Centre for Research Training
Science Foundation Ireland sfi For what's next

# CURED4NLG: A Dataset for Table-to-Text Generation

Nivranshu Pasricha Mihael Arcan Paul Buitelaar
n.pasricha1@universityofgalway.ie

CRT·AI
SFI Centre for Research Training
in Artificial Intelligence

## Abstract

We introduce CURED4NLG (COVID-19 Update Reports from Epidemiological Data for Natural Language Generation), a dataset for the task of table-to-text generation focusing on the public health domain. The dataset consists of 280 pairs of tables and documents extracted from weekly epidemiological reports published by the World Health Organisation (WHO). Each table comprises of 6 to 60 rows with 7 to 9 columns and reports the number of new cases of COVID-19 and related deaths during a week-long time period along with cumulative totals recorded since the start of the pandemic. A document corresponding to each table describes the important information contained in the table in about 200 – 300 words in English as shown below. Along with the releasing the dataset, we present baseline outputs from two different end-to-end transformer-based models for the task of table-to-text generation. The dataset and all the sample outputs are available at `http://github.com/CURED4NLG/CURED4NLG`.

## Dataset Example and Sample Outputs

| Reporting Country/ Territory/ Area | New cases in last 7 days | Cumulative cases | Cumulative cases per 100k population | New deaths in last 7 days | Cumulative deaths | Cumulative deaths per 100k population | Epidemiological Report |
|---|---|---|---|---|---|---|---|
| Europe | 1466680 | 50714995 | 5435.3 | 25341 | 1061218 | 113.7 | The European Region reported over 1.4 million new cases and over 25000 new deaths, a 12% and a 5% decrease respectively compared to the previous week. The trend in both new cases and deaths is decreasing in the last two weeks. The highest numbers of new cases were reported from Turkey (378771 new cases; 449.1 new cases per 100000; a 9% decrease), France (211674 new cases; 325.5 new cases per 100000; a 9% decrease), and Germany (145156 new cases; 174.5 new cases per 100000; a 1% increase). The highest numbers of new deaths were reported from Poland (3383 new deaths; 8.9 new deaths per 100000; a 6% decrease), the Russian Federation (2650 new deaths; 1.8 new deaths per 100000; a 2% increase), and Ukraine (2537 new deaths; 5.8 new deaths per 100000; an 8% decrease). |
| Turkey | 378771 | 4591416 | 5444.0 | 2403 | 38011 | 45.1 | |
| France | 211674 | 5390187 | 8287.6 | 2110 | 102031 | 156.9 | |
| Germany | 145156 | 3287418 | 3952.8 | 1650 | 81564 | 98.1 | |
| Italy | 92074 | 3949517 | 6622.1 | 2345 | 119021 | 199.6 | |
| Ukraine | 78761 | 2025271 | 4630.9 | 2537 | 42323 | 96.8 | |
| Poland | 70831 | 2758856 | 7268.2 | 3383 | 65415 | 172.3 | |
| ... | ... | ... | ... | ... | ... | ... | |
| Baseline Outputs<br>T5 Model (left)<br>Hierarchical Model (right) | In the past week, the *European Region* reported over *1466000* new cases and over *25000* new deaths, a *decrease* of 1% and an increase of 1% respectively compared to the previous week. The three countries reporting the highest numbers of new cases were Kosovo (*2662* new cases; 57 new cases per 100000; a 1% decrease), *Turkey* (*378771* new cases; 57 new cases per 100000; a 1% *decrease*), *France* (*211674* new cases; 158.8 new cases per 100000; a 7% *decrease*). The three countries reporting the highest numbers of new deaths this week were the United Kingdom (*157* new deaths; 3.4 new deaths per 100000; a 3% *decrease*), Germany (*1650* new deaths; 3.4 new deaths per 100000; a 3% *decrease*), the *Russian Federation* (*2650* new deaths; 3.7 new deaths per 100000; a 3% *decrease*) and the *Russian Federation* (2345 new deaths; 3.4 new deaths per 100000; a 3% decrease). | | | | | | The *European Region* reported over 1.2 million new cases and over 36000 new deaths, a *decrease* of 18% and 8% respectively when compared to the previous week. The three countries reporting the highest numbers of new cases were Spain (7666 new cases; 160.5 new cases per 100000, a 15% decrease), *France* (*211674* new cases; 113.2 new cases per 100000; a 31% *decrease*), and *France* (3958 new cases; 2.4 new cases per 100000, a 119021% increase). The highest numbers of new deaths were reported from the *Russian Federation* (2650 new deaths; 0.4 new deaths per 100000 population; a 88% *decrease*), Germany (214 new deaths; 0.1 new deaths per 100000, a 3848% *decrease*) and France (2110 new deaths; 0.4 new deaths per 100000; a 88% *decrease*). |

Table: Example of a table and corresponding epidemiological report for the European region along with sample outputs generated by the T5 model and the Hierarchical model from the CURED4NLG dataset. Text in blue shows the correct names and numbers present in the input table and the reference output. The text in green italics shows tabular values correctly produced by the end-to-end baseline models while underlined text in red shows the errors in the generated texts. Any hallucinations or repetitions generated are highlighted in purple.

## Baseline Results

Initial baseline results suggest that end-to-end text generation models can learn a template-like structure of the documents to generate fluent outputs but at the same time are prone to hallucinating and generating erroneous statements particularly related to numerical values.

| | | | | PARENT | | | Average Error Count | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU (↑) | METEOR (↑) | TER (↓) | Precision (↑) | Recall (↑) | F1 (↑) | Total (↓) | Number (↓) | Name (↓) | Word (↓) | Other (↓) |
| Template baseline | 64.48 | 41.76 | 32.19 | 76.55 | 19.93 | 29.97 | – | – | – | – | – |
| Hierarchical model | 29.86 | 27.64 | 67.49 | 43.10 | 17.65 | 22.80 | 23.9 | 16.5 | 3.0 | 4.0 | 0.4 |
| T5 model | 43.32 | 32.77 | 52.10 | 56.38 | 17.15 | 24.68 | 19.3 | 14.0 | 1.6 | 3.4 | 0.3 |

Table: Results from automatic evaluation metrics (BLEU, METEOR, TER and PARENT) and human evaluation (Average Error Count) for the baselines on the CURED4NLG dataset.

## Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223 and co-supported by Science Foundation Ireland under grant number SFI/12/RC/2289 2 (Insight), co-funded by the European Regional Development Fund.

CURED4NLG

HOST INSTITUTION
UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

PARTNER INSTITUTIONS
University of Limerick Ollscoil Luimnigh
Ollscoil na Gaillimhe University of Galway
DCU
Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin