TaKG: A New Dataset for Paragraph-level Table-to-Text Generation Enhanced with Knowledge Graphs

Qianqian Qi, Zhenyun Deng, Yonghua Zhu, Lia Lee, Michael Witbrock, Jiamou Liu
University of Auckland
{qqi518, zden658, yzhu970, jlee794}@aucklanduni.ac.nz
{m.witbrock, jiamou.liu}@auckland.ac.nz

Abstract

Table-to-text generation refers to a task that generates text using information provided by a given fact table. We introduce TaKG, a new table-to-text generation dataset with the following highlights: (1) TaKG defines a long-text (paragraph-level) generation task as opposed to well-established short-text (sentence-level) generation datasets. (2) TaKG is the first large-scale dataset for this task, containing three application domains and ∼750,000 samples. (3) To address the divergence phenomenon, TaKG enhances table input using external knowledge graphs, extracted by a new Wikidata-based method. We then propose a new Transformer-based multimodal sequence-to-sequence architecture for TaKG that integrates two pretrained language models RoBERTa and GPT-2. Our model shows reliable performance on long-text generation across a variety of metrics, and outperforms existing models for short-text generation tasks.

1 Introduction

Data-to-text generation refers to semantic-preserving conversion from structured data to (unstructured) text. *Table-to-text* generation is a class of data-to-text generation tasks where the input data takes the form of fact tables (Kukich, 1983). Table-to-text generation has widespread applications from biography generation (Lebret et al., 2016) to event summarisation (Wiseman et al., 2017). Thus developing a fluent, truthful and informative table-to-text generation system has attracted considerable attention (Liu et al., 2018; Wang et al., 2020; Liu et al., 2021). A critical factor in building such a system is to prepare reliable and large-scale table-to-text datasets.

However, existing table-to-text generation benchmarks have some clear limitations. First, most existing datasets, such as E2E (Novikova et al., 2017) and ToTTo (Parikh et al., 2020), focus on single-sentence generation tasks, which severely limits their use for tasks that involve the generation of *long texts*, e.g., entire paragraphs. Then, the few datasets that involve long (paragraph-level) text generation, such as MLB (Puduppully et al., 2019) and ROTOWIRE (Wiseman et al., 2017), consist of too few samples (less than 30k). Last, real-world data-to-text generation tasks tend to exhibit the so-called *divergence* phenomenon, where the input data fail to provide all the key information in the target text description (Dhingra et al., 2019; Wiseman et al., 2017; Chen et al., 2019). This is illustrated by an example in Figure 1 for the Dutch painter Jacoba Surie (extracted from WikiBio dataset (Lebret et al., 2016)). Existing table-to-text datasets in general lack of sufficient external knowledge required to generate the target text.

To address these issues, we introduce a new table-to-text generation dataset: TaKG (Table-and-Knowledge Graph) with the following highlights [1]: First, samples in TaKG contain long text (i.e., paragraphs) and their corresponding infoboxes (tables) extracted from Wikipedia. Thus TaKG amounts to a long-text generation task. TaKG contains three domains: biography, place, school, with a total of 745,574 samples, considerably larger than existing table-to-text datasets. To resolve the divergence issue, we employ external knowledge to "fill" the information in text description that is missing from the input infobox. In particular, we exploit another large-scale knowledge graph (KG) repository Wikidata[2]. The KGs are added in TaKG as auxiliary input. Figure 1 (upper right) shows an example KG.

The goal of this paper is two-fold. (1) We first introduce the TaKG dataset. In a nutshell, TaKG defines a task that takes a fact table (i.e., infobox) about a *target entity* and a Wikidata KG as input, and seeks a paragraph-level text description of the target entity. Section 3 provides more details. (2)

[1] TaKG is available on: https://bit.ly/3RR4erL
[2] http://www.wikidata.org/

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 176–187
November 20–23, 2022.

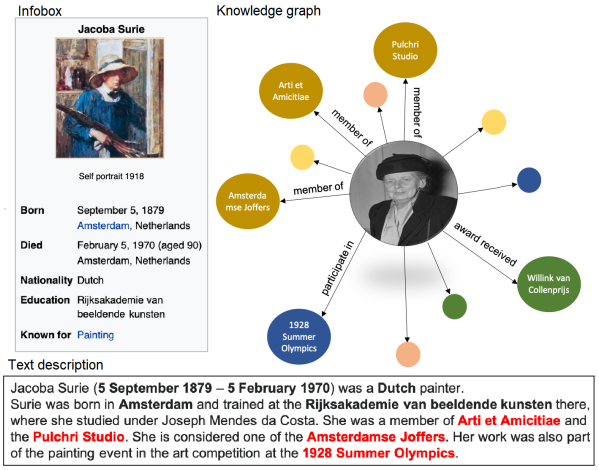

Figure 1: This is an example of generating a biography of Jacoba Surie. The upper left table and bottom text description are extracted from Wikipedia. The blue colour words are the item with hyperlinks. The right upper knowledge graph is retrieved from Wikidata. In lower biography, the red color words indicate the information missing from Wikipedia table but can be found in the knowledge graph.

We then demonstrate how TaKG may be used as a worthy benchmark to train a model for paragraph-level table-to-text tasks. Generating text with multiple data sources is challenging on two axes: table-KG information fusion and high-fidelity natural text generation. To address these challenges, we leverage pretrained language models (PLMs), such as RoBERTa (Liu et al., 2019b) and GPT-2 (Radford et al., 2019), for their abilities to acquire cross-domain knowledge. A seq2seq architecture is proposed for our task and utilizes two PLMs that fuses multiple input sources together. More details are provided in Section 4. We validate our model's ability to generate long-text using the TaKG dataset. To further verify our method's wide applicability, we also demonstrate that, over standard short-text (sentence-level) generation tasks such as WikiBio, our method also outperforms state-of-the-art benchmarks, with large margin on WikiBio up to 7.3% (BLEU) and 9% (Rouge) increment. See Section 5.

Our contributions are summarized below:
1. Creating a large-scale paragraph-level dataset TaKG for table-to-text generation enhanced with knowledge graphs.
2. Long-text generation: Designing a new seq2seq model using two PLMs to accomplish TaKG tasks.
3. Short-text generation: Demonstrating that our new model outperforms benchmarks for sentence-level table-to-text generation tasks.

2 Related Work

Table-to-Text datasets. Existing table-to-text generation datasets are either *single-sentence* or *multi-sentence* generation tasks. The former, such as E2E (restaurant domain) (Novikova et al., 2017), ToTTo (Parikh et al., 2020), and WikiBio (biography domain) (Lebret et al., 2016), are limited in terms of what the task seeks to generate. The latter, such as MLB (26.3k samples) (Puduppully et al., 2019), ROTOWIRE (4.9k samples) (Wiseman et al., 2017), UK-Place (12k samples) and UK-School (5k samples) Chen et al. (2019) contain very few samples and are thus too small-scale.

WikiBio dataset above differs from the other datasets in the sense that each of its samples contains in fact a full paragraph of biography of a person. Nevertheless, the task WikiBio specifies only the first sentence of the paragraph as the ground truth output. Indeed, all benchmarks tested on WikiBio used the dataset as a single-sentence text generation task. Due to the presence of paragraph-level texts in WikiBio, we include WikiBio samples in TaKG by incorporating the *entire paragraph* as ground truth output.

Divergence has been a common issue in multi-sentence generation (Dhingra et al., 2019; Wiseman et al., 2017; Chen et al., 2019). To address this issue, in the UK-Place and UK-School datasets Chen et al. (2019) complements the input tables with some background knowledge. To obtain the background knowledge, the authors take hyperlinked keywords in the Wikipedia infobox and extract one-hop facts of those keywords from Wikidata, a large-scale open-domain knowledge graph repository containing close to 100 million data items (Vrandečić and Krötzsch, 2014). These one-hop facts are then used as background knowledge. However, we point out that this method may produce irrelevant background knowledge that distracts the generation of target text. This is because the hyperlinked keywords in the infoboxes are often not item-specific. For example, 'Painting' and 'Amsterdam' are keywords for the instance illustrated in Figure 1, which are clearly insufficient to deriving specific facts about the Dutch painter Jacoba Surie. In our work, we will integrate samples from UK-Place and UK-School into TaKG while adopting a different way to derive external knowledge graph.

PLM-based data-to-Text generation. With the popularity of Transformer (Vaswani et al., 2017), several large-scale pretrained language models

(PLMs) have been deployed in text generation tasks. Since PLMs are pretrained on a large-scale corpus, their broad applicability with little fine-tuning may suggest that these models have learnt cross-domain knowledge and some common sense from its pretraining step. Recent work has implemented PLMs with multiple input types, e.g., audio (Nagrani et al., 2020), video (Sun et al., 2019), table (Saxena et al., 2020) and knowledge graph (Marino et al., 2021). Nevertheless, no PLM has been designed for the type of tasks presented by TaKG. Inspired by these recent successes, we apply PLMs to train a model for TaKG and demonstrate that it is possible to control PLMs to generate fluent and informative text from tables and knowledge graphs.

3 The TaKG Dataset

Dataset description. TaKG contains three sub-datasets each covers a unique domains: biography, school and place. They are constructed using the samples from WikiBio, UK-Place and UK-School (Chen et al., 2019) respectively. The number of instances of TaKG in different domains are shown in Table 1. Table 2 shows the statistics of the training set in TaKG. The two main columns indicate the average number of word and the average number of relations respectively. For the words statistic, we count by removing the repeated words from table and KG. For example, 'name' is a kind of relation in table, while 'family name' and 'given name' are two relations used in KG. We calculate 'name' as one duplication.

	Train	Dev	Test
TaKG-Biography	582,659	72,831	72,831
TaKG-Place	9,823	1,228	1,228
TaKG-School	3,979	497	498

Table 1: Number of instances for TaKG-Biography, TaKG-Place and TaKG-School.

The divergence phenomenon calls for external knowledge, alongside the fact table, as input to data-to-text generation tasks. Knowledge graph are large knowledge base that facilitates effective representation, storage, and retrieval of knowledge. Wikidata is an exemplary large-scale open-domain knowledge graph which stores comprehensive knowledge regarding famous individuals, places, and organisations (Vrandečić and Krötzsch, 2014). We thus leverage Wikidata to extract our knowledge graphs as extra input in TaKG.

Unlike Chen et al. (2019) which guides the extraction of knowledge through hyperlinks, we design a new method that ensures completeness and relevance of the extra information. As WikiBio is collected using Wikipedia pages, for each WikiBio instance, we first use the provided unique Wikipedia URL IDs to get the corresponding page titles. Then these titles are used as center entities to retrieve KGs from Wikidata. For UK-Place and UK-School, we use the 'articletitle' attribute in the table to get Wikipedia URL first and then follow the same procedure as WikiBio. We ignore some of the relations in KGs, such as 'image', 'signature' and 'audio'.

Task formulation. We now formally define our table-to-text generation task. The input table includes n fields with corresponding content text pairs $\{R_1, R_2, ..., R_n\}$ which are the description of the target entity. Each R_i includes tokens of field $f_1, f_2, ..., f_l$ and tokens of content $c_1, c_2, ..., c_m$. The knowledge graph retrieved from Wikidata can be denoted as $\{E_1, E_2, ..., E_k\}$, where each E_i consist of tokens of entity attribute $a_1, a_2, ..., a_s$ and tokens of value $v_1, v_2, ..., v_j$. The output is a sequence of tokens $o_1, o_2, ..., o_r$ which are the text description of the item from Wikipedia. Our task is constraining PLM in generating text from table data and KG, which can be formulated as:

$$o^*_{1:r} = \underset{o_{1:r}}{\operatorname{argmax}} \prod_{t=1}^{r} P(o_t | o_{1:t-1}, R_{1:n}, E_{1:k}), \quad (1)$$

in which, after linearisation process, table data and linked entities in knowledge graph are represented as $R_i = \langle f_{i,1:l}; c_{i,1:m} \rangle, E_i = \langle a_{i,1:s}; v_{i,1:j} \rangle$. A TaKG-Biography example is shown below which corresponds to Figure 1; other examples are shown in Appendix A.1:

- **Target Entity:** Jacoba Surie
- **Fact Table:**
  - **Born:** September 5, 1879, Amsterdam, Netherlands
  - **Education:** Rijksakademie van beeldende kunsten
  - **Known for:** Painting
  - ...
- **Knowledge Graph:**
  - Jacoba Surie | Occupation | printmaker, draftsperson, painter, lithographer, photographer
  - Jacoba Surie | Member of | Arti et Amicitiae, Amsterdamse Joffers, Sint Lucas (artist society)
  - ...
- **Text Description:**
  - Jacoba Surie (5 September 1879 – 5 February 1970) was a Dutch painter. Surie was born in Amsterdam and trained at the Rijksakademie van beeldende kunsten there, where she studied under Joseph Mendes da Costa. She was a member of Arti et Amicitiae and the Pulchri Studio ...

	Avg.# words			Avg.# relations		
	Table	KG	Duplication	Table	KG	Duplication
TaKG-Biography	44.14	39.50	8.17	12.44	13.71	4.74
TaKG-Place	51.11	16.51	5.12	19.40	5.19	1.59
TaKG-School	81.59	19.34	7.66	48.00	5.66	2.06

Table 2: Data Statistics for TaKG training set.

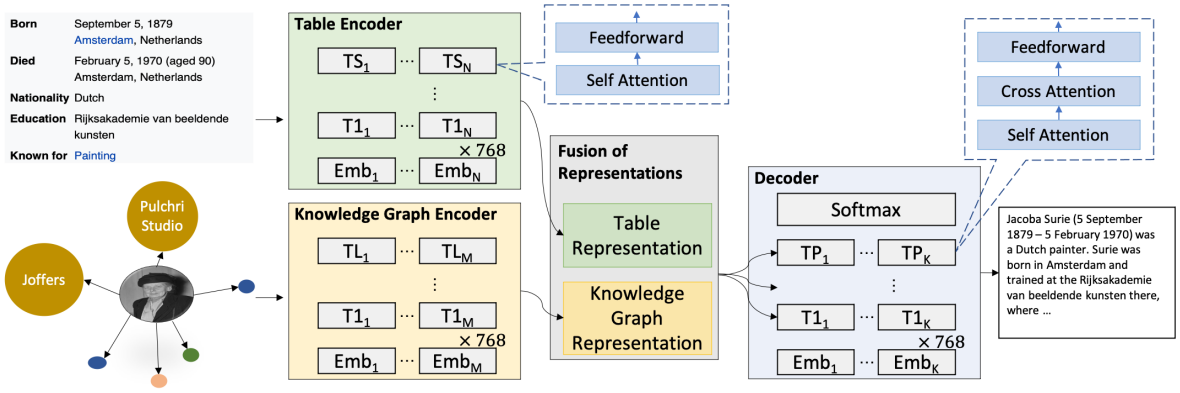

Figure 2: The overall architecture of the proposed Transformer-based seq2seq model. The tabular data and KG data are fed into Table encoder and KG Encoder separately. Then we make a concatenation of the last hidden states from the two encoders. The decoder take the concatenated hidden states as input and generate the description. All the encoders and decoder are initialized from PLMs.

4 Our Model for TaKG

We put forward a Transformer-based seq2seq framework for data-to-text generation with table data and knowledge graph as input. The parameters of encoders are initialized from RoBERTa and all of them are fine-tuned during training. The output hidden state from encoders are concatenated and sent to a Transformer-based decoder. The decoder is initialised from pretrained autoregressive models: GPT-2. After each Self-Attention layer in the decoder, we add a Cross-Attention layer which makes decoder pay attention to both encoded inputs and pre-content outputs.

The overall framework is described in Figure 2. The table pairs $\{R_1, R_2, ..., R_n\}$ include N tokens after tokenization. These tokens are fed into the embedding layer $\{Emb_1, ..., Emb_N\}$, then the embedded table is forwarded to S Transformer layers. In each encoder, the Transformer layers consist of Self-attention Layer and Feedforward Layer. The last hidden state from table encoder is denoted as $EN_T = \{EN_{T1}, EN_{T2}, ..., EN_{TN}\}$ which includes the encoded table information. In the same way, we encode KG data $\{E_1, E_2, ..., E_k\}$ using knowledge graph encoder and get the encoded KG information $EN_G = \{EN_{G1}, EN_{G2}, ..., EN_{GM}\}$.

To integrate the different data representations, we concatenate the table representation and KG representation. Then the concatenation Concat$\{EN_T, EN_G\}$ is sent to the **Cross-attention layers** in the decoder, in which there are P Transformer decoder layers. In contrast to the encoder, the decoder inserts a *Masked Multi-Head Attention* sub-layer which processes the output of the decoder stack to maintain an auto-regressive property. Sequence masking is added in the decoder to omit post-context tokens for current token. For instance, a sentence 'Surie was born in Amsterdam.' is given in the decoder, and we want to apply Self-Attention for 'born' (let 'born' be a query). In this case, we only put attention to 'Surie' and 'was' but not to 'in' or 'Amsterdam'. This method is implemented via attention mask. We get the score matrix before softmax function, and then use the attention mask matrix on the score matrix to set the undesirable token score to a negative number (-100). So that, after applying Softmax, these unwanted scores will become zero, and we keep the actual scores for present and previous tokens except future tokens.

Note that there exists a Cross-attention layer in between Self-attention layer and Feedforward layer in the decoder. The mechanism of Cross-attention is using the generated token as Query (Q) to do attention with Key (K) and Value (V) from another input source which is the concatenation of encoded input in our model. Cross-attention lets

input extracted features and previously generated output tokens attend each other and recombined to a new feature representation that sends to the next layer. In our model, we implement the decoder using GPT-2 with additional Cross-attention layer after each Self-attention layer.

Our method is proposed for entity-based table-to-text generation task which has one or multiple center entities. The possible application scenario includes historical events, news report and storytelling. For news report and storytelling, we can retrieve background information for multiple entities.

5 Experiments

In our work, there are two types of tasks: **sentence-level** generation means to generate one sentence from input data, **paragraph-level** generation generates long text (more than one sentence). We borrow the idea of linearisation (Mager et al., 2020) on table and KG data. Since GPT-2 can generate text with common sense, to some extent it is not necessary to re-train a language model from scratch. Observe that we have the same text-generation goal as the pretraining target of GPT-2 had. Hence we select RoBERTa as encoder and GPT-2 as the decoder. Via fine-tuning RoBERTa and GPT-2 for text generation, the proposed model treats RoBERTa as a feature extractor and GPT-2 as a black-box with encoded text input and text output. For RoBERTa and GPT-2, we use the pre-generated vocabulary and fine-tune the embedding layer. In particular, we have three types of layer settings: *1-layer*, *2-layer* and *12-layer*. Here, *1-layer* means fine-tuning the first layer of RoBERTa and GPT-2 in the studied seq2seq model. Similarly, *2-layer* and *12-layer* mean the corresponding layers to be fine-tuned. There are added Cross-attention (Vaswani et al., 2017) layers in GPT-2, which are trained from scratch. Decoding strategy also needs to be imposed during data-to-text generation. Here, we use Nucleus sampling (p=0.9) and Top-k (k=30) sampling methods (Holtzman et al., 2019) in decoding.

5.1 Evaluation

Automatic Evaluation The typical way to evaluate the quality of text generation is to compare the similarity between candidate text and reference texts. Other than the two commonly used automatic evaluation matrix: BLEU (Papineni et al., 2002) and ROUGE (ROUGE, 2004), we also employ evaluation from semantics, divergence, diversity, grammatic and readability aspects. There are two methods in semantic evaluation, the first calculates the cosine similarity of the semantic representation of text. Here, we use DistilRoBERTa-base , which is a distilled version (Sanh et al., 2019) of RoBERTa-base model (Liu et al., 2019b), to get the semantic representation vector of text. BERTScore (Zhang et al., 2019) is another method for semantic similarity evaluation. PARENT (Dhingra et al., 2019), a divergence index, aligns the n-grams of the references and the generated text into semi-structured data, and then calculate their precision and recall value. Self-BLEU (Zhu et al., 2018) is a metric to evaluate the diversity of the generations. It calculates the BLEU score between generations, and the average score indicates the diversity level, which is the higher the less diverse. LanguageTool [3] is a tool to check grammatical errors of generated text. Grammatical Error Rate denotes the number of grammatical errors per 100 words. For Readability (Smeuninx et al., 2020), we select Coleman–Liau [4] index (Coleman and Liau, 1975) that indicates US grade level.

Human Evaluation We conduct a human evaluation to assess the text (whether the text demonstrate good usage of English, in terms of grammar and fluency, and is easy to read) and accuracy (whether the information contained in text matches well with that in Wikipedia text). We randomly sample 20 samples from the test set of TaKG-Biography, and ask 20 participants to evaluate the text generated from our model and one baseline model: Structure-aware(Liu et al., 2018). We provide the first paragraph from Wikipedia as ground truth and the goal of the participants was to rate the text based on the readability and accuracy. We have trained the Structure-aware for 10 epochs and selected the best model based on training loss.

5.2 Experiment results

This section shows the experiment results of paragraph-level generation task and sentence-level generation task. Paragraph-level generation task is conducted on TaKG-Biography, TaKG-Place and TaKG-School, and sentence-level generation task is on WikiBio.

[3] https://languagetool.org

[4] Coleman–Liau is calculated as $CLI = 0.0588L - 0.296S - 15.8$, where L and S are the average numbers of letters and the average number of sentences per 100 words.

	BLEU	STS-RoBERTa	BERT Score	PARENT		Diversity ↓	Grammatical Error Rate ↓	Readability
				Table	KG			
Table	28.09	0.72	0.89	**0.36**	-	0.74	8.59	11.57
KG	17.33	0.65	0.88	-	0.05	**0.70**	8.66	11.52
T5 with Table	23.03	0.64	0.70	0.09	-	0.73	11.06	10.12
One Encoder	28.26	0.70	0.88	0.09	0.05	0.72	**8.16**	14.21
T5 with Table & KG	25.33	0.72	**0.90**	0.09	**0.06**	0.75	10.30	10.47
Table & KG	**29.26**	**0.73**	**0.90**	**0.36**	**0.06**	0.75	8.54	11.63

Table 3: Evaluation results of paragraph-generation task after the proposed model has been fine-tuned for 10 epochs. The metrics with ↓ stands for the performance with the smaller value is better, and the Wikipedia text readability (coleman_liau) score is 11.38.

