supervisedly on the record-description pairs.

- **Reference for Style Embodiment: Slot-filling.** The second approach serving as a reference is a traditional slot-filling method that first removes the content words in the exemplar sentence $\boldsymbol{y}_e$ to make a template, and fills in the slots with respective values in the record $\boldsymbol{x}$. As all content-independent tokens in $\boldsymbol{y}_e$ are preserved, the method is expected to perform well on style embodiment, but fail on content fidelity due to the possible mismatch between the exemplar sentences and desired content $\boldsymbol{x}$. We manually crafted a large set of slot-filling rules for each of the two datasets, respectively (see Appendix C for more details).

- **Multi-Attribute Style Transfer (MAST)** (Subramanian et al., 2019). We compare with a recent style transfer approach capable of manipulating multiple attributes. To apply to our task, we treat the field values in record $\boldsymbol{x}$ as separate attributes. The method is based on back-translation, as detailed in Appendix D.

- **Adversarial Style Transfer (AdvST)** (Logeswaran et al., 2018). As another style transfer approach for multiple attributes, the model incorporates back-translation with adversarial training to disentangle content and style representations.

Model Configurations

We studied both LSTM (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) architectures. For LSTM, we use a single layer with the Luong attention (Luong et al., 2015) and copy mechanism (Gu et al., 2016). For Transformer, use the recent copy-augmented variant following (Su et al., 2019) with 3 blocks. During training, we first set $(\lambda = 0, \eta = 0)$ to pre-train the model so that it captures the full characteristics of the exemplar sentence. We then switch to $(\lambda = 0.2, \eta = 1.0)$ for full training. Adam optimization (Kingma and Ba, 2014) is used with an initial learning rate of 0.001. At inference time, we use beam search with the width 5 and the maximum decoding length 50.

5.3 Automatic Evaluation

Metrics

Automatic evaluation of the task is an open and challenging problem. We use several quantitative metrics for the two goals of the task, namely content fidelity and style embodiment.

	Method	Restaurant Recommendations: Content %Incl.-new	Restaurant Recommendations: Content %Excl.-old	Restaurant Recommendations: Style m-BLEU	NBA Reports: Content Precision	NBA Reports: Content Recall	NBA Reports: Style m-BLEU
Reference	AttnCopy-S2S	78.88±2.08	99.71±0.06	13.95±0.52	81.62±3.25	75.65±7.42	45.5±0.71
	Slot-filling	61.23	66.2	100	56.69	71.34	100
Baselines	MAST	36.28±0.25	37.06±0.16	**91.76±0.28**	23.06±3.90	27.37±3.88	**95.43±2.71**
	AdvST	51.64±4.45	57.06±4.44	76.02±5.27	67.37±0.66	66.79±1.43	64.67±4.81
Ours	Transformer w/o Coverage	60.03±2.16	74.65±2.69	77.81±3.83	62.58±2.88	70.22±3.58	81.75±2.32
	+ Coverage	61.84±1.31	81.14±2.73	80.29±0.35	67.74±0.79	**74.35±1.22**	81.97±2.87
	LSTM w/o Coverage	60.83±1.29	81.45±1.10	78.91±1.05	68.74±3.07	69.35±3.30	79.88±2.44
	+ Coverage	**65.02±4.16**	**82.53±0.70**	82.92±3.18	**69.54±1.16**	73.27±1.18	80.66±1.89

Table 2: Results of automatic evaluation, averaged over 3 runs ± one standard deviation. The distance between the record and the exemplar is set to ≤ 5 for exemplar retrieval (see the text). Methods in the first block are two reference approaches (Section 5.2), i.e., `AttnCopy-S2S` for content fidelity and `Slot-filling` for style embodiment. For our method, we evaluate the variants with and without the coverage constraint (Section 4.3). The table highlights the best results in the blocks of Baselines and Ours under different metrics.

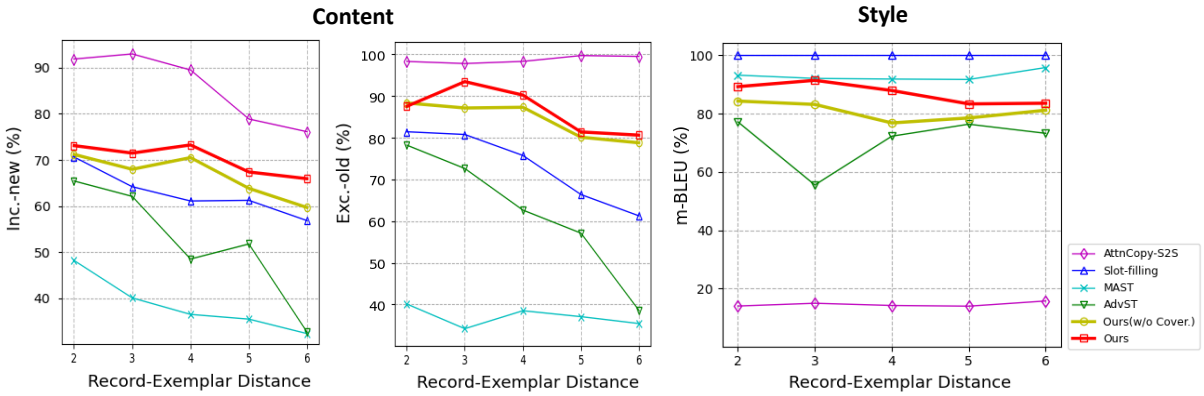

Figure 3: Effect of record-exemplar distance on model performance on the restaurant dataset. **Left**: Content fidelity performance, including "%Inc-new" and "%Exc-old". **Right**: Style embodiment performance by "m-BLEU".

- **Content fidelity.** For the NBA data, we follow the original work (Wiseman et al., 2017) and use information extraction (IE) to measure content fidelity. Given a generated sentence $\hat{\boldsymbol{y}}$ and the input data record $\boldsymbol{x}$, we extract field values from $\hat{\boldsymbol{y}}$ with an IE tool and compute the precision and recall against $\boldsymbol{x}$. We use the IE model provided in (Wiseman et al., 2017), which achieves 81% precision and 86% recall on the test set.

  We found IE on the restaurant data is too difficult to serve as a reliable metric, because the descriptions are less structured. We thus instead train a BERT-based binary classifier (see Appendix E for more details) to evaluate whether a field value is expressed in the generated sentence, which achieves 94% classification accuracy on the test set. We apply the classifier and compute both the percentage of desired $\boldsymbol{x}$ field values expressed in the generation (*%Incl.-new*) and the percentage of original content in $\boldsymbol{y}_e$ (or equivalently, $\boldsymbol{x}_e$) removed from the generation (*%Excl.-old*). The higher both numbers, the more faithfully the generation describes $\boldsymbol{x}$.

- **Style embodiment.** Imitating the exemplar style involves inheriting the sentence structure, word choices, and other surface forms of $\boldsymbol{y}_e$. Inspired by the text style transfer literature (Subramanian et al., 2019; Yang et al., 2018), we measure the BLEU score between the generated and the exemplar sentences. To reduce the influence of the change of content tokens, we mask in both sentences all obvious content tokens, e.g., player/team names and numbers, by replacing them with a special token `<M>`. We denote the metric as *m-BLEU*. This guarantees the reference approach, namely the slot-filling method, achieves an m-BLEU score of 100.

Study: Balance between Content and Style

Table 2 shows the automatic evaluation results on the two datasets. In this study, for exemplar retrieval (Section 4.2), we set the distance between a record and an exemplar to be no larger than 5 both during training and when constructing test

Model	Restaurant Recommendations			NBA Reports		
	Content Fidelity	**Style Embody**	**Fluency**	**Content Fidelity**	**Style Embody**	**Fluency**
Slot-filling	3.36	**5.00**	**4.70**	2.79	**5.00**	**4.86**
AdvST	3.56	4.24	4.02	2.88	4.00	4.09
Ours, LSTM w/o Coverage	3.91	4.38	4.58	3.43	4.13	4.59
Ours, LSTM	**4.28**	4.73	4.54	**3.88**	4.53	4.52
	Ours Better	**No Prefer.**	**Ours Worse**	**Ours Better**	**No Prefer.**	**Ours Worse**
Slot-filling	**64.1%**	18.6%	17.3%	**67.5%**	17.5%	15.0%
AdvST	**70.4%**	14.3%	15.2%	**68.8%**	17.5%	13.8%
Ours, LSTM w/o Coverage	**52.0%**	26.7%	21.3%	**51.3%**	32.5%	16.3%

Table 3: Results of human evaluation. Each metric achieves an average Pearson correlation coefficient $\geq$0.73, showing a reasonable inter-annotator agreement. **Top:** Scoring three aspects on a 5-point Likert scale. **Bottom:** Ranking the generations from pairs of models. We use our LSTM-based full model to compare with other methods.

cases. That is, the record and the exemplar sentence can have 5 mismatched fields, which thus requires strong flexibility of the generation model to be able to automatically adapt the exemplar in order to describe the record accurately.

