Multilingual Table-to-Text Generation with Question-Answer Plans

Anonymous ACL submission

Abstract

Multilingual NLP is challenging due to the lack of training data for low-resource languages. However, some of these languages have up to tens of millions of speakers globally, making it important to improve NLP tools for them. Table-to-Text NLG is an excellent measure of models' reasoning abilities, but is very challenging in the multilingual setting. System outputs are often not attributable, or faithful, to the data in the source table. Intermediate planning techniques like Question-Answer (QA) blueprints have been shown to improve attributability on summarisation tasks. This work explores whether QA blueprints make multilingual Table-to-Text outputs more attributable to the input tables. The challenging multilingual Table-to-Text dataset, TATA, which includes African languages is extended with QA blueprints. Seq2seq models are then fine-tuned on this dataset, with and without blueprints. Results show that QA blueprints improve performance for models fine-tuned and evaluated only on English examples, but do not demonstrate gains in the multilingual setting. This is due to inaccuracies in machine translating the blueprints from English into target languages when generating the training data, and models failing to rely closely on the blueprints they generate. An in-depth analysis is conducted on why this is challenging.

1 Introduction

The majority of NLP research, models and datasets focus on English (Ruder, 2022). Yet there are many widely-spoken languages which are severely underrepresented (low-resource languages, or LRLs). For example, Igbo is a language found predominantly in Nigeria that is considered low-resource despite being spoken by around 44 million people.[1] It is important that more research is done to investigate techniques which make NLP work better for a more diverse set of languages.

The task is to generate **fluent** and **accurate** descriptions of the data in the input table. A seq2seq model takes the linearised table as input, and the goal is to output a verbalisation, like in Figure 1.
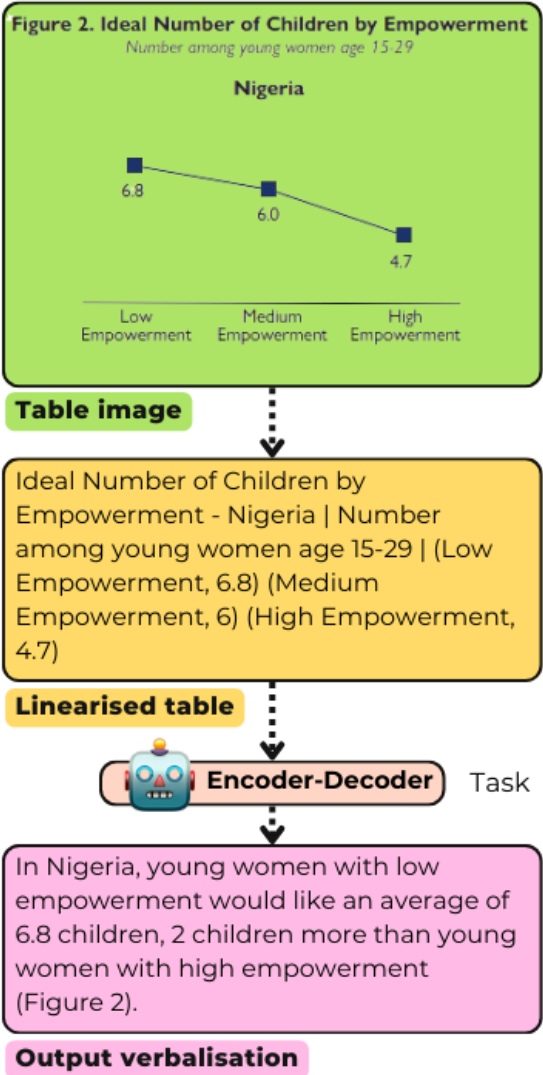

Figure 1: An example of a chart image, its linearised form (created by human annotators), and an output verbalisation (produced by a Seq2Seq neural network).

---

[1] https://celt.indiana.edu/portal/Igbo/index.html

Conditional generation with Question-Answer blueprints, a technique which has proven effective for improving the faithfulness of summaries, is applied to the task of Table-to-Text generation for the first time. Further, it is applied to a challenging multilingual Table-to-Text dataset containing African languages.

We evaluate whether QA blueprints improve the understandability (fluency), but in particular the attributability (faithfulness) of output verbalisations.

The results show that QA blueprints do increase attributability of outputs for models finetuned and evaluated on only English data. However, in the multilingual setting the technique is less effective. The challenge is twofold; inaccuracies in machine translating blueprints generated in English into target languages makes the training dataset less than perfect, baking in a fundamental disadvantage before training even begins. Furthermore, models struggle to produce verbalisations which rely heavily on their blueprints. This is analysed in detail in section 5.

We test the performance of a range of metrics on the task, such as FACTKB, and retrain the learned metric STATA introduced by Gehrmann et al. (2022) and make it publicly available, see subsection 4.3.

2 Related Work

Neural language models tend to hallucinate and repeat themselves, and they struggle to remain faithful to input data and identify important information. Solving these problems is challenging because deep neural networks are opaque by nature, making it difficult to understand their reasoning and find the root of errors (Narayan et al., 2023).

