A comparison between LLMs on the task of Text Summarization

Martina Ianaro
University of Bologna
martina.ianaro3@unibo.it

Anna Vitali
University of Bologna
anna.vitali7@unibo.it

Simone Branchetti
University of Bologna
simone.branchetti3@unibo.it

Abstract

In the modern age we are exposed to more and more information daily. If at first information is useful when confronting a problem, too much of it can lead to a sort of "paralysis" where a person can't decide what to do because he has too many options to choose from. This phenomenon takes the name of Information Overload and one proposed solution to this problem is text summarization. By condensing the information contained in one or more texts the user can absorb the same information quicker or just focus on the most important information. This short paper will focus on the task of text summarization by comparing how LLAMA3 and mBART perform on the same dataset. We will also talk about the problem of validation regarding this task. The results show that for every measure LLAMA3 performs better than mBART, even if the actual values are still poor.

1 Task Description

Text summarization is the task of producing a condensed version of all the information contained in one or more source documents. The task of text summarization is an important one, both for machine and human applications. In the former case, automatic summaries can be used to improve IR systems for example with query expansion (Strzalkowski et al., 1998) and in the latter text summarization can improve the problem of information overload (Bawden et al., 1999) that plagues our modern society. In practice, this task is divided into two subtasks: single document summarization and multiple document summarization. This paper will focus on single document summarization where only one document is considered as the source for the resulting summary produced. In general, the process of text summarization, as described by (El-Kassas et al., 2021) is articulated in three phases:

- Pre-processing: where the original text is cleaned, stop-words are removed and words are tokenized in order to produce a more structured version of the input.
- Processing: this phase is where the actual summarization happens using one or more summarization techniques.
- Post-Processing: here the results of the previous phase are improved by removing repetitions and/or reordering sentences.

The result of this process is, hopefully, a shorter text than the input which maintains the same information, is coherently written and doesn't fabricate any new information.

A brief state of the art for this task is as follows: the task has been researched for a long time, with articles on the topic dating as far back as the mid nineties (Cremmins, 1993; Maybury, 1995). Many researchers focused on different techniques to create summaries like Statistical-Based Methods (Afsharizadeh et al., 2018), Concept-Based Methods (Sankarasubramaniam et al., 2014), Topic-Based Methods (Sahni and Palwe, 2018), Sentence Centrality Based Methods (Roul and Arora, 2019) and Semantic-Based Methods (Mohamed and Oussalah, 2019). In more recent times, the research has shifted to machine learning assisted summarization, where a large language model performs the actual summarization. In (Goyal et al., 2022), the authors confront the performance of some custom trained models with GPT-3 on this task and they state that human readers prefer GPT's summaries by a large margin. This is not an isolated case, pushing the research in the direction of few-shots or zero-shot prompting.

Even though research around this task is very active, researchers have not found a consistent and objective way to measure the quality of the summaries produced by LLMs. The issue of validation is pressing because, without a unifying measure of quality, papers like (Fabbri et al., 2021) are needed once in

a while to compare different models and solutions. The existing ways of validating the quality of a summary are human evaluation (costly, non transferable and time consuming) or machine validation, of which the most popular measure is the ROUGE family. Recall-Oriented Understudy for Gisting Evaluation is a set of metrics used for the validation of summaries that boil down to confronting a produced summary with a set of human-produced summaries. In particolar, the metrics used in this paper from this family are ROUGE-1, ROUGE-2 and ROUGE-L. While the first two measure the co-occurrence of unigrams and bigrams respectively, ROUGE-L focuses on the longest commmon subsequence. The other measure we employed was the BLEU score (while commmonly used for machine translation, it compares the output to some pre-written target). The issues with all these measures is that creating a summary that is as similar as possible a human generated one is not really a measure of quality. Two summaries can be written very differently but they can contain the same information.

2 Dataset Description

The availability of datasets for performing text summarization in Italian is currently limited; only three datasets have been published, and they are accessible via the Hugging Face repository[1]: Fanpage(Landro et al., 2022a), Il Post(Landro et al., 2022b), and ITACASEHOLD(Licari et al., 2023).

For our research, we selected the Fanpage dataset, which features an average summary length of 1.96 sentences and 43.85 words(Landro et al., 2022c). Comprising 84.365 rows, this dataset was divided into training (67,492), validation (8,436), and test sets (8,437), and it was subject to analysis for the text summarization task by Nicola Landro et al(Landro et al., 2022c).

3 Architecture Overview

In order to approach this task we wanted to use established models fine-tuned for our purposes that we could run without access to the latest hardware. We settled on using two different models in order to compare their results: LLAMA3 and mBART.

3.1 LLAMA3

LLAMA 3 is one of Meta's latest open source large language model. It is advertised as a good model for several tasks, text summarization included. We specifically used the 8B parameters version of this model. We used the 8B parameters version of this model, as opposed to the 80B one because of the lesser requirements. The Llama family is a collection on models trained on trillions of tokens to be multipurpose models that can be fine-tuned to fit many tasks. Their biggest issues are their dimensions: they need a large amount of data to properly perform and they are quite large in size (only getting larger with newer models) meaning to properly run they need many resources. The main question we had regarding this model was its adaptability to different languages because our dataset is in Italian and this model has been trained on English text. With this model we aim to discover if fine-tuning enough can bridge the gap between languages and deliver promising results in our task.

3.2 mBART

The other model we explored is mBART, a sequence-to-sequence generative pretraining scheme first introduced in (Tang et al., 2020). This is a multilingual model built upon BART (Lewis et al., 2019), with the intent to expand its functionalities to more languages than just English. BART is a transformer with a bidirectional encoder and an auto-regressive decoder. We finetuned mBART using our dataset in order to find out its performance on our task. Since Italian is in the list of new languages we wanted to explore how this model adapts to a task in just one of the languages it supports.