5.2.1 Paragraph-level Generation with TaKG-Biography

We choose to use fine-tuned Transformer-based seq2seq model T5 (Raffel et al., 2019) as the baseline mode. Two **T5-small** are fine-tuned with table and concatenation of linearized table and KG separately. Besides, we use a standard seq2seq model (**One Encoder**) with the concatenation of linearized table and KG as input.

	BLEU	STS-RoBERTa
1-layer	27.12	**0.72**
2-layer	**28.09**	**0.72**
12-layer	3.33	0.30

Table 4: Comparisons of fine-tuning models on TaKG for 10 epochs with three layer settings.

As a preliminary experiment, to select the optimal number of layers of our proposed model, we compare the performance of our model at *1-layer*, *2-layer* and *12-layer* settings, respectively. BLEU score and STS-RoBERTa score are used as the evaluation metrics. For each setting, we train the model for 10 epochs. As shown in Table 4, *2-layer* get the best performance from both of the BLEU score and STS score. When we increase the layer number to 12, the BLEU score and STS score decreases to 3.33 and 0.30. In addition, the *12-layer* model requires longer time and more memory as the number of training layers increases. Thus, we select to use the *2-layer* model in the paragraph-level text generation task. Then we test the selected models with three different types of input: TaKG-Biography table, TaKG-Biography KG and complete TaKG-Biography.

The evaluation results of models fine-tuned after 10 epochs are shown in Table 3. We observe that our model using complete TaKG-Biography get the best evaluation score on BLEU, semantic (STS-RoBERTa, BERTScore), divergence (PARENT) and achieve comparable results on Grammatical Error Rate and Readability. The readability scores (US grade level 11-12) suggest that all of the models can produce text in the same readability level as Wikipedia text except T5 model (US grade level 10-11). One encoder performs better than fine-tuned T5 models on BLEU and Grammatical Error Rate.

5.2.2 Paragraph-level generation with TaKG-Place and TaKG-School

Since TaKG-Place and TaKG-School are far smaller than TaKG-Biography, we use the *1-layer* setting for the experiments in this section. From Figure 5 and 6, our proposed method using complete TaKG-Place and TaKG-School outperforms the ablation version that only considers the table data in almost all evaluation indexes except diversity and grammatical error rate. This validates the feasibility of using different data sources to improve the quality of generative text. One Encoder model get the lowest score in BLEU, BERTScore and in Diversity. From the diversity scores, the more information is provided to our model, the more deterministic text is generated. Note that, grammatical error rates are kept at a low level, which states the reliability of our method in generating text. Different from TaKG-Biography, when the exhibited models are applying on TaKG-Place and TaKG-School, they need to be fine-tuned with more epochs to learn the knowledge. From the evaluation results, our model obtains little increase on BLEU (0.01 on TaKG-Place and 0.52 on TaKG-School) comparing to the model with table input. From Table 2, for TaKG-Place and TaKG-School, the average words and relations in table are three times larger than these in KG. This is the main reason for limited performance improvement on the two datasets. The results from One Encoder prove that using one encoder for the concatenation of table and KG capture weaker representation than using separated encoders.

	BLEU	BERT Score	PARENT		Diversity ↓	Grammatical Error Rate ↓	Readability
			Table	KG			
Table	22.87	**0.88**	0.06	-	0.76	1.68	9.91
One Encoder	22.05	0.87	0.07	0.07	**0.68**	2.80	10.82
Table & KG	**22.88**	**0.88**	**0.08**	**0.08**	0.78	2.30	9.80

Table 5: Evaluation results of proposed model fine-tuned with UK-Place dataset on paragraph-level generation task for 20 epochs. The metrics with ↓ stands for the performance with the smaller value is better, and the Wikipedia text readability (coleman_liau) score is 10.97.

	BLEU	BERT Score	PARENT		Diversity ↓	Grammatical Error Rate ↓	Readability
			Table	KG			
Table	17.29	**0.88**	0.04	-	0.78	1.31	12.42
One Encoder	17.01	0.87	**0.04**	0.03	**0.72**	2.35	13.00
Table & KG	**17.81**	**0.88**	**0.04**	**0.04**	0.78	2.01	13.14

Table 6: Evaluation results of proposed model fine-tuned with UK-School dataset on paragraph-level generation task for 80 epochs. For metrics with ↓, a smaller value is better. The Wikipedia text readability (coleman_liau) score is 13.78.

	BLEU	STS-RoBERTa	BERT Score	PARENT	Diversity ↓	Grammatical Error Rate ↓	Readability (coleman_liau)
1 epoch	45.69	0.78	0.93	0.10	**0.834**	8.445	10.69
10 epoch	50.36	**0.80**	**0.94**	**0.11**	0.848	**8.355**	10.80
20 epoch	**50.52**	**0.80**	**0.94**	**0.11**	0.849	8.360	10.82

Table 7: Evaluation results of sentence-level generation task with WikiBio in terms of fine-tuning with different epoch. The metrics with ↓ stands for the performance with the smaller value is better, and the Wikipedia text readability (coleman_liau) score is 12.44.

5.2.3 Sentence-level Generation with WikiBio

Four state-of-the-art comparison methods are compared in our experiments to validate the performance of our method. Chen et al. (2019) uses background information and infobox to generate text with a RNN and Multi-Layer Perceptron (MLP) mixed model: **KBAtt**. In (Liu et al., 2018), they describe **Structure-aware** which consists of a field-gating encoder and a description generator with dual attention to generate description given factual table. **Factual Attribute** (Liu et al., 2019a) employs the force attention as well as the reinforcement learning to enrich loyal descriptions for tables. **Tree-like Planning** (Bai et al., 2020) applies a pointer network and a tree-like tuning encoder to capture more relevant attributes in the table. These methods are compared to our model that are fine-tuned after **1, 10, 20 -th epoch**.

	BLEU	Rouge
KBAtt (Chen et al., 2019)	44.59	-
Structure-aware(Liu et al., 2018)	44.89	41.21
Factual Attribute (Liu et al., 2019a)	45.47	41.54
Tree-like Planning (Bai et al., 2020)	47.09	42.82
1-epoch	45.69	41.73
10-epoch	50.36	46.46
20-epoch	**50.52**	**46.69**

Table 8: BLEU and Rouge score comparisons between proposed model and benchmark models on WikiBio dataset.

From the results reported in Table 8, our model with *1-layer* setting has a significant improvement in terms of BLEU and Rouge evaluation metrics, which validates the feasibility of integrating two pretrained language models, i.e., incorporating RoBERTa as encoder and GPT-2 as decoder. Specifically, the studied model achieves better performance than *Structure-aware* and *FA+RL* with fine-tuning only 1 epoch. When fine-tuned with 10 epochs, the demonstrated model outperforms the best comparison methods Tree-like Planning by 3% in both BLEU and Rouge. The results under 20-epoch only show a slight increase compared to the results under 10-epoch which means model get fast convergence within 10 epochs.

We also evaluate the performance of our model fine-tuned with different epochs from semantics, divergence, diversity, grammar and readability aspects as reported in Table 7. From Table 7, similar observation of fast convergence can be more easily observed in different metrics. For diversity and readability, our method gets the score of 0.834 and 10.69 (coleman_liau), which means our method not only can produce more natural language text to describe the constructed table data, but also guarantee the diversity of the generated text. Both background knowledge learnt from PLMs and external knowledge retrieved from Wikidata effectively enrich the expression of sentences.

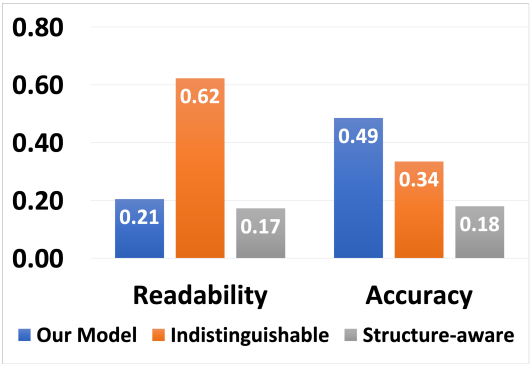

Figure 3: Human-based evaluation result. This figure shows on what proportion of samples different models achieve better scores.

5.2.4 Human Evaluation

Figure 3 shows the human evaluation results. Our model wins over the baseline in terms of both readability and semantic consistency (with ground truth). In terms of readability, 62% of the generated text are indistinguishable and our model performs better on 4% more samples than Structure-aware. In terms of semantic consistency, participants rate that our model performs better than baseline model on 49% of the samples while Structure-aware performs better on 18% samples.

5.3 Case Studies

Case study 1. Figure 4 shows an example from TaKG-Biography on the target entity '4mat'. The input, ground truth and output description are stated in the left table. The output text is generated by our proposed model trained over the complete TaKG-Biography dataset. We label the same or cognate tokens that happen in both inputs and generated text using the same text colour. It can be observed that the generation covers both table and KG inputs, for example, 'composer' and 'sound designer' is copied from the KG data, and 'british' is inferred from 'united kingdom' in both table and KG data. On another note, 'game'(highlighted in yellow colour) appears in both ground truth and generated text, but not provided by inputs. This benefits from the background knowledge learnt by pretrained language models.

The heat map in right part of Figure 4 is the visualization diagram of Cross-Attention matrix from the last Transformer block in the decoder. The darker blue colour means the more attention has been put into from output content to input tokens. The tokens in orange colour indicates the input table data and tokens in green colour is the KG information. Due to the limited space, we list the first 80 tokens from inputs and 30 tokens from outputs. From the attention map, KG data has been put more attention since fact table provide incomplete information.

Case study 2. An example of the comparison between text generated from different models is shown in Figure 5. The left part includes table input and KG input, and the right part are the generated text. Since the table input only provide birth date and name, for the models which only take table as input, they make up the description about occupation and achievement. On the contrary, when the model only takes KG as input, it generates wrong birth date as this information is missing in the KG. For models that make use of both table and KG, they are able to generate text similar to the label description. However, when there is only one encoder for table and KG, the models (One Encoder, T5) are easily making up stories. In biology generation, fabrication is not acceptable. Compare to these baseline models, the proposed model gets the best result, generating exact text that is same as ground truth. From the results, we find that it is not easy to ground T5 for text generation with structured data input via fine-tuning. Besides, using the same encoder for different types of input data works worse than using separated encoders. The main reason is that each encoder can learn the particular patterns from the designated data type.

6 Conclusion

In this study, we introduce TaKG (745,574 samples), the first large-scale KG-enhanced table-to-text dataset. Different from existing well-established sentence-level generation datasets, TaKG defines a paragraph-level generation task. Each sample of TaKG includes three parts: fact table (Wikipedia), knowledge graph (Wikidata) and paragraph-level description of an entity (Wikipedia).