Numerous techniques have been applied in attempts to reduce hallucination and make outputs more faithful and attributable to the source document (Narayan et al., 2023). One family of such techniques is centred around content selection and planning, whereby the model is trained to identify and extract relevant information from the input, and generate an "intermediate plan" of *what to say* (selection) and *in what order* (planning) before generating the actual output, which is conditioned on the plan.

One idea is to build the intermediate plans from question and answer (QA) pairs, where the questions and answers are generated from the target

summaries. This has been shown to allow more control of a model's output, and more explanation of a model's features than entity chains (Narayan et al., 2023).

Content selection and planning has also proved effective in improving Table-to-Text outputs (Puduppully et al., 2019). In 2019 such a model achieved state-of-the-art BLEU scores on the RotoWire (Wiseman et al., 2017) and ToTTo (Parikh et al., 2020) datasets. This work used an LSTM-based encoder-decoder architecture with an attention mechanism and has since been surpassed by simply finetuning $T5_{XL}$ (Kale and Rastogi, 2020). This showed that the modern text-to-text pretraining-finetuning paradigm employed by models such as mT5 works very well for Table-to-Text tasks.

3 QA Blueprint Creation

3.1 Generating & Heuristically Selecting QA Pairs

A similar process for creating the QA blueprints is followed to that of Narayan et al. (2023), but with some differences. The following is done for each English reference (which are single sentences):

First, propositions are extracted from the reference. FLAN-T5-Large (Chung et al., 2022) finetuned[2] for "propositionizing" sentences by Chen et al. (2023) is used generate a minimal sentence for each proposition in the reference.

For example, for the following sentence:

In Nigeria, young women with low empowerment would like an average of 6.8 children, 2 children more than young women with high empowerment (Figure 2).

the following propositions are extracted:

- *In Nigeria, young women with low empowerment would like a average of 6.8 children.*

- *In Nigeria, young women with low empowerment would like to have an average of 2 more children than those with high empowerment.*

The QA pairs are generated directly from each proposition, instead of the reference itself. This guarantees that there will not be overlapping or redundant information, i.e. each QA pair expresses

---

[2]https://huggingface.co/chentong00/propositionizer-wiki-flan-t5-large

2

a different fact, assuming there is no information overlap between propositions.

For each proposition, 5 diverse QA pairs are generated using T5-Large finetuned[3] on SQuAD (Rajpurkar et al., 2016) for QA generation by Manakul et al. (2023).

Next, QA pairs where the question does not end with a question mark, or the question or answer are just empty strings, are dropped.

The goal of the QA blueprints is to focus the model's output on factual data which is attributable to the input table. If numbers are hallucinated at this point, this error will propagate into the final output verbalisation and defeat the blueprint's purpose. Therefore, pairs containing a hallucinated number, that is not present in the source table, are dropped.

QA pairs where the answer is fully contained within the question are dropped.

For any duplicated answers, the QA pair where the question has the greatest lexical similarity with the proposition (calculated via a word-level F1 score between the question and reference), is chosen. The other is dropped.

Finally, the QA pair with the greatest lexical similarity with the reference is selected. For each of the propositions in the above example respectively, these are:

- *"Question: In Nigeria, young women with low empowerment would like an average of how many children? Answer: 6.8."*

- *"Question: In Nigeria, how many more children would young women with low empowerment like to have than those with high empowerment? Answer: 2."*

On average, a QA blueprint generated via this strategy contains two or three QA pairs.

3.2 Building the Blueprint

The chosen QA pairs are then concatenated as follows: Let a denote an answer, and q denote a question. The blueprint b takes the form $a_1;q_1;...;a_n;q_n$. Although it is less natural, answers come before questions following the procedure by Narayan et al. (2023) that yielded superior results. Answers and questions are separated by a full stop ".", and QA pairs are separated by a pipe "|".

Special tokens *"Blueprint:"* and *"Verbalisation:"* are prefixed to the blueprint b and verbalisation v respectively, and these are concatenated to form the references.

In this setup, the encoder-decoder model takes the linearised table t as input and learns to predict the blueprint b as $p(b|t)$, then generate the output v as $p(v|t, b)$. It is noted by Narayan et al. (2023) that this relies on the blueprints being correct, and acknowledged that errors at this stage will propagate down the pipeline and affect the final generated verbalisation.

3.3 Making Blueprints Multilingual

Since the goal is to produce multilingual generations, two blueprint setups are created. The first just uses the English blueprints. So, when the model generates output v as $p(v|t, b)$, b is always in English and v is in the target language. The second still generates the blueprints in English but translates them to the target language. So, both b and v are in the target language. For the English blueprint dataset only, a language tag, for example *"(Yorùbá)"* is also inserted after *"Verbalisation"* to help indicate to the model what the target language for the verbalisation is.

Machine translating QA pairs from English to the other languages will propagate small errors in translation further down the pipeline and hinder the effectiveness of the blueprints. To evaluate this risk, an analysis of translation quality is done on the training set. Since the dataset is parallel (i.e., each sample appears in every language), there is ample data to test the quality of machine translations in this specific context.

Each English sample's reference is translated, using Google Translate,[4] into the seven other languages which appear in the training set, and these translations are compared to the corresponding samples written in the target languages with automatic MT metrics. The results are recorded in Table 1.

Translation quality for the more broadly spoken languages such as French and Portuguese is high, with lower quality for the African languages. This is to be expected, as mT5 will have seen more English and French samples during its training than, say, Igbo. Swahili enjoys the best translations of the African languages, achieving higher scores than even Arabic. All languages can be translated with

---

	Hausa	Igbo	Swahili	Yorùbá	French	Arabic	Portuguese
CHRF	0.53	0.56	0.68	0.19	0.80	0.57	0.66
BLEU	0.30	0.37	0.47	0.06	0.68	0.33	0.47

Table 1: CHRF and BLEU scores for Google Translate-powered translations of English samples in TATA compared to corresponding samples in the target languages.

reasonable quality, with the exception of Yorùbá, which achieves the lowest scores by a large margin.

See Appendix D for an example translated blueprint in Swahili.

3.4 Repetition Penalty

It is observed that the finetuned models are all highly susceptible to very repetitive outputs, especially non-English languages, and the effect is worsened when using blueprints. For example (repetition in bold): *"Blueprint: 31%. What percentage of women with high education are looking for an infant scale? | Verbalisation (Swahili): Miongoni mwa wanawake wadogo, 31% pekee ya wanawake walio na kipimo **cha juu cha juu cha juu cha juu...**"*

Excessive repetition in the blueprint sometimes results in the the actual verbalisation never being generated. This means the blueprint cannot be cut before comparing the output to the reference, which is highly undesirable as the blueprint itself is not meant to be evaluated against the reference.

Therefore, a repetition penalty is applied during inference (not during finetuning) to alleviate this. The repetition penalty is a form of penalised sampling, as formulated in Keskar et al. (2019) and simply penalises tokens which have already been generated. See Appendix C for details on how the penalty θ is chosen.

4 Automatic Evaluation

Human evaluators would produce the best judgements on model outputs for TATA. However, this is infeasible due to time, logistical and expense constraints. Evaluation is therefore done entirely through a combination of generic and learned automatic NLP metrics.

4.1 CHRF & BLEU

CHRF and BLEU are calculated and reported for the sake of interest. However, when interpreting them it must be kept in mind that they achieve extremely low correlations with human evaluations on TATA (<0.16) (Gehrmann et al., 2022).

Standard automatic evaluation metrics are so unsuitable for this task partly becasue of the reasoning required, often over multiple cells, which simply cannot be captured by these metrics, and also due to the massive verbalisation space, i.e., the number of "correct" verbalisations is utterly vast (especially for large tables) compared to, for example, the number of correct translations of some sentence from one language to another. This is why metrics which consider the table itself, such as PARENT (Dhingra et al., 2019) perform better on table-to-Text.

4.2 FACTKB

FACTKB differs fundamentally from metrics such as CHRF and BLEU as it was trained specifically to evaluate whether outputs are factually faithful to their sources.

FACTKB is intended to be used with *(summarisation, article)* input pairs. We explore whether, and to what extent, it can be used to produce scores for *(verbalisation, linearised table)* pairs. FACTKB uses RoBERTa$_{Base}$ (Liu et al., 2019), which is not multilingual. Therefore it can only be used to evaluate the models finetuned on English examples.

To determine whether FACTKB can be used to evaluate Table-to-Text on TATA, scores are computed on examples from the human evaluations dataset, and the Pearson correlation between the score and the human judgement of attributability (0 or 1) is calculated. FACTKB achieves a Pearson correlation with human evaluations of **0.22**. This outperforms the next-best existing metric, CHRF (0.16), but is still quite low.

BERTScore (Zhang et al., 2020) is also considered. It only achieves a Pearson correlation of **0.12** with human evaluations on TATA, which is lower than CHRF, so it is excluded.

4.3 STATA

Following instructions from Gehrmann et al. (2022), a learned metric is trained on human annotations of references and model outputs for TATA. It is called STATA, or *Statistical Assessment of Table-to-Text in African languages.*

The human annotations dataset marks outputs as "understandable" if they are fluent and grammatically correct, and "attributable" if they satisfy the "understandable" criteria, and also only contain data which correctly reflects the data in at least one cell in the input table.

Gehrmann et al. (2022) showed that STATA performs significantly better than existing metrics for evaluating TATA. Gehrmann et al. (2022) train three variants of the metric, but designate STATA$_{QE}$ (a quality estimation model which considers the linearised input and the candidate, but not the references) as the dataset's main metric as it has the highest Pearson correlation, 0.66, with the human assessments. Due to the computational constraints mentioned in Appendix A, we use mT5$_{Large}$ instead of mT5$_{XXL}$. This yields a Pearson correlation with human assessments of attributability of **0.59**, significant at the p < 0.01 confidence level. While not as good as if the metric had been trained with mT5$_{XXL}$, this is still far better than any existing metric. Since STATA$_{QE}$ was shown to perform the best, it is the only STATA variant we replicate.

It should be noted that the STATA scores reported cannot be directly compared to those from Gehrmann et al. (2022), as they are trained with different mT5 checkpoints. They should instead be interpreted in relation to one another, and CHRF can be used for comparison with Gehrmann et al. (2022) if desired.

5 Results

5.1 English Subset Results

In Table 2, CHRF and BLEU are calculated using the predicted and reference verbalisations (traditional reference-based metrics), and FACTKB and STATA are calculated using the predicted verbalisation and the linearised input table (Quality Estimation, or reference-less, metrics).

The results show that the blueprints do improve the attributability of outputs for the model finetuned on English examples. For the Small model, STATA increases from 0.513 to 0.523. This is significant because the range of values outputted by STATA is relatively narrow; from about 4.9 to 7.2 based on observation. This narrow range is common for learned metrics, and means that an increase of 0.1 is quite large. If the metric is trained with a larger model, the range of observed values will increase as the model becomes more confident. FACTKB also increases by 0.04 for the blueprint model.

mT5$_{Large}$ is also finetuned on the English subset, however it overfits very quickly and does not achieve representative results, due to the dataset's tiny size. The full TATA dataset is small to begin with, and the English subset (about eight times smaller than the whole dataset, at 902 training examples) appears to simply be too small to finetune the Large model on. Therefore, these results are discarded and not reported.

5.1.1 Examples & common patterns

Some frequent idiosyncrasies emerge upon manually examining model outputs. Models will commonly produce phrases like *"increased from 15 percent to 15 percent"*. This is a common verbalisation pattern in the training data which the model learns, but it is clear that it has not learnt what *"increase"* means in this context, as the numbers are the same. Similar mistakes in characterising a comparison between two numbers occur relatively often, even if both numbers do appear in the input table. For example, *"mortality rate is 19, compared to 19"* This is indicative of a lack of reasoning. Note that the repetition penalty, which is applied, does not stop this.

The majority of verbalisations begin with *"The percentage of..."*, *"The proportion..."* or a few other common phrases. This is just a reflection of common sentence structures in the training data and is not an issue per se, but it does raise the concern that even a more powerful and capable model will probably not generate highly diverse verbalisations from this dataset.

Verbalisations are also sometimes not fully formed. For example, *"The proportion of children under age 5 who are wasted or too short for their age."* or *"Although the proportion of women who have received antenatal care from a skilled provider."* These would be valid as the first half of verbalisations, but are not completed and therefore do not make sense.

Generally, the title and unit parts of the input tables seem to appear the most consistently and accurately in the verbalisations, with the actual data points less so. This is consistent with an observation made by Gehrmann et al. (2022).

5.2 Multilingual Results

The Small baseline model trained by Gehrmann et al. (2022) achieves a CHRF of 0.33, which is very close to the 0.32 achieved by our baseline Small model (Table 3), so the results are considered to be

Model	CHRF	BLEU	FACTKB	STATA
Small	0.33	0.15	0.24	0.513
Small (Blueprints)	0.30	0.11	0.28	0.523

Table 2: Results for mT5$_\text{Small}$ finetuned and evaluated on the English subset of TATA.

Model	CHRF	BLEU	STATA
Small	0.32	0.16	0.552
Small (Eng blu.)	0.29	0.09	0.525
Small (Trans blu.)	0.30	0.12	0.542
Large	0.33	0.13	0.552
Large (Eng blu.)	0.24	0.04	0.519
Large (Trans blu.)	0.27	0.11	0.544

Table 3: Results of finetuned multilingual mT5 models on the test set (all languages). Models with "blu." are models with blueprints.

replicated.

The English blueprint setup performs poorly (Table 3). Inspecting outputs of models trained on the English blueprints, which have to be able to generate outputs that contain multiple languages (English and the target language) reveals that they sometimes mix languages up. In the following example, colours denote different languages: *"Among Tanzania, one-third of Tanzanian women in the United States had all the three or more antenatal care visits, asilimia 29 ya wanawake in the United States had all the three or more antenatal care visits."*

"asilimia 29 ya wanawake" is a Swahili phrase meaning *"29 percent of women"* which has been incorrectly generated in the middle of an English sentence. (In this example, the United States is also hallucinated and a clause is repeated, but those are separate issues).

The translated blueprints fare decisively better than the English ones across all metrics, but are still slightly worse than no blueprints (Table 3).

5.2.1 Model size

Gehrmann et al. (2022) saw a large performance jump on TATA from mT5$_\text{Small}$ to mT5$_\text{XXL}$ (13B). However, we do not observe Large (1.2B) improving over Small. One likely hypothesis is model-wise double descent, the phenomenon whereby performance degrades as the model size increases to a point, then begins to improve again as the model size is increased further (Nakkiran et al., 2019). Even the full TATA dataset is small, and the Large model converges quickly or overfits. Increasing the model size by a factor of 10 would still result in a fast convergence, but the model will likely achieve a much lower loss before this happens. We are unable to test this hypothesis due to compute restraints.

5.3 Per-Language Analysis

For more granular insights, the multilingual models are evaluated on each language in the test set individually. (The Blueprint columns in this table are translated blueprints as these were shown to perform better in Table 3). These per-language evaluation results are noisy (Table 4). Broadly, blueprints rarely improve any of the metrics, and performance between the Small and Large model is very close.

The lower-resource African languages also tend to benefit the most from an increase in model size. Igbo, Hausa and Yorùbá all achieve higher STATA scores with the Large model, and are the only languages to do so. This suggests that scale is of particular importance for the low-resource languages.

5.4 Blueprint Analysis

Table 5 shows how closely predicted blueprints match reference blueprints in the dev set (the candidates and references are split on *"Verbalisation:"* and only the blueprints compared). The dev set is used as no blueprints are generated for the test set, because at test time only the verbalisations are compared.

These scores are very low. It should be noted that achieving a high CHRF or BLEU on the blueprints, or indeed the entire output, is not the explicit goal of training. There are many valid blueprints for any given table, especially when generating short verbalisations from large tables, where there is lots of data to choose from. Arguably more important, is whether blueprints are related to the input table, and whether verbalisations are related to their blueprints.

Table 6 provides a measure of this. It shows how well the models are able to generate blueprints, and to what extent the verbalisations use these blueprints. To do this, CHRF and BLEU are calculated between the linearised input and the blueprint

6

Lang	Small	Small Blueprints	Large	Large Blueprints
En	**0.19** / 0.33 / **0.551**	0.15 / 0.34 / 0.529	0.17 / **0.37** / 0.538	0.10 / 0.29 / 0.549
Sw	**0.21** / **0.39** / **0.589**	0.17 / 0.36 / 0.569	0.16 / 0.38 / 0.581	0.13 / 0.30 / 0.585
Yo	*0.03* / 0.13 / 0.567	*0.03* / *0.14* / 0.563	*0.03* / *0.14* / **0.577**	0.02 / *0.14* / 0.561
Fr	**0.17** / 0.36 / 0.528	0.11 / 0.33 / 0.526	0.14 / **0.38** / 0.526	0.12 / 0.31 / **0.529**
Pt	**0.17** / *0.39* / 0.527	0.16 / 0.34 / 0.518	0.15 / *0.39* / 0.512	0.15 / 0.32 / **0.531**
Ha	**0.17** / *0.33* / 0.526	0.12 / *0.33* / 0.523	0.12 / *0.33* / **0.546**	0.12 / 0.29 / 0.515
Ar	**0.14** / 0.32 / **0.539**	0.11 / *0.33* / 0.523	0.12 / *0.33* / 0.533	0.12 / 0.31 / 0.519
Ig	**0.20** / **0.35** / 0.596	0.17 / 0.32 / 0.584	0.16 / 0.34 / **0.605**	0.15 / 0.27 / 0.558

Table 4: Language-specific performance of mT5 multilingual models (BLEU / CHRF / STATA). Bold figures represent the best result for each metric in each row. Italic figures are ties.

Model	CHRF	BLEU
Small Trans Blueprints (Multilingual)	0.27	0.07
Small Blueprints (English)	0.23	0.05

Table 5: CHRF and BLEU between predicted and gold blueprints on dev set.

to quantify how much information from the table is present in the blueprint. CHRF and BLEU are also calculated between the blueprint and the verbalisation to quantify how closely the output relies on the blueprint for content selection.

Again, note that the goal is obviously not to have blueprints which are exactly the same as the input tables, nor verbalisations which are exactly the same as the blueprints. So, just maximising these metrics is not desirable here. The models' scores are to be interpreted with respect to the dataset scores, as an indicator the extent to which the models exhibit desirable attributes of the dataset.

The **English dataset** and **Multilingual dataset** rows in Table 6 represent scores calculated directly on the respective training datasets, and showcase the best-case scores. The **English model** (mT5$_{Small}$ finetuned on the English subset with blueprints) and **Multilingual model** (mT5$_{Small}$ finetuned on the full dataset with translated blueprints) rows represent scores calculated on the respective test sets, and the models' generated blueprints and verbalisations (the generation is split on *"Verbalisation:"* and the blueprint and verbalisation separated).

Even the English model is not able to generate blueprints that are as closely related to the linearised table as the references (CHRF of 0.24 versus 0.28). Furthermore, the similarity between the blueprints and the outputs is very low (CHRF 0.39, BLEU 0.20) compared to the dataset (CHRF 0.61, BLEU 0.24). This indicates that not only is the model struggling to generate blueprints that are as good as gold, but it is also failing to remain as

faithful to its blueprint (as showcased in the second example in 8). In other words, the model's verbalisations have significantly less in common with its blueprints than the dataset's.

The multilingual model actually does a better job of producing verbalisations which rely on the blueprints. There is a smaller percentage drop in **Blueprint → Verbalisation** CHRF than for English, and BLEU actually improves slightly. It is clear from these findings that the model's verbalisations do not draw from its blueprints enough, as exemplified in the examples in Table 8. One way of encouraging the model to rely more on its blueprints would be to use a form of constrained decoding, which could help focus the model on using words which it generates in its blueprint.[5] This is not explored and is left to future work.

Still, there is a fundamental disadvantage in the multilingual setup before training even begins. Note how much lower the multilingual dataset's **Blueprint → Verbalisation** score (CHRF 0.43, BLEU 0.13) is than the English dataset's (CHRF 0.61, BLEU 0.24). This indicates that, as predicted in subsection 3.3, inaccuracies in translating the English blueprints into the target languages have made it challenging to create high-quality multilingual blueprints.