4 Experimental Setup

4.1 LLAMA3

For the LLAMA3 Model the training process lasted 10 epochs, with a batch size per device set to 8 and a total batch size of 32. The total number of training steps was 60. The learning-rate was set to 1e-7, the optimizer used was AdamW and we adopted a number of gradient accumulation steps of 4. Additionaly a maximum sequence length of 6000 tokens was chosen by studying the length of articles and summaries before and after cleaning.

The LLAMA3 model was taken from the Hugging Face repository[2]. During the training execution, quantization was leveraged via the LoRA technique, configuring the model to operate at 4-bit accuracy over 1'000 items of train set 500 items of

[1] https://huggingface.co/datasets

[2] https://huggingface.co/unsloth/llama-3-8b

validation set and evaluating the performance on 100 items of the test set.

4.2 mBART

The configuration included a batch size of 22 and a micro-batch size of 8, with a training process of 10 epochs. We implemented a warm-up ratio of 0.1 to gradually introduce the learning rate and set the logging steps to 10 for regular updates during training. The gradient accumulation steps were calculated as the integer division of the batch size by the micro-batch size. Our learning rate was set to 1e-7, and we used Torch's AdamW optimizer. We have set the maximum length of the source sequence to 1024 tokens and the maximum length of the target sequence to 512 tokens.

The mBART model (Tang et al., 2020) was sourced from the Hugging Face repository[3]. It's fine-tuned over 1'000 items of train set and evaluated over 500 items of validation set. The evaluation metrics are computed over a test set of 100 items.

4.3 Hardware configuration

In order to conduct our tests, we had to obtain access to a suitable machine. We landed on StairwAI[4], which is managed by Professor Andrea Omicini and his team. The hardware configuration for our experiments, thus, consisted of an NVIDIA V100 GPU with 32GB of high RAM and CUDA version of 12.2, ensuring that we had sufficient computational resources to handle the training and evaluation processes efficiently.

4.4 Evaluation Metrics

To evaluate the models performance, we used several metrics: ROUGE-1, ROUGE-2, ROUGE-L, and BLEU. These metrics provided a comprehensive assessment of the model's accuracy in generating text. Specifically, ROUGE measures the overlap between the generated summary and the reference summary, while BLEU evaluates the generated text against the original text by scoring based on word overlap.

5 Results and Analysis

In this section, we compare the performance of two state-of-the-art models, LLAMA3 and mBART, for the task of text summarization.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
LLAMA3	0.095	0.042	0.068	0.014
mBART	0.217	0.092	0.153	0.039

Table 1: Performance metrics for the model over 100 items of cleaned test set.

The results, which are reported in table 1, highlight that mBART outperforms LLAMA3, despite its poor overall performance. There is a 0.312 gap above ROUGE-1. The difference is closing to 0.05 over ROUGE-2 and 0.085 for ROUGE-L. BLEU confirms the superior performance of mBART, which surpass LLAMA3 by 0.025.

Overall, the empirical evaluation highlights the superior performance of mBART in text summarization tasks, suggesting its greater capability in producing concise and relevant summaries compared to LLAMA3.

6 Conclusions

Text summarization is a very useful task for the general public and industries alike because it reduces the time needed to process information. In this paper we tested how two different large language models (an encoder based one and a decoder based one), with minimal fine-tuning fare against a task as complex as this one. The results are numerically lackluster because of a typical problem of text summarization: its validation measures are not really standardized and they all rely on an already written summary. Between LLAMA3, which is trained in English and we finetuned on an Italian task, and mBART, which is trained in over 50 languages (italian included) and we finetuned on the task, the latter achieved better results

References

Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, and Ayoub Bagheri. 2018. Query-oriented text summarization using sentence extraction technique. In *2018 4th international conference on web research (ICWR)*, pages 128–132. IEEE.

David Bawden, Clive Holtham, and Nigel Courtney. 1999. Perspectives on information overload. In *Aslib proceedings*, volume 51, pages 249–255. MCB UP Ltd.

Edward Cremmins. 1993. Valuable and meaningful text summarization in thoughts, words, and deeds.

[3] https://huggingface.co/facebook/mbart-large-50

[4] https://andreaomicini.apice.unibo.it/xwiki/bin/view/Project/StairwAI

Summarising Text for Intelligent Communication. Dagstuhl, Germany.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Nicola Landro, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022a. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5).

Nicola Landro, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022b. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5).

Nicola Landro, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022c. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5):228.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Daniele Licari, Praveen Bushipaka, Gabriele Marino, Giovanni Comandé, and Tommaso Cucinotta. 2023. Legal holding extraction from italian case documents using italian-legal-bert text summarization. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 148–156, New York, NY, USA. Association for Computing Machinery.

Mark T Maybury. 1995. Generating summaries from event data. *Information Processing & Management*, 31(5):735–751.

Muhidin Mohamed and Mourad Oussalah. 2019. Srl-esa-textsum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing & Management*, 56(4):1356–1372.

Rajendra Kumar Roul and Kushagr Arora. 2019. A nifty review to text summarization-based recommendation system for electronic products. *Soft Computing*, 23(24):13183–13204.

Aashka Sahni and Sushila Palwe. 2018. Topic modeling on online news extraction. In *Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017*, pages 611–622. Springer.

Yogesh Sankarasubramaniam, Krishnan Ramanathan, and Subhankar Ghosh. 2014. Text summarization using wikipedia. *Information Processing & Management*, 50(3):443–461.

Tomek Strzalkowski, Jin Wang, and G Bowden Wise. 1998. Summarization-based query expansion in information retrieval. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.