Tackling Hallucinations in Neural Chart Summarization

Master thesis Seminar

Presented By: Saad Obaid
Supervisors: Vera Demberg, Ondrej Dusek, Iza Skrjanec

Contents

1. Natural Language Generation
2. The Problem of Hallucinations
3. Neural Chart Summarization
4. Text-to-Text Transfer Transformer
5. Automatic Evaluation Metrics
6. Identifying problems, experiment and results
7. Further Work

Natural Language Generation (NLG)

- Process to produce text to meet specified goals.
  - Image Captioning
  - Machine translation
  - Abstractive summarization
  - Story generation
  - Data-to-text
- Recent advances: GPT-2, T5, BART, etc.

Data-to-text

- Data-to-text generation refers to the task of generating textual output from non-linguistic input - (Reiter and Dale, 1997)
  - Task oriented dialog systems (Wen et al., 2015)
  - Weather forecast generator (Sripada et al., 2003)

Table Title: Cristhian Stuani
Section Title: International goals

No.	Date	Venue	Opponent	Result
2	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	5-0

<page_title> Cristhian Stuani </page_title>
<section_title> International goals </section_title>
<table> <cell> 2. <col_header> No. </col_header> </cell>
<cell> 13 November 2013 <col_header> Date </col_header>
</cell> <cell> Amman International Stadium, Amman,
Jordan <col_header> Venue </col_header> </cell> <cell>
Jordan <col_header> Opponent </col_header> </cell>
<cell> 5-0 <col_header> Result </col_header> </cell>
</table>

On 13 November 2013 Cristhian Stuani netted the second in a 5–0 win in Jordan.

Source: ToTTo: A controlled table-to-text generation dataset.

The Problem of Hallucinations

Psychological term

- An experience involving the apparent perception of something not present.

Hallucinations in NLG

- In the context of NLG, Hallucination means generating unfaithful or meaningless text.
- Formally defined for automatic summarization task - Maynez et al. (2020):
    - A summary ***S*** of a document ***D*** contains a hallucination if it contains information not found in ***D*** that is factually correct.

Types of hallucinations

- Intrinsic:
  - Generated output that contradicts the source sentence.
- Extrinsic:
  - Generated output that cannot be verified from a source content.

Segment	Value of assets in billion U.S. dollars
Renewable energy	15.94
Power	26.73
Healthcare	30.51
Aviation	41.65
Capital	117.55

Title: General Electric 's total assets in FY 2019 , by segment (in billion U.S. dollars)

This statistic represents General Electric 's total assets in the fiscal year of 2019 , with a breakdown by segment . In its healthcare segment , the company had assets to the value of around 91 billion U.S. dollars .

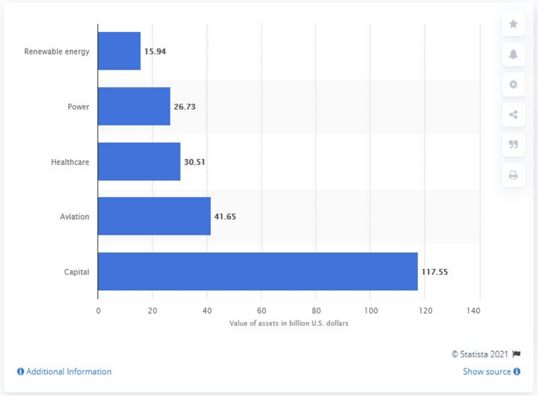
This statistic represents General Electric 's total assets in the fiscal year of 2019 , with a breakdown by segment . In its healthcare segment , the company had assets to the value of around 30 billion U.S. dollars . General Electric Company is an American multinational conglomerate founded in 1892.

Causes

- Source-Reference divergence
  - 62% of first sentences in the WIKIBIO dataset have additional information not present in the data - Dhingra et al. (2019)
- Training and modeling choices
  - Model learns the wrong correlations in the training sample.
  - Models prioritize parametric knowledge over input knowledge - Longpre et al. (2021)
- Decoding strategies
  - Decoding strategies that improve the diversity and fluency are correlated with increased hallucinations - (Dziri et al., 2021)

Neural Chart Summarization

Task where the goal is to explain a chart and summarize key takeaways from it in natural language.

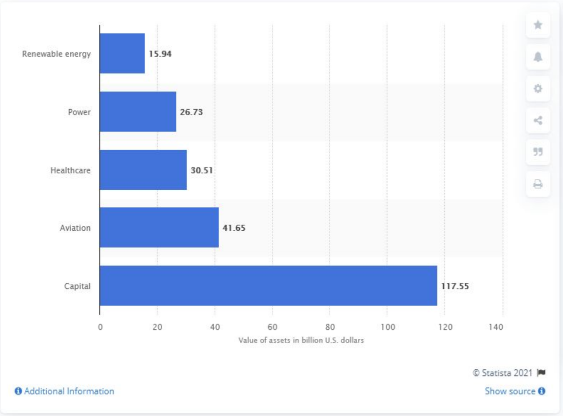


Segment	Value of assets in billion U.S. dollars
Renewable energy	15.94
Power	26.73
Healthcare	30.51
Aviation	41.65
Capital	117.55

Title: General Electric 's total assets in FY 2019 , by segment (in billion U.S. dollars)

This statistic represents General Electric 's total assets in the fiscal year of 2019 , with a breakdown by segment . In its healthcare segment , the company had assets to the value of around 30.5 billion U.S. dollars .

How to model this task?



Segment	Value of assets in billion U.S. dollars
Renewable energy	15.94
Power	26.73
Healthcare	30.51
Aviation	41.65
Capital	117.55

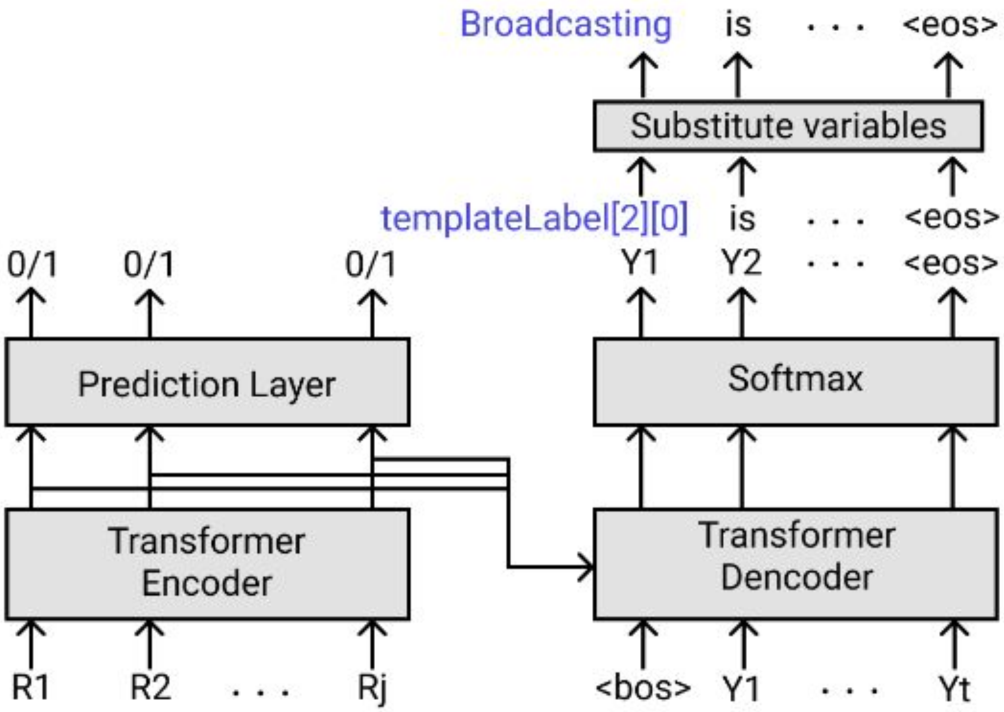
Title: General Electric 's total assets in FY 2019 , by segment (in billion U.S. dollars)

This statistic represents General Electric 's total assets in the fiscal year of 2019 , with a breakdown by segment . In its healthcare segment , the company had assets to the value of around 30.5 billion U.S. dollars .

Related Work

Dataset	Task modelled as	Model Architecture	Size
Chart2Text - Obeid and Hoque (2020)	data-to-text	Transformer + substitution module	8,147
Autochart - Zhu et al. (2021)	image-to-text	-	23,543
SciCap - Hsu et al. (2021)	Data-to-text, image-to-text	LSTM, CNN+LSTM	2 million
Chart-to-text - Kanthara et al. (2022)	Data-to-text, image-to-text	T5, BART, ResNet-LSTM	34,811
Barch - Skrjanec et al., (2022)	Data-to-text	LSTM, Transformer, KGPT	1,063

Chart2Text - Obeid and Hoque (2020)



Model Output:

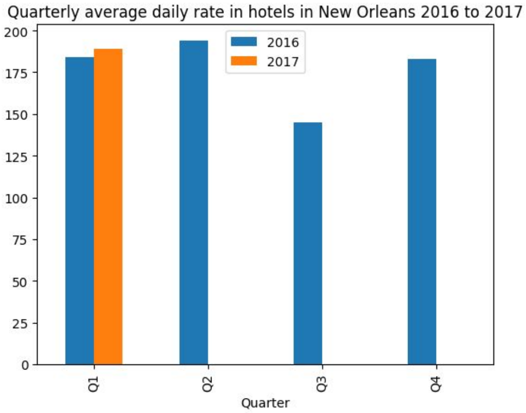
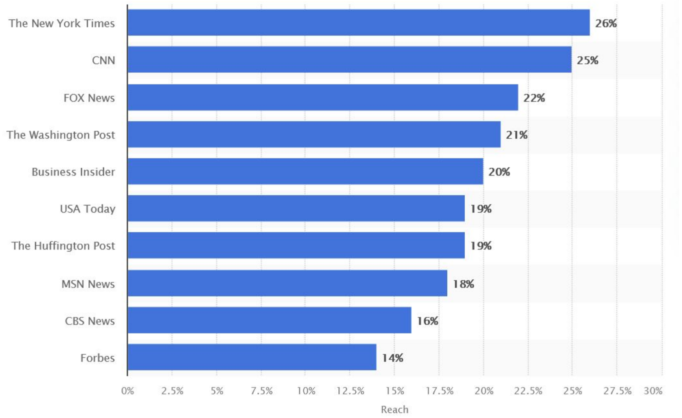
The templateLabel[2][0] templateTitle[2] templateTitle[4] is the largest source of templateTitle[2] for templateTitleSubject[0]. In 2018/2019, the football club earned approximately templateValue[2][0] templateScale euros from domestic and international competitions templateLabel[2][0], more than twice of what they earned in 2011/2012. The second biggest templateTitle[2] templateTitle[4] was templateLabel[3][0] – sponsorships and merchandising.

After Variable Substitution:

The Broadcasting revenue stream is the largest source of revenue for Liverpool FC . In 2018/2019 , the football club earned approximately 299.3 million euros from domestic and international competitions Broadcasting , more than twice of what they earned in 2011/2012 . The second biggest revenue stream was Commercial – sponsorships and merchandising .

Figure 3: Demonstration of data variable substitution.

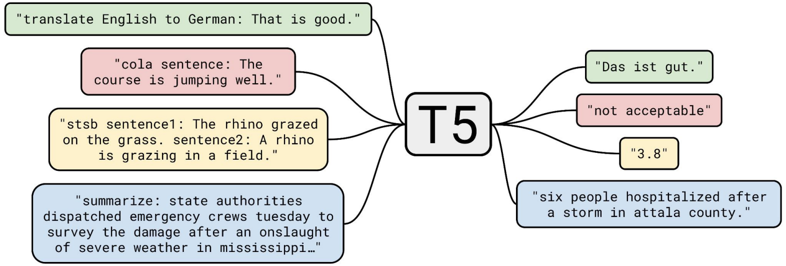
Source: Obeid, J., & Hoque, E. (2020). Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model.

<table>
<tr>
<td>Chart2Text - Obeid and Hoque (2020)</td>
<td>
</td>
<td>Quarter|Q1|0|bar_chart 2016|184|1|bar_chart 2017|189|2|bar_chart Quarter|Q2|0|bar_chart 2016|194|1|bar_chart 2017|0|2|bar_chart Quarter|Q3|0|bar_chart 2016|145|1|bar_chart 2017|0|2|bar_chart Quarter|Q4|0|bar_chart 2016|183|1|bar_chart 2017|0|2|bar_chart</td>
</tr>
<tr>
<td>Chart-to-text - Kanthara et al. (2022)</td>
<td>
</td>
<td>Most popular news brands in the United States as of June 2018 , by reach | The New York Times | CNN | FOX News | The Washington Post | Business Insider | USA Today| The Huffington Post| MSN News| CBS News| Forbes| 26% | 25% | 22% | 21% | 20% | 19% | 19% | 18% | 16% | 14% |</td>
</tr>
</table>

Text-To-Text Transfer Transformer

T5: Text-to-Text Transfer Transformer

- Single model for wide variety of tasks.
- Sequence-to-sequence architecture.
- Child of BERT and GPT-2
  - Encoder: Masked Language Modelling
  - Decoder: Autoregressive



Source: Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21.140 (2020): 1-67.

Automatic Evaluation Metrics

Evaluating NLG models

- Evaluating NLG is hard compared to classification
  - Several factors like quality, fluency, verbosity, consistency are considered.
- Evaluated across multiple metrics.
- Most common metrics are BLEU and ROUGE.
  - Problem: only measures surface overlap and not quality.

BLEURT

- Trained regression model based on BERT.
- Measures semantic similarity.
    - Input: reference-hypothesis pair
    - Output: Score between [-1,1] that indicates to what extent hypothesis convey the meaning of references.

NUBIA: NeUral Based Interchangeability Assessor for Text Generation

- Interpretable metric.
- Utilizes three models to produce a score.
- Score based on three factors:
  - Semantic Similarity (Roberta-STS)
    - Scores the difference between two sentences.
  - Logical Entailment (Roberta-NLI)
    - Entailment, Contradiction or Neutral (Undecided)
  - Grammaticality (GPT-2)
- Useful for evaluating faithfulness

Perplexity

- Exponentiation of negative-log likelihood.
- Informative for grammaticality or fluency when evaluated for large pre-trained transformers like GPT-2, T5, etc.

Identifying problems, experiment and results

Input format

- Linearized data table
- Input table lacks chart related information like title, *x-axis* and *y-axis* labels and legends

[Chart2Text](#)-small - Obeid and Hoque (2020)	Quarter|Q1|0|bar_chart 2016|184|1|bar_chart 2017|189|2|bar_chart Quarter|Q2|0|bar_chart 2016|194|1|bar_chart 2017|0|2|bar_chart Quarter|Q3|0|bar_chart 2016|145|1|bar_chart 2017|0|2|bar_chart Quarter|Q4|0|bar_chart 2016|183|1|bar_chart 2017|0|2|bar_chart
[Chart-to-text](#)- big Kanthara et al. (2022)	Most popular news brands in the United States as of June 2018 , by reach | The New York Times | CNN | FOX News | The Washington Post | Business Insider | USA Today| The Huffington Post| MSN News| CBS News| Forbes| 26% | 25% | 22% | 21% | 20% | 19% | 19% | 18% | 16% | 14% |

Data	Title	Reference	Hypothesis
Quarter|Q1|0|bar_chart 2016|184|1|bar_chart 2017|189|2|bar_chart Quarter|Q2|0|bar_chart 2016|194|1|bar_chart 2017|0|2|bar_chart Quarter|Q3|0|bar_chart 2016|145|1|bar_chart 2017|0|2|bar_chart Quarter|Q4|0|bar_chart 2016|183|1|bar_chart 2017|0|2|bar_chart	Quarterly average daily rate in hotels in New Orleans 2016 to 2017	This statistic shows the quarterly average daily rate in hotels in New Orleans in 2016 and 2017 . In the first quarter of 2017 , the average daily rate for hotels in New Orleans in the United States was 189 U.S. dollars .	This statistic shows the quarterly average daily rate of hotels in Seattle in 2016 and 2017. In the first quarter of 2017, the average daily rate of hotels in Seattle in the United States was 189 U.S. dollars.

Data	Reference	Hypothesis
Most popular news brands in the United States as of June 2018 , by reach | The New York Times | CNN | FOX News | The Washington Post | Business Insider | USA Today| The Huffington Post| MSN News| CBS News| Forbes| 26% | 25% | 22% | 21% | 20% | 19% | 19% | 18% | 16% | 14% |	This statistic gives information on the most popular news brands in the United States as of June 2018 , ranked by reach . According to the study , the brand with the highest reach was The New York Times , which reached 26 percent of consumers in June 2018 .	The New York Times CNN was the most popular news brand in the United States as of June 2018, with a reach of 66 percent among U.S. consumers. The New York Times was followed by Fox News and The Washington Post, which both had a 19 percent reach among U.S. consumers.

Hypothesis 1

- Formatting the input in a better and informative manner by adding more chart related information will result in higher n-gram scores thus indicating it reduces the hallucinations.

Proposed input format

title + x-y labels + x-y values	title + x - y labels/legends + x-y values
Most popular news brands in the United States as of June 2018 , by reach x-y labels news brand - Reach, x-y values The New York Times 26% , CNN 25% , FOX News 22% , The Washington Post 21% , Business Insider 20% , USA Today 19% , The Huffington Post 19% , MSN News 18% , CBS News 16% , Forbes 14%	Quarterly average daily rate in hotels in New Orleans 2016 to 2017 \n labels Quarter - 2016 - 2017 values Q1 184 189 , Q2 194 - , Q3 145 - , Q4 183 -

Experiment design

- Train three models using T5.
- On Chart2Text-small dataset (Obeid and Hoque):
  - *model_t5_original* trained on linearized data as formatted by the authors.
  - *model_t5_modified* trained on proposed linearized input.
- On chart-to-text big dataset (Kanthara et al.):
  - *model_t5_big* trained on proposed linearized input.
- Splits:
  - Train 0.75, Validation 0.15, Test 0.15.

Results

Model	BLEU ↑	ROUGE-2 ↑	PPL ↓	BLEURT ↑
model_t5_original	26.19	33.56	-	-
model_t5_modified	33.90	44.84	-	-
Obeid and Hoque (2020) transformer	18.54	-	-	-
model_t5_big	39.82	55.03	8.27	0.17
Kanthara et al. (2022) T5	37.01	-	10.0	0.15

Model_t5_big was trained for 170k iterations (14 epochs). Kanthara el al. was trained for 500k iterations (82 epochs) with the same hyperparameters.

NUBIA

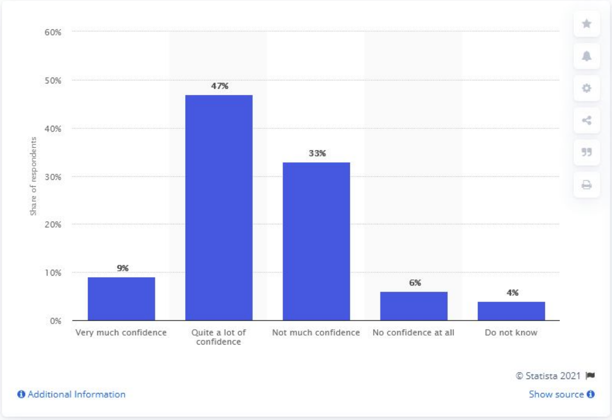
Model	Logical Agreement ↑	Contradiction ↓	Neutrality	Semantic Similarity ↑	NUBIA ↑
model_t5_original	5.56	67.84	26.5	3.04/5	35.4
model_t5_modified	33.27	22.31	44.41	3.5/5	46.9
model_t5_big	39.34	21.36	39.38	3.68/5	55.67

Data	Reference	Hypothesis
Quarterly average daily rate in hotels in New Orleans 2016 to 2017 \n labels Quarter - 2016 - 2017 values Q1 184 189 , Q2 194 - , Q3 145 - , Q4 183 -	This statistic shows the quarterly average daily rate in hotels in New Orleans in 2016 and 2017 . In the first quarter of 2017 , the average daily rate for hotels in New Orleans in the United States was 189 U.S. dollars .	This statistic shows the quarterly average daily rate in hotels in New Orleans in 2016 and 2017. In the first quarter of 2017, the average daily rate for hotels in New Orleans in the United States was 189 U.S. dollars.