So, why do the multilingual blueprint models achieve slightly higher STATA scores than the English-only blueprint models if the multilingual blueprints are lower quality? These multilingual

---

[5]https://huggingface.co/blog/constrained-beam-search

| | Linearised input → Blueprint | | Blueprint → Verbalisation | |
	CHRF	BLEU	CHRF	BLEU
English dataset	0.28	0.02	0.61	0.24
English model	0.24	0.02	0.39	0.20
Multilingual dataset	0.25	0.02	0.43	0.13
Multilingual model	0.23	0.01	0.36	0.16

Table 6: CHRF and BLEU between linearised inputs, blueprints and verbalisations in the training data and model outputs.

models are still trained on around eight times more training data, as they see the full dataset, not just the English subset. This allows the multilingual models to learn the general task better, despite the language-specific challenges, and is also the reason the multilingual model achieves higher BLEU and CHRF between predicted and reference blueprints in Table 5.

6 Conclusions

The evidence suggests QA blueprints are effectively improve the attributability of Table-to-Text outputs in English. More work needs to be conducted to validate this, with both a larger English Table-to-Text dataset such as ToTTo, and a larger model, such as mT5$_{XXL}$ or the newly-released Gemma-7b.[6]

In the multilingual setup, English blueprints degrade performance significantly and sometimes cause models to mix up multiple languages in their verbalisations.

Translated blueprints fare better, but still worse than no blueprints at all. This is due to inevitable inaccuracies in machine translating the blueprints from English to the target languages in the dataset for training. These imperfect translated blueprints mean that the multilingual models do not have high-quality gold examples to learn from.

As a result, the models struggle to generate blueprints which are as closely related to the input tables as the dataset. This problem, as measured by BLEU, is more severe for multilingual models than English.

For the English results, an increase in STATA scores is observed, while generic automated metrics such as BLEU and CHRF decrease. This further confirms that these metrics are not suitable for evaluating TATA due to their very low correlations with human evaluations. Although FACTKB performs

slightly better, is not recommended for evaluating TATA in future work, as its correlation with humans is still quite low, and this is not the task it was designed for. The learned metric STATA should be the definitive judge of model performance on TATA when human evaluators are not available. We release our version of STATA online in the hope that it will make doing research on TATA more accessible. However, this version is trained with mT5$_{Large}$. It should ideally be retrained with mT5$_{XXL}$, released and standardised, else comparing scores across papers will be impossible.

It is also observed that increasing model size results in larger gains for the low-resource languages.

Multilingual Table-to-Text generation remains a very challenging task for neural models. Based on our findings, recommendations for future work are as follows: Trial using LLMs to generate more synthetic training data in several languages. If high-quality synthetic data can be generated, and the dataset size increased, this will boost model performance (although LLMs will probably only be able to generate good examples in the higher-resource languages). Additionally, constrained decoding could also be explored as a way to make verbalisations utilise the blueprints more. Finally, TATA can be turned into a more constrained task by having human annotators highlight the cells in each table that are used in its reference verbalisation, as the ToTTo dataset does. This greatly reduces the valid answer space, especially for larger tables.

7 Acknowledgements

---

[6]https://blog.google/technology/developers/gemma-open-models/

References

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. Dense x retrieval: What retrieval granularity should we use? *Preprint*, arXiv:2312.06648.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

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. *Preprint*, arXiv:1906.01081.

Sebastian Gehrmann, Sebastian Ruder, Vitaly Nikolaev, Jan A. Botha, Michael Chavinda, Ankur Parikh, and Clara Rivera. 2022. Tata: A multilingual table-to-text dataset for african languages. *Preprint*, arXiv:2211.00142.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

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

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization. *Preprint*, arXiv:2301.12307.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2019. Deep double descent: Where bigger models and more data hurt. *CoRR*, abs/1912.02292.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Preprint*, arXiv:2207.00397.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sebastian Ruder. 2022. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Model Training Details

Gehrmann et al. (2022) use mT5$_{\text{Small}}$ and mT5$_{\text{XXL}}$. However, mT5$_{\text{XXL}}$, and indeed mT5$_{\text{XL}}$ are both massive models (13B and 3.7B parameters respectively). Available compute resources made finetuning these models infeasible. The largest model that could be trained is mT5$_{\text{Large}}$ (1.2B), so this is used instead.

Both mT5$_{\text{Small}}$ and mT5$_{\text{Large}}$ are finetuned with the following hyperparameters and setup: a constant learning rate of 0.001, dropout of 0.1, per-device batch size of 4, 5 epochs for the Small model and 3 for Large, weight decay of 0.001 for Large, linearised tables (inputs) and references are both truncated to 512 tokens, and validation loss is measured every 100 or 500 steps depending on batch size and after training the checkpoint with the lowest loss is selected.

B StATA Training Details

mT5$_{Large}$ and mT5$_{Small}$ are finetuned for regression with an RMSE (Root Mean Squared Error) loss function. The metric is released on the Huggingface Hub (https://huggingface.co/adenhaus/mt5-large-stata) in the hope that it will prove useful to other researchers who wish to work on TATA.

A spare token in mT5's vocabulary is chosen as the regression token. At each step, the RMSE of this token's logit and the label, which is 0 or 1, is taken. This is the loss.

