Just Pick One

A Comparative Analysis of BART and PEGASUS for Abstractive News Summarization

CS6120 - Summer 2025

Jack Einbinder

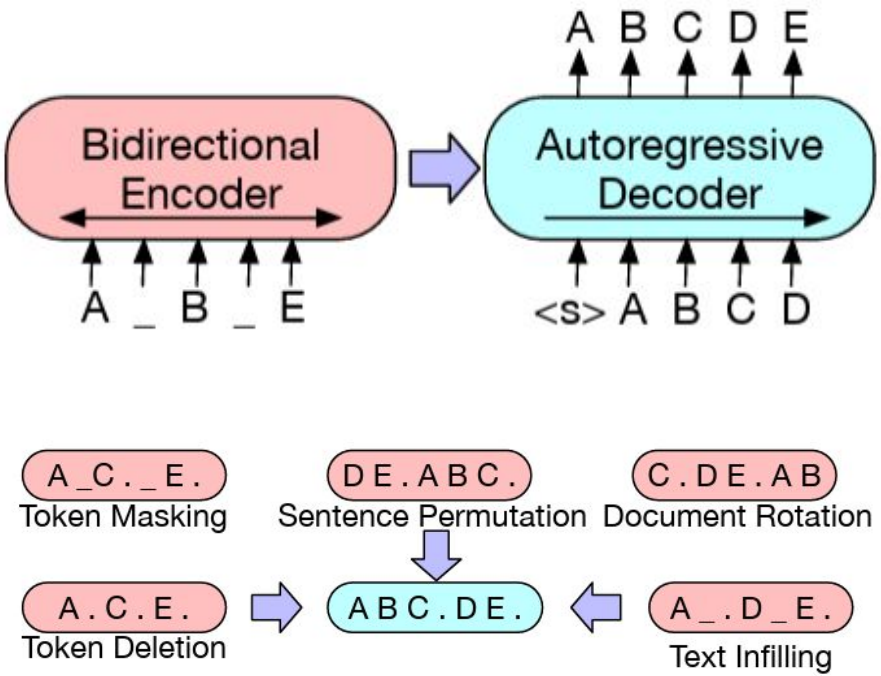
Background

- Summarization is the process of condensing a source document into a concise sequence that captures all of its essential content.
- Early summarization techniques were extractive
- Modern summarization focuses on abstractive summarization, enabled by Transformer-based models
- BART and PEGASUS are the current SOTA in abstractive summarization

CNN/Daily Mail Dataset

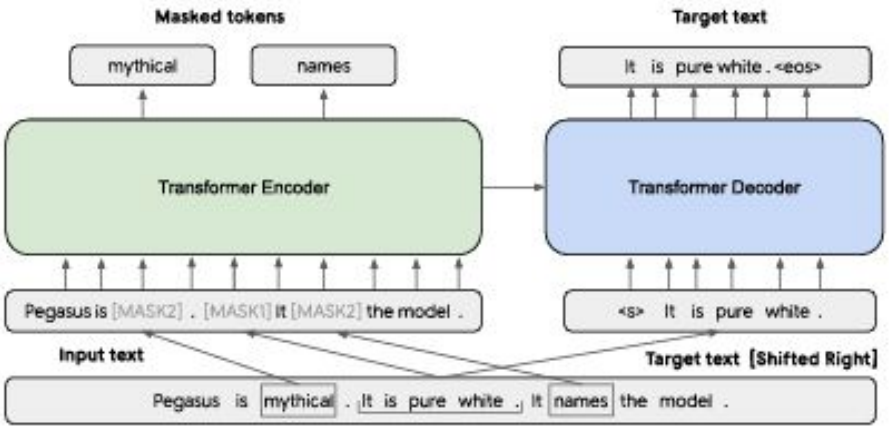
- Compiled by Hermann et al in 2015 for abstractive summarization
- Contains over 300K English-language news articles written with corresponding summaries
- Used as a labeled dataset for the supervised training of abstractive summarization models
- Foundational in training models for abstractive summarization, enabling both BART and PEGASUS