We then propose a new Transformer-based sequence-to-sequence architecture for TaKG that integrates two pretrained language models RoBERTa and GPT-2. For paragraph-level generation, to generate text with multiple structured data sources, we use the simple yet effective concatenation-based fusion to combine the multiple structured data representation. Our model shows the ability to generate reliably long texts using multiple data sources (table and KG) with

Table	name: 4mat origin: united kingdom article-title: 4mat
KG	title: 4mat sex or gender: male occupation: composer, sound designer genre: electronic music, chiptune country of citizenship: united kingdom given name: matthew family name: simmonds languages spoken, written or signed: english
Generation	4mat is a british music composer and sound designer who is best known for his work on the british gamecom`` early show '', `` computer games '', and for his work on the programming of `` computer games ''. he has also collaborated with many other works including the game of `` x-play '', `` space games '' and `` robot games ''.
Ground Truth	matthew simmonds , also known as 4mat or 4-mat , is a british electronic musician and video game composer best known for his chiptunes written in tracker software . he began his career in the demoscene of the early 1990s composing on the amiga .

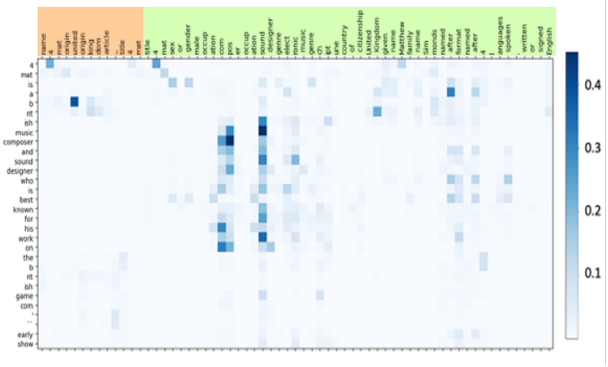

Figure 4: On the left, we have the data from the complete TaKG-Biography. The same or related tokens that happen in both input and generation are highlighted using text colour. The word 'game' with yellow colour background is the information not provided from input, but happens in both generation and ground truth text. On the right, we have a heat map of attention weight from the last cross-attention layer in the decoder. Tokens in orange colour represent table data, and tokens in green colour means KG data. From this heat map, the darker blue colour indicates, the more attention has been put. Due to the page limit, we attach the enlarged heat map in Appendix A.2.

Table Input	
birth_date	14 february 1976
name	muriel vaudey
Knowledge Graph Input	
sex or gender	female
given name	Muriel
family name	Vaudey
country of citizenship	France
languages spoken, written or signed	French
occupation	ski mountaineer
sport	ski mountaineering

Label	**muriel vaudey -lrb- born february 14 , 1976 -rrb- is a french ski mountaineer .**
Structure-aware (Liu et al., 2018)	muriel vaudey -lrb- born february 14 , 1976 -rrb- is an american actress . she is best known for her role as the vaudey vaudey in the television series ``vaudey'' .
Table	muriel vaudey -lrb- born february 15, 1976 in st. paul, minnesota -rrb-, is a canadian former cyclist. she competed in the women's individual road race event at the 1984 summer olympics.
KG	muriel vaudey -lrb- born may 21, 1989 -rrb- is a french ski mountaineer.
One encoder with Table and KG	muriel vaudey -lrb- born february 14, 1976 -rrb- is a french ski mountaineer. vaudey was born in lille. she competed first in the Labrador section of the nationalqqara selection list.
T5 with Table	muriel vaudey -lrb- born february 14, 1976 -rrb- is a russian actor. he is known for his role as sarah sahna in the sahna '' film sahna ''.
T5 with Table and KG	muriel vaudey -lrb- born february 14, 1976 -rrb- is a french ski mountaineer. she is the first french female ski mountaineer to win the french national championship.
Table and KG	muriel vaudey -lrb- born february 14, 1976 -rrb- is a french ski mountaineer.

Figure 5: The left part shows table input and KG input. The generations from different models are shown on the right, and label text is shown in the first row of it. The text highlighted in red indicates wrong generation including false date or fictitious story.

the evaluation on BLEU, PARENT and semantic similarity score (STS-RoBERTa). To further verify the ability of the proposed method, we conduct the experiments on sentence-level text generation using WikiBio. Our method outperforms the best benchmark models with large margin on WikiBio with 7.3% (BLEU) and 9% (Rouge) increment.

References

Yang Bai, Ziran Li, Ning Ding, Ying Shen, and Hai-Tao Zheng. 2020. Infobox-to-text generation with tree-like planning based attention network. In *IJCAI*, pages 3773–3779.

Shuang Chen, Jinpeng Wang, Xiaocheng Feng, Feng Jiang, Bing Qin, and Chin-Yew Lin. 2019. Enhancing neural data-to-text generation models with external background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3022–3032.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Tianyu Liu, Fuli Luo, Pengcheng Yang, Wei Wu, Baobao Chang, and Zhifang Sui. 2019a. Towards comprehensive description generation from factual attribute-value tables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5985–5996.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13415–13423.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. Gpt-too: A language-model-first approach for amr-to-text generation. *arXiv preprint arXiv:2005.09123*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. 2020. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10317–10326.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. *arXiv preprint arXiv:1906.03221*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4498–4507.

Nils Smeuninx, Bernard De Clerck, and Walter Aerts. 2020. Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and nlp. *International Journal of Business Communication*, 57(1):52–85.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

A Appendix

A.1 Examples from TaKG-place and TaKG-school

We show the examples from TaKG-place and TaKG-school below.