Data	Reference	Hypothesis
Most popular news brands in the United States as of June 2018 , by reach x-y labels news brand - Reach, x-y values The New York Times 26% , CNN 25% , FOX News 22% , The Washington Post 21% , Business Insider 20% , USA Today 19% , The Huffington Post 19% , MSN News 18% , CBS News 16% , Forbes 14%	This statistic gives information on the most popular news brands in the United States as of June 2018 , ranked by reach . According to the study , the brand with the highest reach was The New York Times , which reached 26 percent of consumers in June 2018 .	The statistic presents the most popular news brands in the United States as of June 2018, ranked by reach. According to the findings, The New York Times was the most popular news brand with 26 percent reach among U.S. adults.

Additional information in the training summaries

This statistic depicts the operating profit of the H & M Group worldwide from 2009 to 2019 . In 2019 , the global operating profit of the H & M Group was about 1.8 billion U.S. dollars.H & M is a leading global fashion company with strong values and a clear business concept . H & M constantly strives to have the best customer offering in each individual market – which includes giving customers the best price .



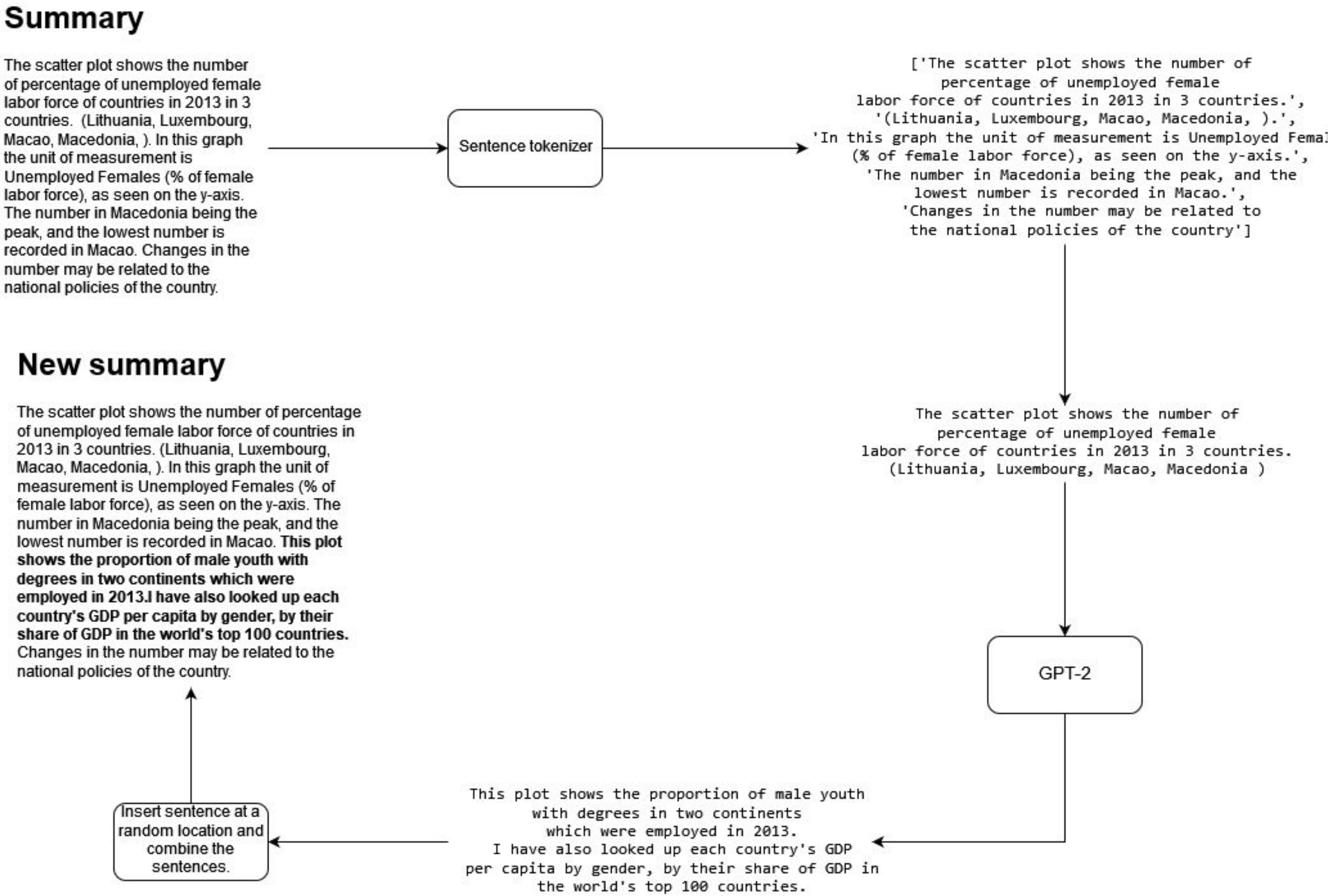
As of March 2020 , the biggest share of the Norwegian respondents (47 percent) had quite a lot of confidence that the national health care system can handle the coronavirus (COVID-19) outbreak in a good way . By comparison , only four percent had no trust in the health care system at all . The first case of COVID-19 in Norway was confirmed on February 26 , 2020 . For further information about the coronavirus (COVID-19) pandemic , please visit our dedicated Facts and Figures page .

Hypothesis 2

- The model learns additional information from the training summaries and generates additional information during inference.

Experiment design

- Train two models using T5 on Autochart dataset (Zhu et al).
- *t5_autochart_original*:
    - trained on linearized data as represented in slide 28.
- *t5_autochart_modified*:
    - trained on autochart summaries that are augmented with some noise generated from GPT- 2.
- Train: 8000, Validation: 1296, Test: 1297


Summary
The scatter plot shows the number of percentage of unemployed female labor force of countries in 2013 in 3 countries. (Lithuania, Luxembourg, Macao, Macedonia,). In this graph the unit of measurement is Unemployed Females (% of female labor force), as seen on the y-axis. The number in Macedonia being the peak, and the lowest number is recorded in Macao. Changes in the number may be related to the national policies of the country.
Sentence tokenizer
['The scatter plot shows the number of percentage of unemployed female labor force of countries in 2013 in 3 countries.', '(Lithuania, Luxembourg, Macao, Macedonia,).', 'In this graph the unit of measurement is Unemployed Femal
(% of female labor force), as seen on the y-axis.', 'The number in Macedonia being the peak, and the lowest number is recorded in Macao.', 'Changes in the number may be related to the national policies of the country']
The scatter plot shows the number of percentage of unemployed female labor force of countries in 2013 in 3 countries. (Lithuania, Luxembourg, Macao, Macedonia)
GPT-2
This plot shows the proportion of male youth with degrees in two continents which were employed in 2013. I have also looked up each country's GDP per capita by gender, by their share of GDP in the world's top 100 countries.
Insert sentence at a random location and combine the sentences.
New summary
The scatter plot shows the number of percentage of unemployed female labor force of countries in 2013 in 3 countries. (Lithuania, Luxembourg, Macao, Macedonia,). In this graph the unit of measurement is Unemployed Females (% of female labor force), as seen on the y-axis. The number in Macedonia being the peak, and the lowest number is recorded in Macao. **This plot shows the proportion of male youth with degrees in two continents which were employed in 2013.I have also looked up each country's GDP per capita by gender, by their share of GDP in the world's top 100 countries.** Changes in the number may be related to the national policies of the country.

Results

Model	BLEU ↑	ROUGE-2 ↑	BLEURT ↑
t5_autochart_original	58.88	62.03	0.185
t5_autochart_modified	46.09	48.92	-0.109

Model	Logical Agreement ↑	Contradiction ↓	Neutrality	Semantic Similarity ↑	NUBIA ↑
t5_autochart_original	33.07	31.04	35.51	3.23/5	79.89
t5_autochart_modified	24.03	24.45	51.50	3.33/5	88.49

Conclusion of Experiments

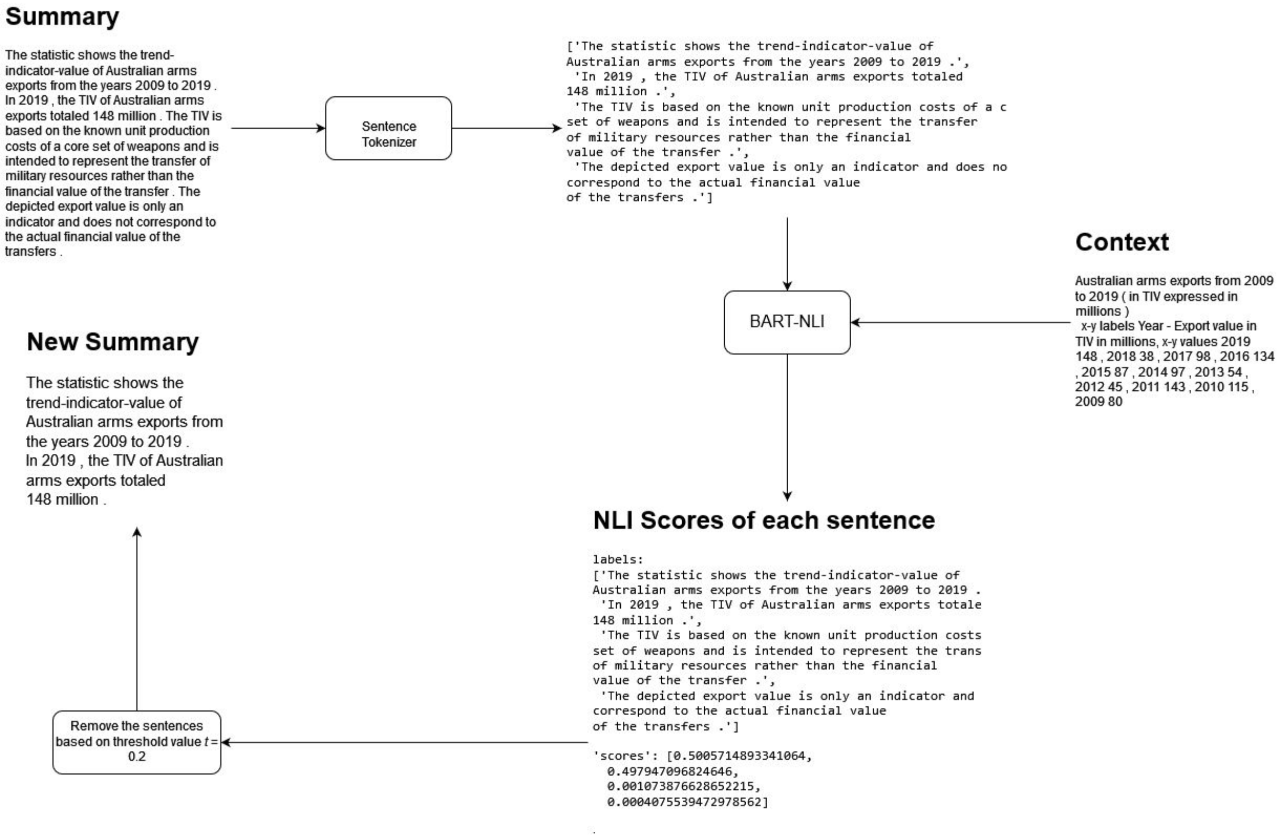
- Adding more chart related information results in better scores and fewer training epochs (80% less).
- For our second experiment, BLEU, ROUGE, BLEURT agree but NUBIA does not.
  - Possible argument: Overall NUBIA score is due to high neutrality. We are only interested in agreement and not irrelevancy.
  - Further work needs to be carried out to understand how NUBIA computes the scores.

Further Work

Utilizing NLI

- NLI: Task of determining whether the given hypothesis logically follows from the premise.
  - What NUBIA does when checking logical agreement.
- In our case:
  - Hypothesis: Segmented Summary
  - Premise: Data table
- Can be used at three levels
  - Pre-processing
  - Post-processing
  - Training and model

Preprocessing and postprocessing

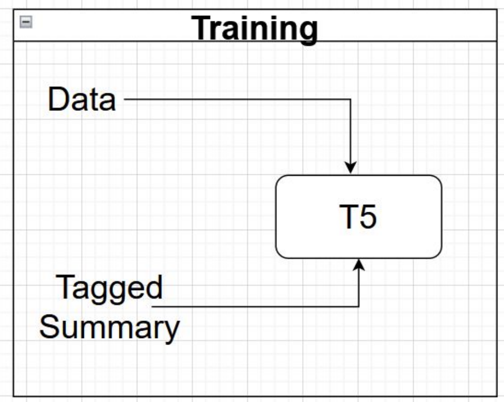


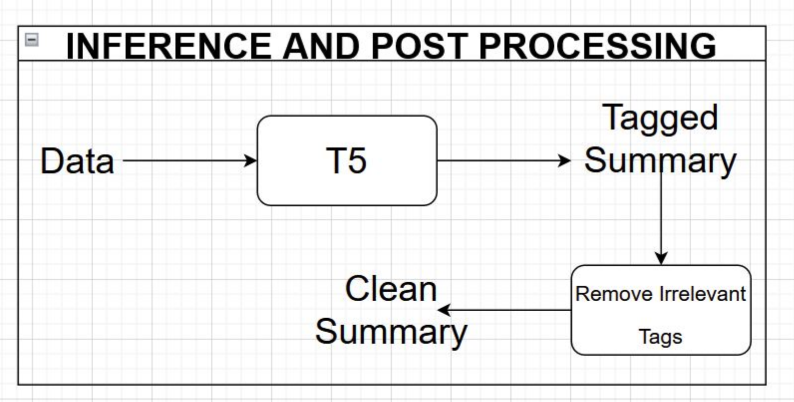
Any problems with this approach?

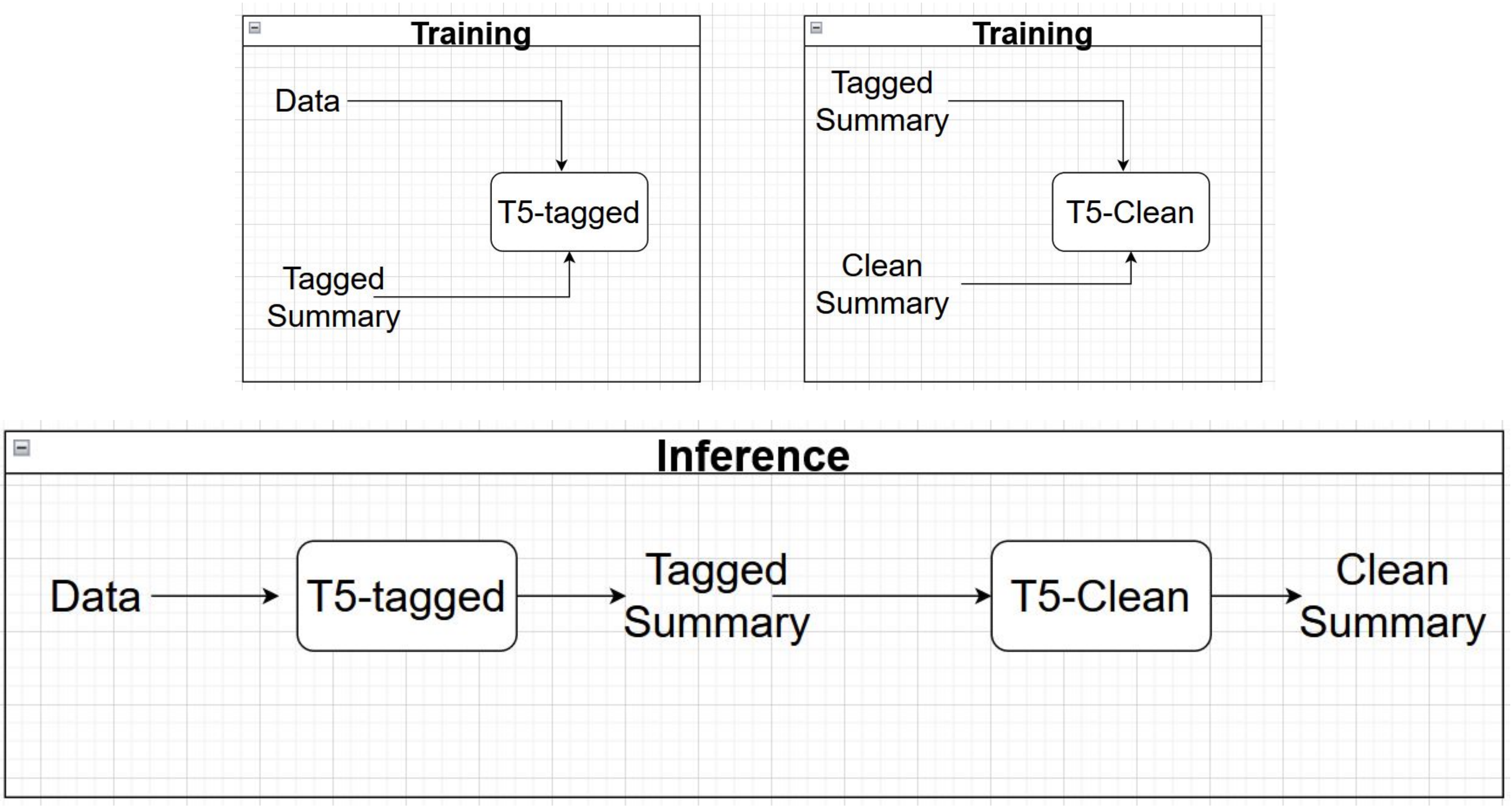
Model and Training

- Use NLI to create a new parallel corpus:
  - Tagged summary - (Clean)Summary pairs
  - Tags: <relevant> and <irrelevant>

Data	Summary	Tagged Summary	Clean Summary
Forecast of the national debt of selected euro countries until 2020 labels Country - 2018 - 2019 - 2020 values Greece 181.1% 174.9% 168.9% , Italy 132.2% 133.7% 135.2% , Portugal 121.5% 119.5% 116.6% , France 98.4% 99% 98.9% , Spain 97.1% 96.3% 95.7% , Cyprus 102.5% 96.4% 89.9% , Ireland 64.8% 61.3% 55.9% , Germany 60.9% 58.4% 55.6%	This statistic shows a forecast of the national debt of selected euro countries from 2018 to 2020 in relation to the gross domestic product (GDP) . The national debt figures include the debt of the central state , the states , the communities and the parishes , as well as social security . In Greece , the national debt is estimated to amount 168.9 percent of the GDP in 2020 .	<relevant>This statistic shows a forecast of the national debt of selected euro countries from 2018 to 2020 in relation to the gross domestic product (GDP) .</relevant> <irrelevant>The national debt figures include the debt of the central state , the states , the communities and the parishes , as well as social security .</irrelevant> <relevant>In Greece , the national debt is estimated to amount 168.9 percent of the GDP in 2020 .</relevant>	This statistic shows a forecast of the national debt of selected euro countries from 2018 to 2020 in relation to the gross domestic product (GDP). In Greece , the national debt is estimated to amount 168.9 percent of the GDP in 2020 .