BART - (Bidirectional and Auto-Regressive Transformer)



- Combines two key technologies
  - Bidirectional encoder (BERT)
  - Autoregressive Decoder (GPT)
- Corrupts input text using various strategies
  - Token Masking
  - Sentence Permutation
  - Document Rotation
  - Token Deletion
  - Text Infilling
- Decoder seeks to reconstruct the original text

PEGASUS - (Pre-training with Extracted Gap-sentences for Abstractive Summarization)



MLM (also depicted here) is not used in the final model

- Same architecture as BART
  - Encoder/decoder
- Introduces a novel pre-training objective
  - Gap Sentences Generation (GSG)
- Current SOTA in abstractive summarization
- Slightly outperforms* BART on the CNN/Daily Mail dataset

ROUGE - (Recall-Oriented Understudy for Gisting Evaluation)

- Developed by Lin at University of Southern California in 2004
- Addresses the problem of automating the evaluation of summarization quality
- Introduces four summarization metrics:
  - ROUGE-N
  - ROUGE-L
  - ROUGE-W
  - ROUGE-S
- ROUGE-1, ROUGE-2 and ROUGE-L are standard in the literature

Objective

- Evaluate and compare the quality of news summaries generated by PEGASUS and BART models pre-trained on the CNN/Daily Mail dataset
- Determine the higher quality model using intrinsic and extrinsic methods
  - Intrinsic Evaluation Metrics
    - ROUGE-1
    - ROUGE-2
    - ROUGE-L
  - Extrinsic Evaluation Metrics
    - Manual Review (1-5)
- Conduct an ablation study testing hyperparameter strategies on the higher performer to determine an optimal configuration

Methodology

BART Default vs PEGASUS Default

- Compare performance on 100 summaries using ROUGE metrics
- Compare 10 summaries each manually
- Explore hyperparameter configurations of winning model in ablation study

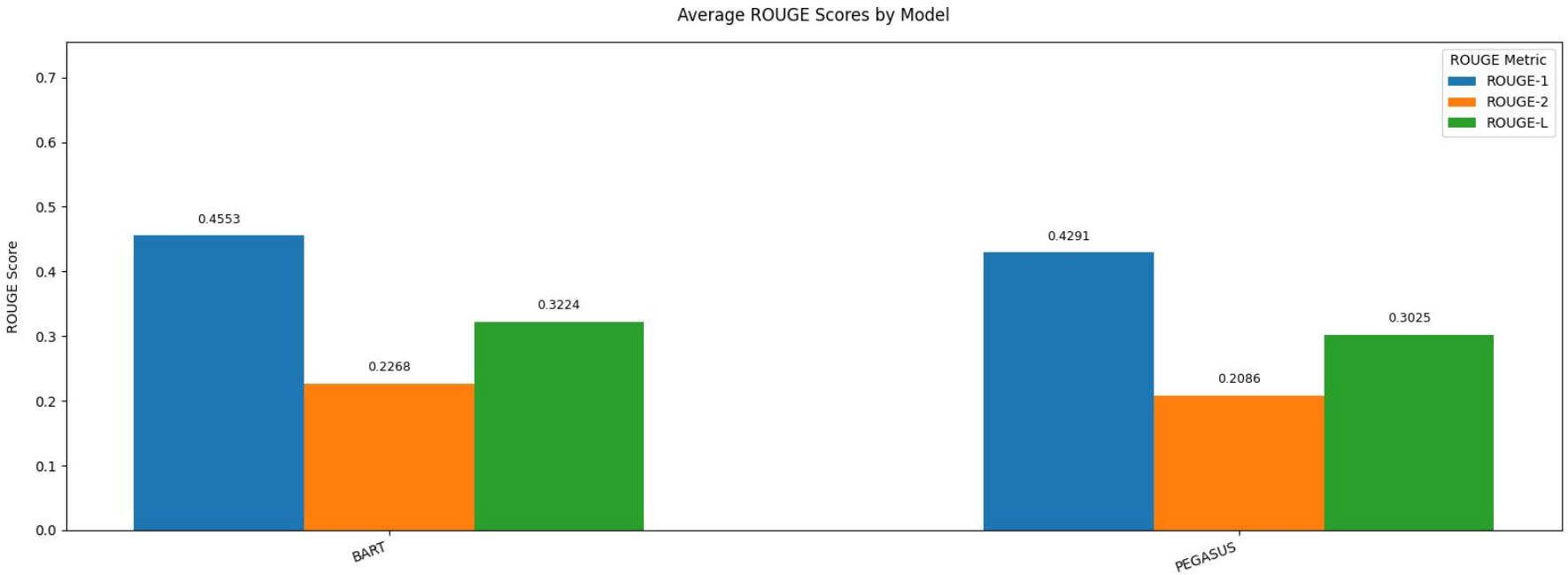
Ablation Study (10 model configurations)