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \quad (1)$$

In Equation 1, $\hat{y}$ is the logit of the special token and y is the label. N is the number of observations. At inference time, generation is constrained to this token, and its logit x is converted into a probability with the sigmoid, or logistic function (Equation 2). This is the final metric.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

The human annotations file contains rows with the following fields and data types: output (string), model (string), interpretable (float), attributable (float), cells (float), reasoning (float), id (string), set (string), lang (string) and linearized_input (string).
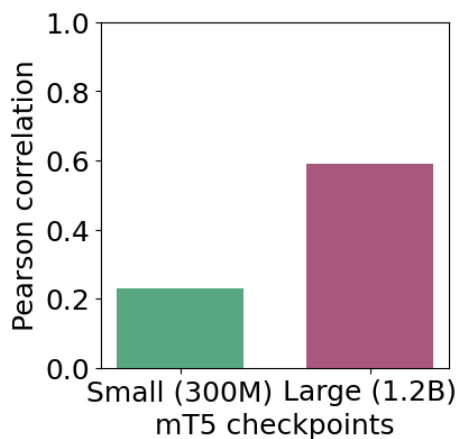
Figure 2: Pearson correlations of mT5-Small and mT5-Large STATA models with U+A human evaluations.

The file required some cleaning before it could be used for training. All rows where "attributable" is not either 0.0 or 1.0 are dropped. 73.7% of samples have a 0 in the "attributable" column, and 26.3% have a 1. This phenomenon, where the dataset is unbalanced in favour of the negative class, is common in classification or similar datasets. However, since the model learns the data quite effectively as-is, no data augmentation or other techniques are applied to balance the classes. The file is then randomly sampled into train (80%), validation (10%) and test (10%) splits. This makes the training set 4,900 rows and the validation and test sets 612 rows each. This dataset is also released on the Huggingface Hub (https://huggingface.co/datasets/adenhaus/stata) so that STATA can be easily retrained by other researchers, particularly those with more compute resources who can train it with mT5$_{XXL}$.

To evaluate the metric, scores are computed on the test set and the Pearson correlation between them and the human assessments of attributability (1, or 0) is computed. Figure 2 shows the correlations for the Small and Large model.

B.1 Hyperparameters & other details

Models are finetuned with the following setup: a constant learning rate of 0.0001, per-device batch size of 32 for Small and 8 for Large, 15 epochs, dropout of 0.1, inputs are truncated to 2048 tokens, validation loss is measured every 0.1 epochs, and after training the checkpoint with the lowest RMSE loss is selected.

C Repetition Penalty Details

This is simply Huggingface's *repetition_penalty*[7] parameter which can be passed to *model.generate*.

An analysis of varying repetition penalties is performed on the English-finetuned mT5$_{Small}$ with blueprints, as the model and dataset are small which allows for a faster analysis. Figure 3 shows that as the repetition penalty is increased, performance on all metrics increases until 1.4, after which the metrics begin to decline. A θ of 1.0 is equivalent to no penalty. However, by inspecting outputs it is observed that a penalty of 1.2 is enough to stop most of the highly problematic repetitive outputs. The penalty is somewhat a blunt instrument, so it is not over-used here. Although the scores are higher, a penalty of 1.2 is cautiously chosen for the rest of the experiments as it mitigates the main problem, without altering the normal generations excessively. This agrees with the findings of

---

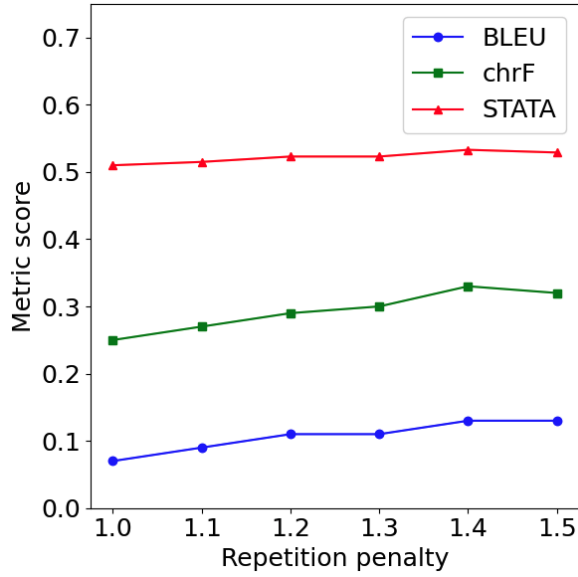[7] https://huggingface.co/docs/transformers/internal/generation_utils

Figure 3: Multilingual model performance on the test set (including all languages) with varying repetition penalties.

Keskar et al. (2019), who also note that a θ of 1.2 strikes a good balance between keeping repetition to a minimum but maintaining fluent and sensible outputs.

Where highly repetitive blueprints do still occur at test time, these obviously bring down the metrics slightly, as a blueprint is being compared to a reference verbalisation which is not intended. But these bad candidates are not removed.