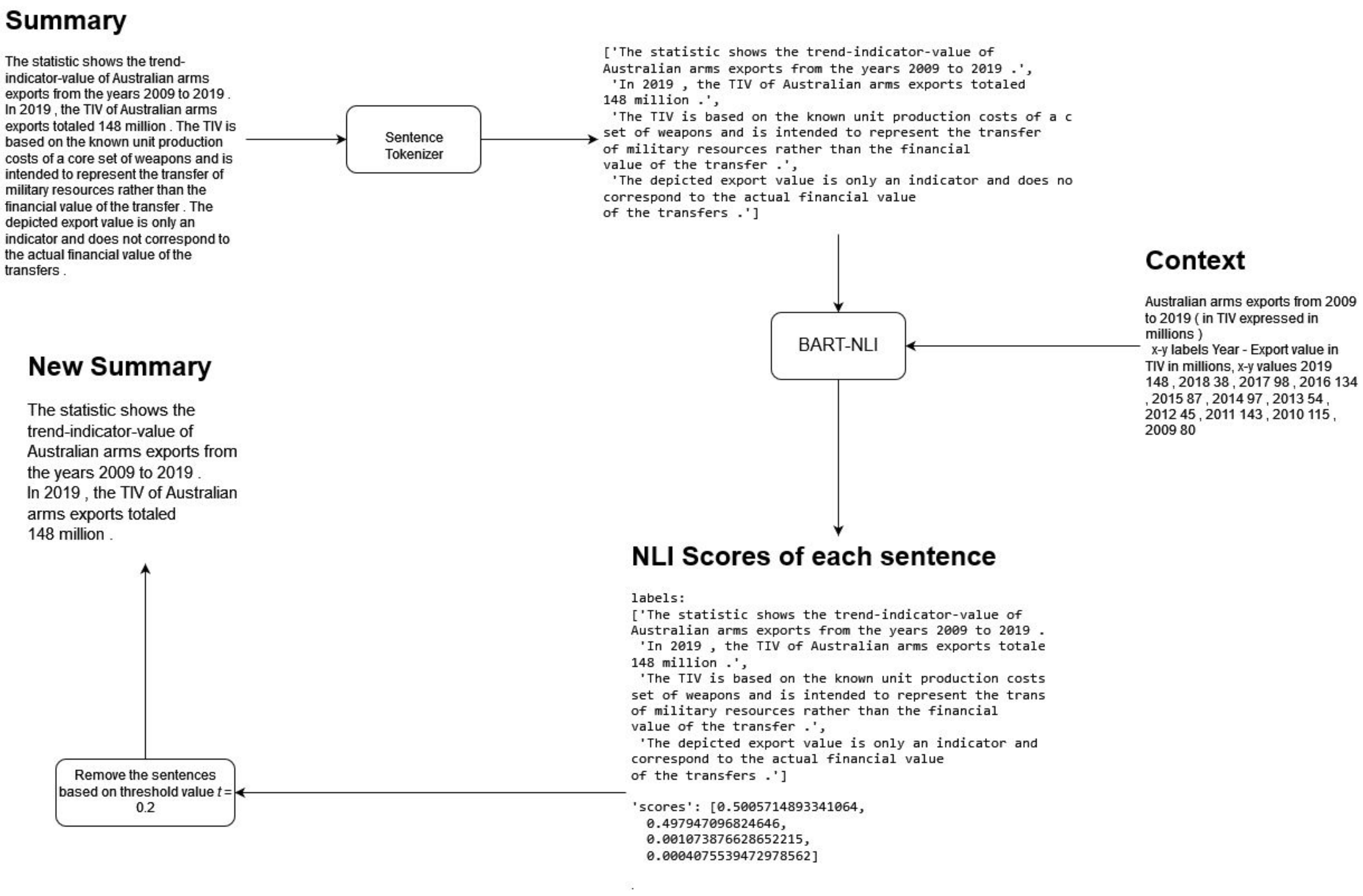





Training
Data
T5-tagged
Tagged Summary
Training
Tagged Summary
T5-Clean
Clean Summary
Inference
Data
T5-tagged
Tagged Summary
T5-Clean
Clean Summary

Too many expectations

- Expecting the pretrained NLI model to identify relevant and irrelevant sentences correctly.
- Expecting the T5 model to generate a summary as well as tag the sentences correctly.


Summary
The statistic shows the trend-indicator-value of Australian arms exports from the years 2009 to 2019 . In 2019 , the TIV of Australian arms exports totaled 148 million . The TIV is based on the known unit production costs of a core set of weapons and is intended to represent the transfer of military resources rather than the financial value of the transfer . The depicted export value is only an indicator and does not correspond to the actual financial value of the transfers .
Sentence Tokenizer
['The statistic shows the trend-indicator-value of Australian arms exports from the years 2009 to 2019 .', 'In 2019 , the TIV of Australian arms exports totaled 148 million .', 'The TIV is based on the known unit production costs of a c set of weapons and is intended to represent the transfer of military resources rather than the financial value of the transfer .', 'The depicted export value is only an indicator and does no correspond to the actual financial value of the transfers .']
Context
Australian arms exports from 2009 to 2019 (in TIV expressed in millions) x-y labels Year - Export value in TIV in millions, x-y values 2019 148 , 2018 38 , 2017 98 , 2016 134 , 2015 87 , 2014 97 , 2013 54 , 2012 45 , 2011 143 , 2010 115 , 2009 80
BART-NLI
New Summary
The statistic shows the trend-indicator-value of Australian arms exports from the years 2009 to 2019 . In 2019 , the TIV of Australian arms exports totaled 148 million .
NLI Scores of each sentence
labels: ['The statistic shows the trend-indicator-value of Australian arms exports from the years 2009 to 2019 . 'In 2019 , the TIV of Australian arms exports totale 148 million .', 'The TIV is based on the known unit production costs set of weapons and is intended to represent the trans of military resources rather than the financial value of the transfer .', 'The depicted export value is only an indicator and correspond to the actual financial value of the transfers .']
'scores': [0.5005714893341064, 0.497947096824646, 0.001073876628652215, 0.0004075539472978562]
Remove the sentences based on threshold value t = 0.2

Model: BART-NLI on huggingface
Setting: Multiple classes
Threshold: 0.3

Testing BART-NLI

- Out of the 50 examples tested:
    - 31 were 100% correctly tagged.
    - 19 were incorrectly tagged
- What kind of sentences were incorrectly tagged?