- Compare performance on 100 summaries using ROUGE metrics
- Compare 3 summaries each manually

Results

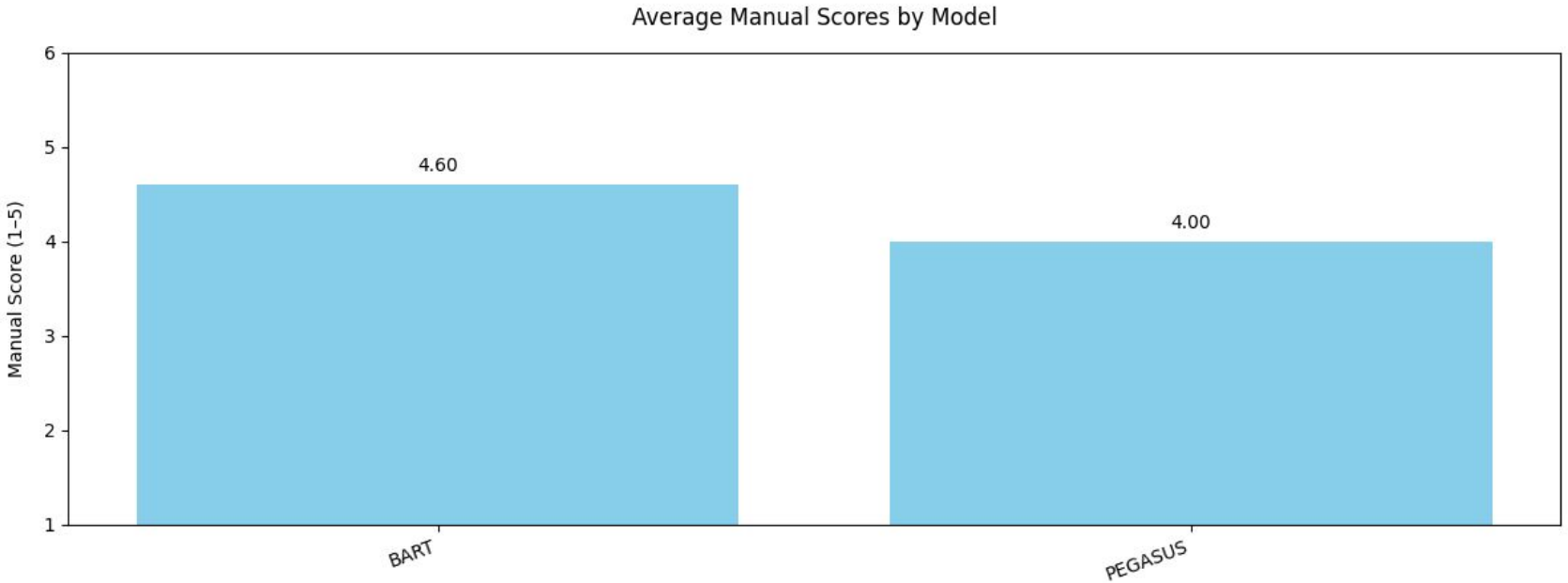
All evaluation data, including model summaries, and code is available on Github

BART vs PEGASUS - ROUGE



BART Default outperformed PEGASUS Default slightly across all ROUGE metrics.

BART vs PEGASUS - Manual Evaluation



BART summaries were slightly more fluent, contained more detail, and avoided repeated phrases.

BART vs PEGASUS - Example Summary

Ground-truth	BART	PEGASUS
Didier Drogba given the Barclays Spirit of the Game award . The 37-year-old's foundation has done impressive work in Africa . Some of Chelsea's stars attended a charity ball which raised £400,000 . CLICK HERE for all the latest Chelsea news .	*Didier Drogba has been awarded with a Barclays Spirit of the Game trophy. The Chelsea forward set up the 'DidierDrogba Foundation in Africa' He hopes to inspire the next generation of footballers in Africa to fall in love with the game. The 37-year-old scored against Leicester on Wednesday.*	*Didier Drogba has set up the 'Didier Drogba Foundation in Africa'<n>**The 'Didier Drogba Foundation,' contribute financial and material support in education and health including school bags for the school children .<n>The 'Didier Drogba Foundation,' contribute financial and material support in education and health including school bags for the school children .**<n>Chelsea's stars such as Eden Hazard, Petr Cech and Branislav Ivanovic were out in force earlier this month as they raises £400,000 for the foundation at a charity ball .*

BART produced a higher quality, more succinct summary with more fluent language. PEGASUS captured essential details but generated a clunkier summary with a repeated sentence (red).

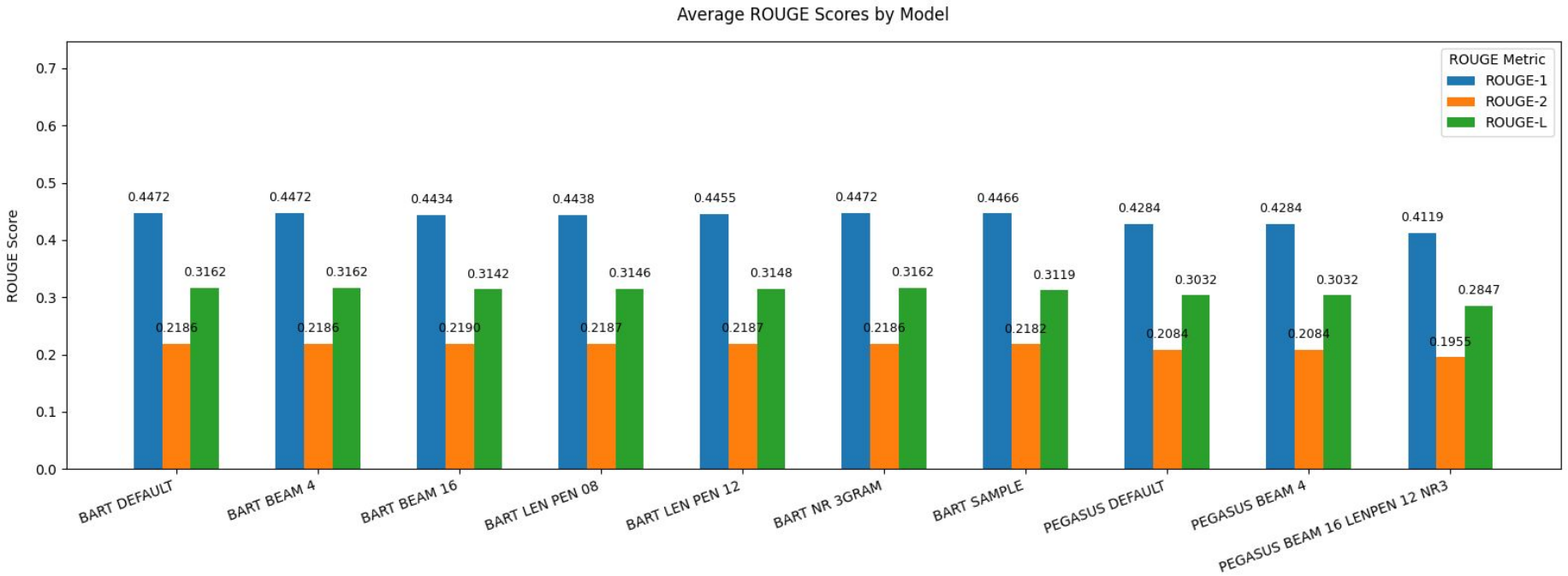
Ablation Study Model Configurations - BART

Model Name	Details
BART DEFAULT	Default BART model from HuggingFace.
BART BEAM 4	Beam search enabled with num_beams = 4. The model explores the 4 highest-probability sequences at each step.
BART BEAM 16	Beam search enabled with num_beams = 16. The model explores the 16 highest-probability sequences at each step. A length penalty of 1.2 is applied to favor slightly longer outputs. no_repeat_ngram_size = 3 is used to reduce repetition by restricting repeated tri-grams.
BART LENGTH PENALTY 0.8	Beam search enabled with num_beams = 4. A length penalty of 0.8 is applied to favor shorter outputs.
BART LENGTH PENALTY 1.2	Beam search enabled with num_beams = 4. A length penalty of 1.2 is applied to favor longer outputs.
BART NO-REP 3-GRAM	Beam search enabled with num_beams = 4. no_repeat_ngram_size = 3 is used to reduce repetition by restricting repeated tri-grams.
BART SAMPLE	Sampling enabled with do_sample = True. At each step, the model samples from the top 90% of tokens (top_p = 0.9) with a temperature of 1.2 to flatten the probability distribution, making lower-probability tokens more likely.

Ablation Study Model Configurations - PEGASUS

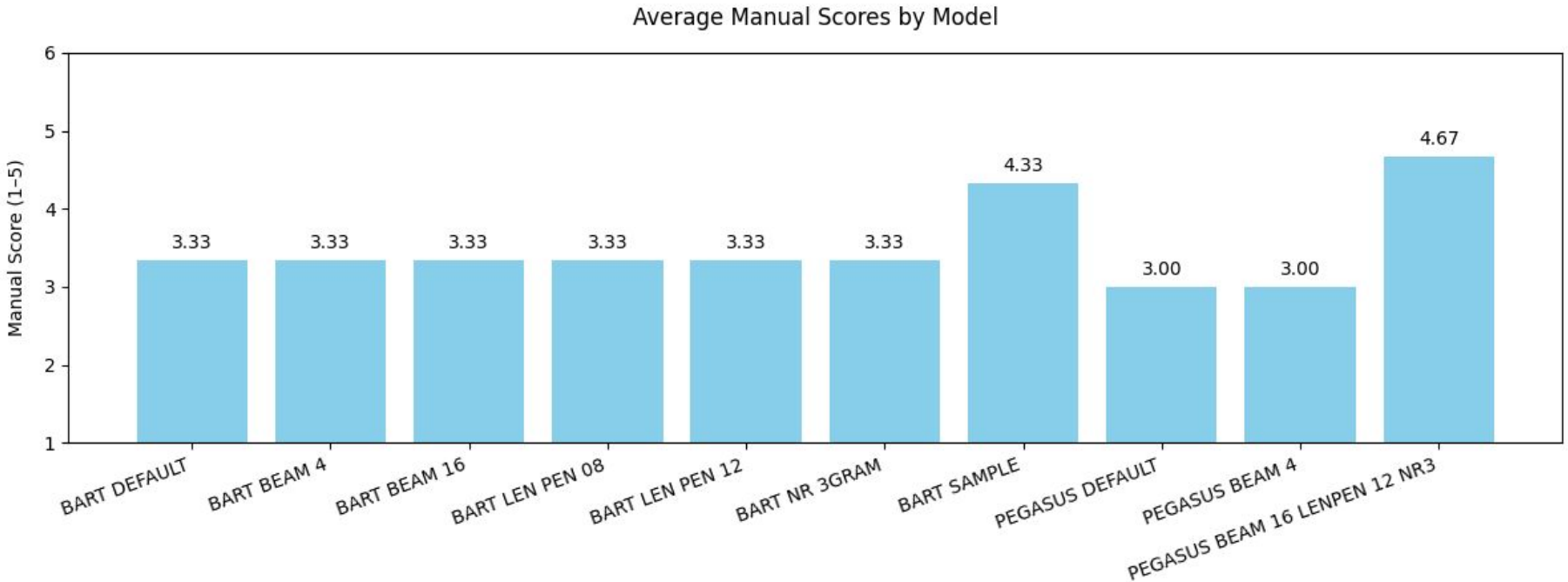
Model Name	**Details**
PEGASUS DEFAULT	Default PEGASUS model from HuggingFace.
PEGASUS BEAM 4	Beam search enabled with num_beams = 4. The model explores the 4 highest-probability sequences at each step.
PEGASUS BEAM 16	Beam search enabled with num_beams = 16. A length penalty of 1.2 is applied to favor longer outputs. no_repeat_ngram_size = 3 is used to reduce repetition by restricting repeated tri-grams.

Ablation Study - ROUGE



ROUGE scores remained remarkably stable across all configurations, with no significant improvements observed from varying hyperparameter strategies.

Ablation Study - Manual Evaluation



Manual review indicates a slight advantage for the BART sampling approach (top_p = 0.9) and overall superior performance from PEGASUS using beam search (num_beams = 16).

Ablation Study - Example Summary

Ground-truth	BART Default	BART Sample	PEGASUS Beam 16
Image went viral after it was posted by user Bobitis on Reddit . *The unnamed youth, from New York, knelt on a chair to reach whiteboard .* *The boy is the son of New York professor and drew map from memory .*	*The 11-year-old boy is seen drawing the intricate map of the world* ***from scratch****. He is the son of a professor in New York who invited him to her class. In the images, the youth is seen kneeling on a chair in order to reach the whiteboard.*	*The 11-year-old is the son of a professor in New York. He is seen kneeling on a chair and drawing the world map* ***from memory****.* ***The map is highly accurate, with some major countries labelled. It includes political borders and even tiny coastal islands.***	*The 11-year-old boy* ***with autism*** *is seen drawing the intricate map of the world from scratch .<n>In the images, the youth is seen kneeling on a chair in order to reach the whiteboard, with* ***his pen held to the extremely detailed map****.<n>****Not only is the map accurate, with some major countries labelled, but the high level of detail extends to the inclusion of political borders and even tiny coastal islands .***

The BART Default model captures almost all of the necessary details, but fails to mention that the 11-year-old subject of the article has autism, and that his map is exceptionally detailed. The BART Sample model contains more relevant information (yellow) and better phrasing (green). The PEGASUS Beam 16 mode produced the best summary with all necessary details and better, more fluent phrasing (green).

Conclusion

- BART and PEGASUS represent the SOTA in abstractive summarization
- PEGASUS boasts superior performance, the reality is more complex
- The default BART configuration outperforms the default PEGASUS configuration (slightly)
- Varying hyperparameter configurations have minimal impact on summary quality
    - Sampling provides some advantage
    - Aggressive beam search provides some advantage with PEGASUS
- The choice between BART and PEGASUS should be guided primarily by personal preference and convenience