1. Example from TaKG-place

- **Target Entity:** Alva
- **Fact Table:**
  - UK Parliament: Ochil and South Perthshire
  - Country: Scotland
  - Sovereign state: United Kingdom
  - ...
- **Knowledge Graph:**
  - Alva | population | 4,600 in 2016
  - Alva | area | 0.598 square mile
  - ...
- **Description:**
  - Alva (Scottish Gaelic: Ailbheach, meaning rocky) is a small town in Clackmannanshire, set in the Central Lowlands of Scotland. It is one of a number of towns situated immediately to the south of the Ochil Hills, collectively referred to as the Hillfoots Villages or simply The Hillfoots. It is located between Tillicoultry and Menstrie. Alva had a resident population of 5,181 at the 2001 census but this has since been revised to 4,600 in 2016. It boasts many features . . .

2. Example from TaKG-school

- **Target Entity:** St Bonaventure's
- **Fact Table:**
  - Established: 1877 (in Forest Gate)
  - Founder: Franciscans
  - Age: 11 to 18
  - ...
- **Knowledge Graph:**
  - St Bonaventure's RC School | country | United Kingdom
  - St Bonaventure's RC School | historic county | Essex
  - ...
- **Description:**
  - St Bonaventure's, known informally as St Bon's, is a voluntary-aided Catholic secondary school for boys aged 11–16 in Forest Gate, London Borough of Newham, England, with a mixed gender sixth form for 16–18-year-old students. It is under the trustee-ship of the Roman Catholic Diocese of Brentwood. St Bonaventure's is the oldest boys' school in Newham, having been established in the West Ham area of Essex by the Franciscan order in 1875, following the Roman Catholic Relief Act 1829. . . .

A.2 Heat map of attention weight

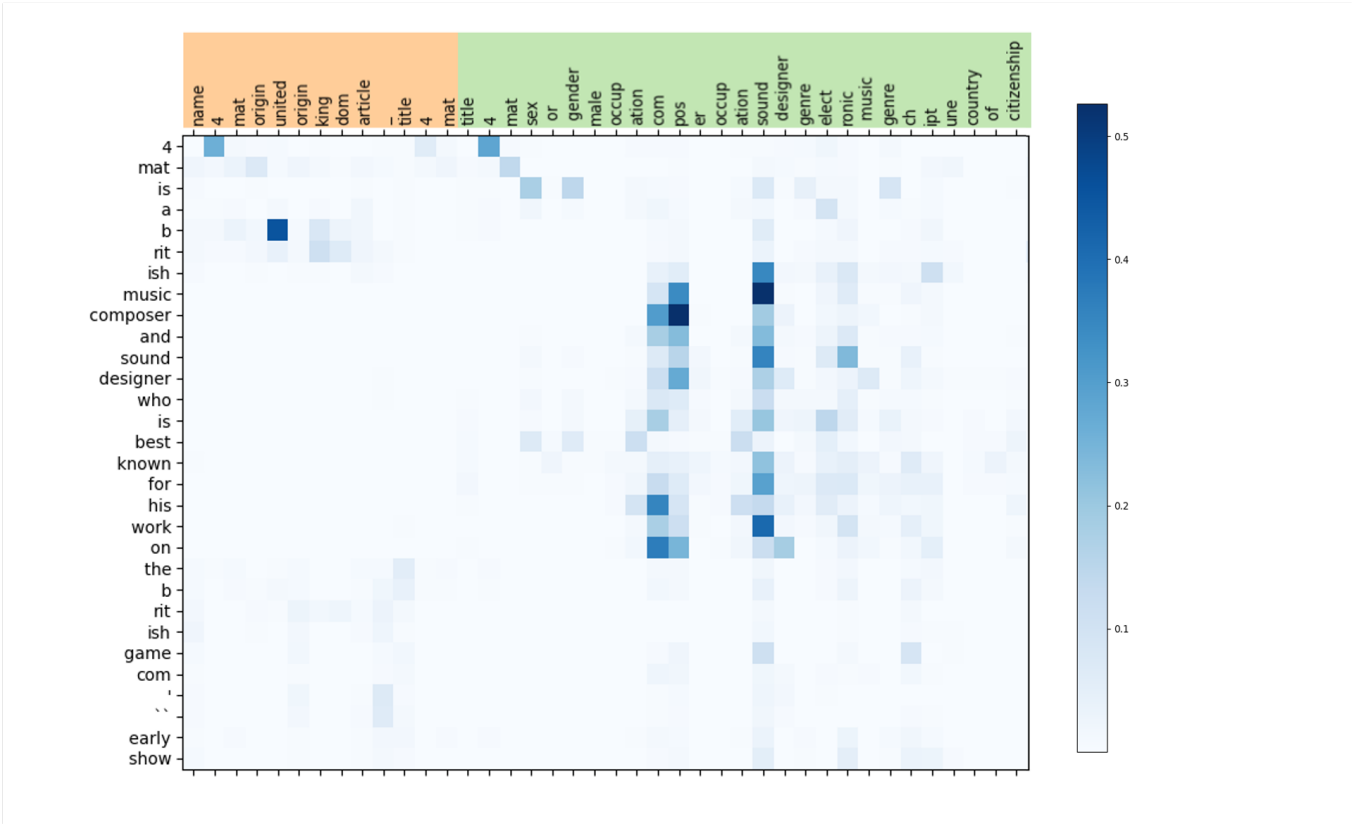

Figure 6: Tokens in orange colour represent table data, and tokens in green colour means KG data. From this heat map, the darker blue colour indicates, the more attention has been put.