As expected, the reference methods excel only in one of the two aspects, respectively. Specifically, `AttnCopy-S2S` expresses the desired content well, yet is incapable of embodying the designated style (e.g., m-BLEU=13.95). On the contrary, the `Slot-filling` method achieves perfect style m-BLEU by definition, but falls short of adaptively described the desired content in an accurate way, as shown by the low content scores. The two style transfer approaches (`MAST` and `AdvST`) also fail in terms of content fidelity performance. This is partly because these models are built on a different task assumption (i.e., modifying independent textual attributes) and are incompetent in manipulating the structured content well.

Our proposed approach is able to better balance between content fidelity and style embodiment. For example, in terms of content fidelity, our approach with an LSTM architecture improves over the `Slot-filling` results by 16.3 on NBA content precision and 12.9 on Restaurant content %Excl.-old. The approach meanwhile keeps a high style m-BLEU score of over 80. Regarding the ablation study, the results show the proposed content coverage constraint (Section 4.3) consistently improves both the content and style performance by a large margin. We note that the LSTM and transformer architectures perform comparably, with LSTM slightly better on the restaurant dataset. We speculate that the copy mechanism of LSTM (Gu et al., 2016) is slightly more effective than that of transformer (Su et al., 2019).

Study: Effect of Record-Exemplar Distance
We then study how well the different methods would perform when given exemplars of varying distances (mismatchness) to the records. Figure 3 show the content and style results under different distances. We can see that, as the exemplars deviate more from the structure of the records, the model performance drops since it is getting harder to automatically adapt the exemplars to express the desired content. For example, the "%Excl.old" score (middle panel) of the methods `Slot-filling` and `AdvST` decreases quickly. Our approach maintains a more stable performance and keeps a better content-style balance. The results also show the proposed content coverage constraint consistently offers enhanced performance.

5.4 Human Evaluation

We also perform human evaluation for a more thorough and accurate comparison. Following the experimental settings in prior work (Subramanian et al., 2019; Logeswaran et al., 2018; Shen et al., 2017), we undertake two types of human evaluation: (1) We ask three human annotators to score generation results in three aspects, namely content fidelity, style embodiment, and sentence fluency, on a 5-point Likert scale. (2) We present to each annotator a pair of generated sentences, one from our model and the other from a comparison method, then ask the annotator to rank the two sentences by considering the above criteria jointly. Annotators can also choose "no preference" if the sentences are equally good or bad. For each study, we evaluated on 80 test instances. We use the LSTM architecture as it outperforms the transformer slightly in the automatic evaluation. We compare with the `Slot-filling` method, `AdvST` (which is bet-

Content Record	**Name** Cocum	**EatType** coffee shop	**Food** Italian	**PriceRange** £20-25	**CustomRating** high	**FamilyFriendly** family friendly
Exemplar 1	Looking for French food near Zizzi? Come try Strada, which has a 3-star customer rating and priced lowly.					
Slot filling	Looking for Italian [...] food near Zizzi? Come try [...] Cocum, which has a high customer rating and priced £20-25.					
AdvST	For Italian [...] place near Zizzi? Come try [...] Cocum, which has a high customer rating with priced £20-25.					
Ours	Looking for an Italian coffee shop? Come try family-friendly Cocum, which has a high customer rating and priced £20-25.					
Exemplar 2	Along the riverside near Cafe Rouge, there is a Japanese food place called The Golden Curry. It has an average customer rating since it is not a family-friendly environment.					
Slot-filling	Along the riverside near Cafe Rouge [...], there is a Italian food [...] place called Cocum. It has an high customer rating since it is not a family-friendly environment.					
AdvST	Along the riverside near the Ranch [...], there is a Italian food [...] place called Cocum. It has [...] high customer rating since it is not a family-friendly environment.					
Ours	Priced £20-25, there is an Italian food coffee shop called Cocum. It has a high customer rating since it is a family-friendly environment.					

Content Record	**PLAYER** Patrick	**PLAYER** Dwight Howard	**PLAYER** Harden	**PTS** 10
Exemplar	Both J.J. Hickson and Timofey Mozgov reached double - figures , scoring 10 and 15 points.			
Slot-filling	Both Patrick [...] and Dwight Howard reached double - figures , scoring 10 and 15 points.			
AdvST	Both J.J. Hickson [...] and Dwight Howard reached double - figures , scoring 10 and 10 points.			
Ours	Patrick , Dwight Howard and Harden reached double - figures , scoring 10 points.			

Table 4: Example outputs by different models given various exemplar sentences. Text of erroneous content and syntax are highlighted in red, where [...] indicates desired content that is missing. Text portions about the writing style in both exemplars and the generated sentences by our model are highlighted in blue.

ter than MAST in automatic evaluation), and our variant without the coverage constraint.

Table 3 shows the results. From the top block, as discussed above, the Slot-filling method performs well in terms of style embodiment and fluency. However, its content fidelity is extremely weak. In contrast, our model achieves a better balance across the three criteria, by obtaining the best performance on content fidelity and reasonably high scores on both style embodiment and fluency. The fluency of our full model is slightly inferior to the variant without the coverage constraint, which is not unexpected since the full model modifies more portions of the exemplar sentences, which would result in minor language mistakes.

The bottom block of Table 3 shows the human ranking results. We can see that our model consistently outperforms the comparison methods with over 50% wins on both datasets.

5.5 Qualitative Study

Table 4 shows samples on two test cases. We can see that the proposed full model performs superior to other approaches in effectively retaining the desired style and describing the content. For example, in the first two examples, other approaches often fail to remove the redundant content (e.g., "near Zizzi" or "riverside") from the generation while neglecting desired fields in the record. The proposed model performs better by adaptively adding and deleting text portions for accurate content description. Similarly, in the third case, both Slot-filling and AdvST fail to convey the new field value "Harden" given the exemplar, and leave in irrelevant information given the second one due to the different record structures between $\boldsymbol{x}$ and $\boldsymbol{x}_e$. In contrast, our full model generates the desired sentence. Appendix F provides more examples including failure cases.

6 Conclusion

We have studied the new problem of record-to-text generation with style imitation. We developed a new approach with an attention-copy mechanism, weakly supervised learning, and a content coverage constraint. Experiments show the approach achieves a good balance between content fidelity and style control, and is flexible to adapt exemplars that do not match the record perfectly. We are interested in applying the style imitation approach to control longer paragraphs given data tables.

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *EMNLP*, pages 502–512.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *ACL*, pages 152–161.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *ACL*.

Jan Milan Deriu and Mark Cieliebak. 2018. Syntactic manipulation for generating more diverse and interesting texts. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 22–34, Tilburg University, The Netherlands. Association for Computational Linguistics.

Longxu Dou, Guanghui Qin, Jinpeng Wang, Jin-Ge Yao, and Chin-Yew Lin. 2018. Data2text studio: Automated text generation from structured data. In *EMNLP*, pages 13–18.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG Challenge. *arXiv preprint arXiv:1901.11528*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *ACL*.

Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy S Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Advances in Neural Information Processing Systems*, pages 10073–10083.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

Hayate Iso, Yui Uehara, Tatsuya Ishigaki, Hiroshi Noji, Eiji Aramaki, Ichiro Kobayashi, Yusuke Miyao, Naoaki Okazaki, and Hiroya Takamura. 2019. Learning to select, track, and generate for data-to-text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2102–2113, Florence, Italy. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *NAACL*.

Glorianna Jagfeld, Sabrina Jenne, and Ngoc Thang Vu. 2018. Sequence-to-sequence models for data-to-text natural language generation: Word- vs. character-based processing and output diversity. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 221–232, Tilburg University, The Netherlands. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ravi Kondadadi, Blake Howald, and Frank Schilder. 2013. A statistical NLG framework for aggregated planning and realization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1406–1415, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5108–5118.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Susan W. McRoy, Songsak Channarukul, and Syed S. Ali. 2000. YAG: A template-based generator for real-time systems. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 264–267, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. 2018. Exemplar encoder-decoder for neural conversation generation. In *ACL*, pages 1329–1338.

Hao Peng, Ankur P Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. In *NAACL*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. In *ACL*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6830–6841.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy. Association for Computational Linguistics.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *ICLR*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *EMNLP*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2018. Learning neural templates for text generation. In *EMNLP*.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *NeurIPS*.

Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, and Lei Li. 2020. Variational template machine for data-to-text generation. In *ICLR*.