FRENCH CHEESE BEATS GOOGLE

Jordan Sassoon, Abhay Mathur
Institut Polytechnique de Paris
Palaiseau, France
{jordan.sassoon, abhay.mathur}@polytechnique.edu

ABSTRACT

Transfer learning has revolutionized the NLP landscape, allowing for massively pre-trained models to apply their understanding of natural language to downstream tasks. Abstractive summarization has been heavily impacted by large-scale models, though with a focus on the English language. Novel French-only pre-trained models set state-of-the-art performances on various NLP tasks, further proving the impact of transfer learning. In this report, we elicit our findings on news article abstractive summarization in French, testing pre-trained models and models trained from scratch. It is evident that pre-training offers a more robust and accurate basis for the task.

1 Introduction

The current landscape of digital news dissemination underscores the need for natural language processing (NLP) models capable of autonomously generating concise and informative headlines for news articles. The principal aim of this challenge is to develop and refine algorithms capable of identifying the essence of a given news article and condensing the same in a headline to aid users effectively. Given the rapid proliferation of news content online, the efficacy of a headline emerges as a pivotal determinant influencing reader engagement and content consumption.

This project focuses on building and evaluating various NLP solutions to abstractive summarization in French. In this variant of text summarization, the target summary is an interpretation of the source corpus that uses words that might not be present in the input. Transfer learning has been extensively applied to summarization, with multiple applications of the BERT [1] model. Massively training a bidirectional encoder transformer on a masked language modeling objective, BERT effectively reshaped many NLP tasks. Applications of BERT in the French language include CamemBERT [2] and FlauBERT [3]. A subsequent model, BART [4], showed that pre-training the autoregressive decoder significantly improves performance in generative tasks. BARThez [5] trains a BART-based model on a large French dataset.

This report continues with a summary and analysis of the dataset we worked with, derived from the News Article Title Generation Kaggle challenge [6]. The following sections describe the methods we applied to the task and the experiments we conducted. The results will be elicited along with a discussion. The conclusion summarizes the report.

2 Dataset

In this section, we provide an overview of the dataset utilized in the challenge. A first general introduction of the data showcases the objective, while the data analysis aims to reveal important information that we can induct as a prior in our pipelines.

2.1 Data

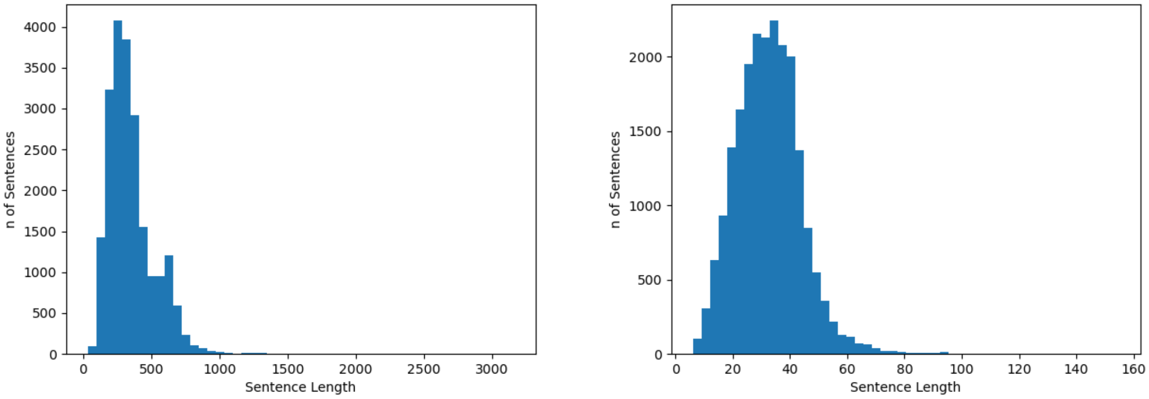
To perform news article summarization, the data is a collection of raw text French articles and their corresponding headlines. The train, validation, and test splits contain 21401, 1500, and 1500 article-headline pairs, respectively.

The dataset is relatively small, compared to datasets used by current state-of-the-art NLP models, therefore we can assume that a massively pre-trained model would have a more extensive understanding of French.

The text comes with escaped characters and French-specific characters, such as accented vowels. In our pipelines, we match the preprocessing to that with which the pre-trained models were trained. Paradoxically, preprocessing often results in lower performance, and many pre-trained models work better with unprocessed text, such as BERT. