Language Models for Dialogue Summarization: Fine-Tuning the Pegasus Model for Specific Summarization Tasks

Laure Hajislam, Clàudia Domènech Farré, Ovindu Chakrawarthige

June 2, 2024

1 Division of tasks

All authors contributed equally to this work. Each of us developed a different script for the training of the model on each dataset. Then, we did the following tasks separately: Ovindu was in charge of putting the code all together and modularizing it. Clàudia was in charge of putting all the information and insights together by writing this report. Laure was in charge of conducting the statistical tests for the results, and doing the results analysis.

Abstract

This study addresses the research question: How can fine-tuning and knowledge transfer fine-tuning techniques enhance the Pegasus model's ability to summarize dialogues from diverse datasets? We fine-tuned Pegasus, developed by Google for abstractive text summarization, using various dialogue based datasets, leveraging the HuggingFace Transformers library. Our methodology involved two stages: initially fine-tuning Pegasus independently on each dataset, followed by sequential fine-tuning to explore knowledge transfer effects. Evaluation using ROUGE metrics revealed that while fine-tuning generally improves performance, the extent of improvement varies by dataset. Results indicate that dataset-specific tuning significantly enhances summarization capabilities, with knowledge transfer fine-tuning showing promise, particularly for structurally similar datasets. This research provides valuable insights into optimizing model performance for real-world applications in fields such as customer service, healthcare, and project management, offering a cost-effective and scalable approach to deploying high-performance summarization models.

2 Introduction and Background

In the realm of natural language processing (NLP), summarizing dialogues presents unique challenges due to the informal, colloquial, and often non-linear nature of conversations. Effective summarization of these dialogues is crucial for various applications in fields like customer service, healthcare, and project management, where extracting essential information quickly can greatly enhance efficiency and decision-making processes.

This project explores the efficacy of fine-tuning and knowledge transfer fine-tuning techniques on the Pegasus model, developed by Google for abstractive text summarization. Our primary research question investigates how these techniques can enhance the Pegasus model's ability to summarize dialogues from diverse datasets, each with distinct characteristics. Fine-tuning involves adapting a pre-trained model on a specific dataset, while knowledge transfer fine-tuning refers to sequentially fine-tuning the model on multiple datasets to leverage learned knowledge from one domain to another.

A wide variety of modern technologies can be accessed for summarization tasks, however, most well performing solutions like GPT-4 are very costly for sustained usage on large datasets. Despite signifi-

cant progress in dialogue summarization (3; 6), there remains a lack of focused research on effectively summarizing informal and non-linear dialogues such as with our NPC dataset (3.2), particularly through cost-effective models like Pegasus. Additionally, the potential of knowledge transfer fine-tuning across diverse dialogue datasets is underexplored, presenting an opportunity to enhance model performance even further. Fine-tuning a significantly lighter weight model like Pegasus allows you to achieve good performance for your niche summarizing use case, while also allowing cost effective deployment even on CPU settings. The novelty of this work lies in its systematic exploration of how fine-tuning and knowledge transfer fine-tuning impact the summarization capabilities of Pegasus across different dialogue datasets. Our findings seek to provide valuable insights into optimizing model performance for real-world applications, ultimately contributing to more efficient information extraction from dialogue-heavy texts.

3 Methodology

3.1 The Pegasus Model

The Pegasus model (9), which utilizes Gap Sentence Generation (GSG) during pre-training, masks entire sentences and requires the model to predict them based on surrounding context. This approach fosters the generation of coherent and contextually relevant summaries, making it a strong candidate for dialogue summarization tasks. By fine-tuning Pegasus on three diverse datasets, we aim to evaluate its performance across varied conversational contexts and assess potential improvements brought by knowledge transfer fine-tuning.

3.2 Datasets

The choice of datasets reflects the project's objective to benchmark the Pegasus model's adaptability across different conversational contexts:

- **SamSum** (4): Simulates typical messenger-like conversations, testing the model's ability to summarize everyday dialogue.
- **NPC (Character Dialogue)** (npc): Contains fictional character interactions, challenging the model to capture nuanced, story-driven exchanges .
- **Empathetic Dialogues** (emp) : Comprises emotion-rich conversations, assessing the model's capability to handle emotional subtleties in dialogues.

These datasets were meticulously selected to cover a broad spectrum of dialogue types, ensuring a comprehensive evaluation of the model's summarization performance across varied dialogue scenarios. Particular focus was also placed on the dataset size, in order to facilitate an effective and meaningful fine-tuning pipeline.