D Blueprint Translation Example

The blueprint, generated in English (see Appendix E for full examples), is:

> 13%. How much of young women in Mali are in the highest tercile for empowerment? |81%. What is the percentage of young women in the Philippines who are in the highest tercile for empowerment? |

This is translated into Swahili as:

> 13%. Je, ni idadi gani ya wanawake vijana nchini Mali walio katika eneo la juu zaidi la kuwezeshwa? | 81%. Je, ni asilimia ngapi ya wanawake vijana nchini Ufilipino walio katika eneo la juu zaidi la kuwezeshwa? |

Translated back into English (again, with Google Translate) to check the quality:

> 13%. What is the proportion of young women in Mali in the highest area of empowerment? | 81%. What percentage of young women in the Philippines are in the highest empowerment zone? |

This is relatively close to the original English blueprint, however it loses the term "tercile" and replaces it with "area" or "zone", which could have many different meanings (e.g. spatial) that do not closely relate to "tercile". Swahili is one of the languages with better translation performance. For those with even poorer translations, the blueprints will not match perfectly to the verbalisations, which are translated by expert humans (Gehrmann et al., 2022).

E Blueprint Examples

See Table 7.

F Output Examples

In the first example in Table 8, neither the vanilla nor the blueprint verbalisations are very good in terms of understandability or attributability. Both refer to the correct concepts, but the former makes no reference to any actual data, and the second entirely hallucinates the data. The blueprint is semi-correct, as "wanted then" is the most common category, at 0.57. Unfortunately, this data is not correctly referenced in the verbalisation at all.

In the second example, a very good blueprint has been formulated. It captures important information from the table and does so correctly, identifying the right figure (88%) and year (2003). However, again the verbalisation fails to use this blueprint effectively, mentioning the correct year, but not the figure. This is still an improvement over the vanilla verbalisation, which references neither.

11

Dataset	Source verbalisation	Generated blueprint	New reference
English	Only 13% of young women in Mali are in the highest tercile for empowerment compared with 81% of young women in the Philippines.	13%. How much of young women in Mali are in the highest tercile for empowerment? |81%. What is the percentage of young women in the Philippines who are in the highest tercile for empowerment? |	Blueprint: 13%. How much of young women in Mali are in the highest tercile for empowerment? |81%. What is the percentage of young women in the Philippines who are in the highest tercile for empowerment? |Verbalisation: Only 13% of young women in Mali are in the highest tercile for empowerment compared with 81% of young women in the Philippines.
Translated blueprints	Asilimia 13 pekee ya wanawake wadogo katika Mali ndio wapo katika kikundi cha juu cha uwezeshaji ikilinganishwa na asilimia 81 ya wanawake wadogo katika Ufilipino.	13%. Je, ni idadi gani ya wanawake vijana nchini Mali walio katika eneo la juu zaidi la kuwezeshwa? |81%. Je, ni asilimia ngapi ya wanawake vijana nchini Ufilipino walio katika eneo la juu zaidi la kuwezeshwa? |	Blueprint: 13%. Je, ni idadi gani ya wanawake vijana nchini Mali walio katika eneo la juu zaidi la kuwezeshwa? |81%. Je, ni asilimia ngapi ya wanawake vijana nchini Ufilipino walio katika eneo la juu zaidi la kuwezeshwa? |Verbalisation: Asilimia 13 pekee ya wanawake wadogo katika Mali ndio wapo katika kikundi cha juu cha uwezeshaji ikilinganishwa na asilimia 81 ya wanawake wadogo katika Ufilipino.
English blueprints	Seules 13% des jeunes femmes au Mali se situent dans le tercile le plus élevé en matière d'autonomisation, contre 81% des jeunes femmes aux Philippines.	13%. How much of young women in Mali are in the highest tercile for empowerment? |81%. What is the percentage of young women in the Philippines who are in the highest tercile for empowerment? |	Blueprint: 13%. How much of young women in Mali are in the highest tercile for empowerment? |81%. What is the percentage of young women in the Philippines who are in the highest tercile for empowerment? |Verbalisation (French): Seules 13% des jeunes femmes au Mali se situent dans le tercile le plus élevé en matière d'autonomisation, contre 81% des jeunes femmes aux Philippines .

Table 7: Example blueprints from the three dataset setups for the same input table with English, Swahili and French as target languages.

Reference	Linearised input	Setup	Blueprint	Verbalisation
17 percent of births in Kenya are unwanted, and 26 percent are mistimed (wanted later).	Planning Status of Births |Percent |(Wanted then, 0.57) (Unwanted, 0.17) (Wanted later, 0.26)	Vanilla		The proportion of births were wanted at the time of delivery.
		Blueprint	wanted births. What was the most common form of births?	The proportion of births were wanted at births, from a low of 7% in 1990 to a high of 11% in 2012.
The 2008-09 data indicate a rise since 2003 in medical antenatal care coverage.	Trends in Receipt of Antenatal Care from a Skilled Medical Provider , Kenya 2003-2008 |Percentage of women with live birth in the past 5 years |(2003, 88) (2008-09-01 00:00:00, 92)	Vanilla		Although the proportion of women who have received antenatal care from a skilled provider.
		Blueprint	88%. What was the rate of antenatal care from a skilled provider in 2003?	The proportion of women with antenatal care from a skilled provider in 2003.

Table 8: Output examples from vanilla and blueprint models, colour-coded to show where relevant information from the table has been used in blueprints and verbalisations.