Relevant	Irrelevant
Additional Information	Trends
	Approximation
	When something sounds contradictory but is not.
	Difference in scale

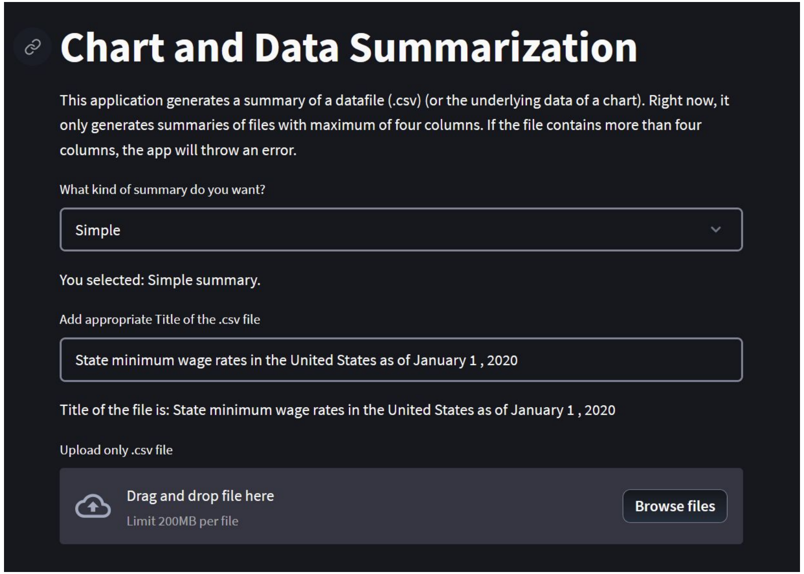
Official international reserves by country 2019 x-y labels Country - Value in billion U.S. dollars, x-y values Argentina 54.1 , Australia * 45.02 , Austria 26.46 , Brazil 386.48 , Canada** 86.3 , Chile 65.57 , China: Hong Kong* 471.34 , China: Mainland* 3383.04 , Colombia 54.5 , Germany 226.59 , Greece 8.65 , India * 430.78 , Ireland 5.29 , Japan 1379.46 , Republic of korea 401.48 , Mexico* 187.99 , Peru 67.66 , Singapore* 272.67 , Spain 78.06 , Switzerland* 829.42 , Thailand* 218.41 , United Kingdom 198.06 , United States 128.94

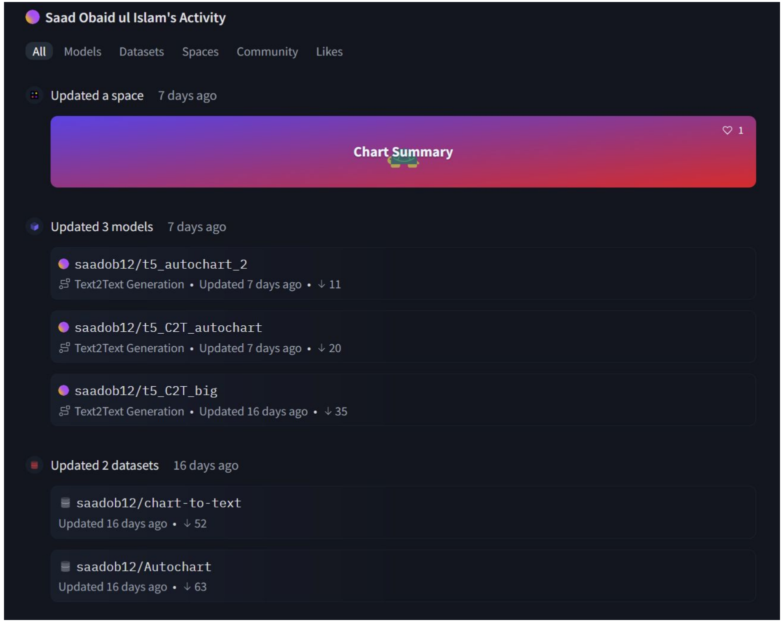
<irrelevant: 0.11783356219530106>Why do countries hold reserves ? </irrelevant> <irrelevant: 0.0819632038474083>Of all the countries in the world , China had , by far , the largest international reserves in August 2019 , with **3.38 trillion** U.S. dollars in reserves and foreign currency liquidity .</irrelevant> <irrelevant: 0.02616974525153637>A simple explanation for China 's accumulation of foreign currency could be its consistently positive and substantial trade balance .</irrelevant> <irrelevant: 0.014350765384733677>Japan was the **only other country with over** a trillion U.S. dollars in reserves , with a total of **1.38 trillion** U.S. dollars .</irrelevant>

Questions?

Try out the models and look at the datasets

https://huggingface.co/saadob12





The end

References

- Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. Natural Language Engineering, 3(1), 57-87
- Wen, T. H., Gasic, M., Mrksic, N., Su, P. H., Vandyke, D., & Young, S. (2015). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. arXiv preprint arXiv:1508.01745.
- Sripada, S., Reiter, E., & Davy, I. (2003). SumTime-Mousam: Configurable marine weather forecast generator. Expert Update, 6(3), 4-10.
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. arXiv preprint arXiv:2005.00661.
- Liu, T., Wang, K., Sha, L., Chang, B., & Sui, Z. (2018, April). Table-to-text generation by structure-aware seq2seq learning. In Thirty-Second AAAI Conference on Artificial Intelligence.
- Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-based knowledge conflicts in question answering. arXiv preprint arXiv:2109.05052.
- Dziri, N., Madotto, A., Zaiane, O., & Bose, A. J. (2021). Neural path hunter: Reducing hallucination in dialogue systems via path grounding. arXiv preprint arXiv:2104.08455.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140), 1-67.
- Obeid, J., & Hoque, E. (2020). Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. arXiv preprint arXiv:2010.09142.
- Zhu, J., Ran, J., Lee, R. K. W., Choo, K., & Li, Z. (2021). AutoChart: A Dataset for Chart-to-Text Generation Task. arXiv preprint arXiv:2108.06897.

- Hsu, T. Y., Giles, C. L., & Huang, T. H. K. (2021). Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624.
- Kanthara, S., Leong, R. T. K., Lin, X., Masry, A., Thakkar, M., Hoque, E., & Joty, S. (2022). Chart-to-Text: A Large-Scale Benchmark for Chart Summarization. arXiv preprint arXiv:2203.06486.
- Iza ˇSkrjanec, Muhammad Salman Edhi, and Vera Demberg. Barch: an english dataset of bar chart summaries. In Proceedings of the Language Resources and Evaluation Conference, pages 3552–3560
- Post, M. (2018). A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771.
- Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
- Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696.
- Kane, H., Kocyigit, M. Y., Abdalla, A., Ajanoh, P., & Coulibali, M. (2020). NUBIA: NeUral based interchangeability assessor for text generation. arXiv preprint arXiv:2004.14667.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.
- Parikh, A. P., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo: A controlled table-to-text generation dataset. arXiv preprint arXiv:2004.14373.