3.3 Model Training

3.3.1 Transformers Library

The training leveraged the robust capabilities of the HuggingFace Transformers library (8) which is operates as a high-level API that can wrap around frameworks like PyTorch. This choice facilitated efficient model initialization, tokenization, and training loop management using high-level functions:

- **AutoModelForSeq2SeqLM and AutoTokenizer**: These classes are pivotal for loading the Pegasus model and its tokenizer, gearing the model for high-performance sequence-to-sequence tasks such as summarization, translation, and question answering.
- **TrainingArguments**: This is an object that can be configured to optimize learning rate scheduling, apply weight decay, and implement gradient accumulation, which is crucial for managing larger batch sizes effectively within GPU memory constraints.

3.3.2 Fine-Tuning Stages

We employed two stages of fine-tuning to perform evaluations:

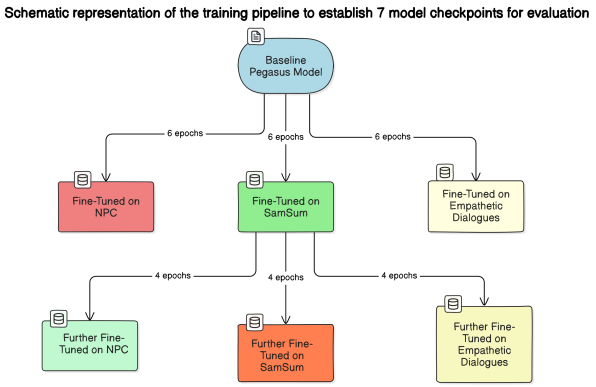

Figure 1: Schematic representation of the training pipeline to establish 7 model checkpoints for evaluation

1. **First Stage:** Pegasus was independently fine-tuned on each dataset for 6 epochs. This initial phase focused on adapting the model to the basic structural and thematic elements of conversational summarization. This creates three different checkpoints, one for each dataset.

2. **Second Stage:** We then leverage one checkpoint (fine-tuned on SamSum) from the first stage, further training for 4 epochs on all three datasets. This stage focuses on testing cross dataset generalizability and explores the impact of knowledge transfer (for the fine-tunes on NPC and Empathetic Dialogues datasets). This creates three more checkpoints.

A schematic representation of this is provided in Figure 1

Note: All training, testing, and inference pipelines were operated on the Snellius HPC cluster (7).

3.4 ROUGE Metric

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (5) metrics were employed to quantitatively assess the quality of summaries produced by the model. These measure the overlap of:

- **ROUGE-1**: Unigrams between the machine-generated summaries and the reference texts.

- **ROUGE-2**: Bigrams, offering insights into the sequential word-pair matching accuracy.

- **ROUGE-L**: The longest common subsequences, indicating the model's ability to maintain the coherence and order of information in summaries.

- **ROUGE-LSum**: Focused on evaluating the summarization of longer segments of text, capturing more comprehensive linguistic structures.

ROUGE metrics were also intermittently calculated during epochs in fine-tuning pipelines. This supplemented gradient loss visualizations to get a better idea of model convergence and potential overfitting.

4 Results and Discussion

The first stage of fine-tuning showed a positive performance improvement across all three datasets. The second stage showed improvements on further training on the SamSum dataset and knowledge transfer based fine-tuning on the NPC dialogues dataset. However, sequentially fine-tuning on the empathetic dialogues dataset caused a decrease in performance.

4.1 Results

4.1.1 Quantitative Investigation

Quantitative Enhancements: The ROUGE scores provide a clear, quantitative measure of improvement. As visualized in the graphs (Figures 2, 4, 3):

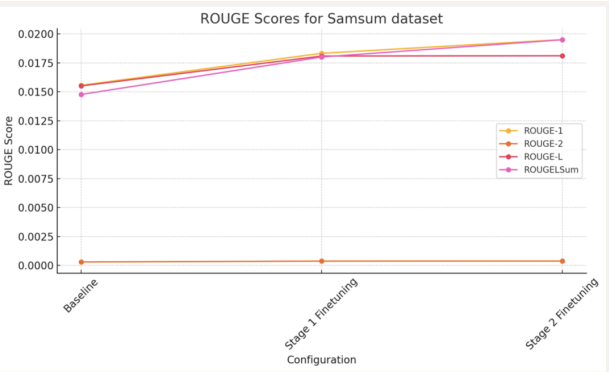

Figure 2: ROUGE scores for SamSUM dataset - The **SamSum dataset** stimulates a steady improvement for both stages of fine-tuning. This indicates that the model is progressively refining its ability to summarize messenger-like conversations with continuous epochs.

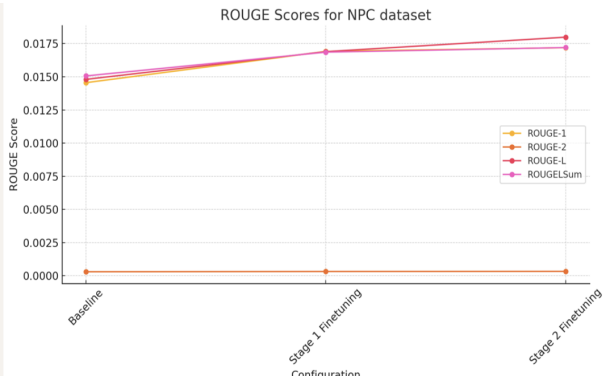

Figure 3: ROUGE scores for NPC Dialogue Dataset - The **NPC (Character Dialogue) dataset** also shows an improvement in performance for both stages. This indicates the feasibility of our sequential fine-tuning strategy and a meaningful transfer of knowledge from the SamSum fine-tuned checkpoint.

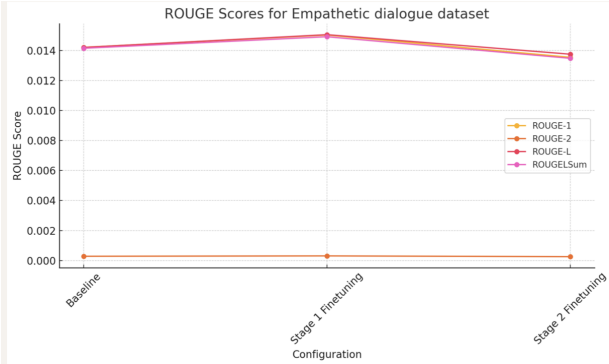

Figure 4: ROUGE scores for Empathetic Dataset - Contrastingly, the **Empathetic Dialogues dataset** presents a minimal increase in the first stage but a noticeable drop in the second stage. This drop suggests potential overfitting during the first stage or an inability of the model to further adapt to the emotional variability present in this dataset.

4.1.2 Qualitative Examples

Below are a literal comparison of inference examples from the various checkpoints for the datasets for human-driven evaluation. While the empiricism of human-driven evaluation is very low, we also identify that it is extremely difficult to effectively evaluate these results due to the low number of epochs of fine-tuning. We believe that training for a higher number of epochs will produce more differentiable results for human-understanding based benchmarks.

SamSum Dataset

- **Original Dialogue:**

  Hannah: Hey, do you have Betty's number?
  Amanda: Lemme check
  Amanda: Sorry, can't find it.
  Amanda: Ask Larry
  Amanda: He called her last time we were at the park together
  Hannah: I don't know him well
  Amanda: Don't be shy, he's very nice
  Hannah: If you say so..

Hannah: I'd rather you texted him
Amanda: Just text him
Hannah: Urgh.. Alright
Hannah: Bye
Amanda: Bye bye

- **Ground Truth Summary:**

  Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.

- **Baseline Model Summary:**

  Amanda: Ask Larry Amanda: He called her last time we were at the park together. Hannah: I'd rather you texted him. Amanda: Just text him.

- **Fine-Tuned on SamSum:**

  Amanda can't find Betty's number. Larry called Betty last time they were at the park together. Hannah wants Amanda to text Larry. Amanda will text Larry.

- **Further Fine-Tuned on SamSum:**

  Amanda can't find Betty's number. Amanda suggests to text Larry. Larry called Betty at the park. Hannah wants Amanda to text Larry.

NPC (Character Dialogue) Dataset

- **Original Dialogue:**

  painter: Yes sir! do you have a preferred color?
  worker: I am just a lowly worker. I think the red paint looks good. I can assist you with painting.
  painter: I will love that. Tell me, how long have you been here?
  worker: Not too long. The paintings in this Castle are magnificent.
  painter: Nice, there are a lot of secret about the castle. Take this maid's room for instance, heard there are so many secrets about it
  worker: Yes, In the village where I work, the other workers speak of some of the hidden secrets of this castle.
  painter: ok...i hope to get some truths however. We should get working, the night is upon us
  worker: Yes, we must get started but while we paint, it would not do any harm to keep an eye out for passages and secrets.
  painter: Great!
  worker: I will start on this side if you like. I will start taping the area for us.
  painter: Ok, I have this area covered. You do this with a level of enviable professionalism.

- **Ground Truth Summary:**

  Painter and worker are painting the maid's room in the castle. The worker will start on this side. The painter will cover the area.

- **Baseline Model Summary:**

  Painting secrets of a castle. Reveal the secrets of a maid's room. Use the Daily Discussion to help you understand today's featured news stories.

- **Fine-Tuned on NPC:**

  The painter and worker discuss the secrets of a castle. Paint the maid's room. The worker does no harm to keep an eye out.

- **Knowledge Transfer Fine-Tuning (Starting from SamSum):**

  The painter and worker speak the castle's secrets. Ready to paint the maid's room. The worker will start on one side.

Empathetic Dialogues Dataset

- **Original Dialogue:**

  I went to a new restaurant and was pleasantly surprised by the food. It was better than I expected.

- **Ground Truth Summary:**

  I went to a new restaurant and was pleasantly surprised by the food.

- **Baseline Model Summary:**

  I went to a new restaurant and was pleasantly surprised by the food. The food was better than I expected. It was better than I expected.

- **Fine-Tuned on Empathetic Dialogues:**

  I went to a new restaurant and was surprised by the food. The food was better than what I expected. It was better.

- **Knowledge Transfer Fine-Tuning (Starting from SamSum):**

  I went to a new restaurant and was surprised by the food. The restaurant was food. The food was better. It was better than I expected.

4.2 In-depth Discussion

Our experiments with fine-tuning the Pegasus model across three distinct dialogue datasets — SamSum, NPC (Character Dialogue), and Empathetic Dialogues — provided valuable insights into the model's adaptability and performance. The results from both stages of fine-tuning highlight the nuanced impact of dataset characteristics and the effectiveness of sequential fine-tuning (knowledge transfer fine-tuning).

SamSum Dataset as seen in Figure 2: The first stage of fine-tuning on the SamSum dataset showed a steady improvement in ROUGE scores, reflecting the model's ability to effectively summarize messenger-like conversations. The structured nature of SamSum dialogues, which often mimic real-world text exchanges, likely facilitated the model's learning process. During the second stage, where further fine-tuning was performed starting from the SamSum checkpoint, the model exhibited continued performance enhancement. This consistent improvement suggests that the SamSum dataset provides a solid foundation for the model to understand and generate coherent summaries, benefiting from additional training epochs. While our training pipeline only facilitated 10 epochs due to resource limitations, we estimate that a 10-20 fold increase in training time would warrant effective convergence.

NPC (Character Dialogue) Dataset as seen in Figure 3: In the first stage, fine-tuning on the NPC dataset led to notable improvements in ROUGE metrics, demonstrating the model's capacity to capture the nuances of story-driven, fictional character interactions. The second stage of fine-tuning, which involved continuing from the SamSum-trained checkpoint, also yielded positive results. This indicates successful knowledge transfer, where the contextual understanding gained from SamSum was beneficial when adapting to the NPC dataset. Theoretically, the sequential fine-tuning allowed the model to leverage previously acquired knowledge, enhancing its ability to summarize dialogues that have a different thematic and structural composition.

Empathetic Dialogues Dataset as seen in Figure 4: The first stage of fine-tuning on the Empathetic Dialogues dataset showed some initial improvements in ROUGE scores. The summary example provided in the qualitative section reflects the non-nuanced and unstructured inference of the model. It even seemingly becomes worse than the baseline through a human-understanding benchmark. However, the second stage of fine-tuning, starting from the SamSum checkpoint, resulted in a performance decline. This decline suggests potential overfitting during the first stage, where the model might have become too specialized on the subtleties of the SamSum dataset, limiting its ability to generalize to this one. Additionally, the high emotional variability and nuanced content in the Empathetic Dialogues dataset may have posed additional challenges.

5 Conclusion

This study investigates the impact of fine-tuning and knowledge transfer fine-tuning on the Pegasus model for dialogue summarization, addressing the unique challenges posed by informal and colloquial conversations across various domains. Our methodology, which involved systematically fine-tuning Pegasus on diverse datasets such as SamSum, NPC, and Empathetic Dialogues, demonstrated that dataset-specific tuning significantly enhances summarization performance, with knowledge transfer fine-tuning showing promising results, particularly with structurally similar datasets. The varying results across different datasets underscore the importance of adaptive fine-tuning strategies tailored to the specific characteristics of the dialogue data. We specifically identify challenges faced by our methodology in addressing emotional nuances in the data. Our approach contributes to the field by offering a cost-effective and scalable alternative for deploying high-performance summarization models, reinforcing the potential for practical applications in areas such as customer service, healthcare, and project management. Ultimately, our findings highlight the relevance of our work in advancing NLP techniques for real-world dialogue summarization tasks.

References

[emp] Empathetic dialogues summary. `https://huggingface.co/datasets/jtatman/empathetic_dialogues_summary`. Accessed: 2024-06-02.

[npc] Light dialogue summarization batch. `https://huggingface.co/datasets/npc-engine/light-batch-summarize-dialogue`. Accessed: 2024-06-02.

[3] Feng, X., Feng, X., and Qin, B. (2022). A survey on dialogue summarization: Recent advances and new frontiers.

[4] Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

[5] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[6] Singhal, D., Khatter, K., A, T., and R, J. (2020). Abstractive summarization of meeting conversations. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–4.

[7] SURF (2024). Snellius: the National Supercomputer. `https://www.surf.nl/en/services/snellius-the-national-supercomputer`. Accessed: 2024-06-02.

[8] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing.

[9] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2019). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.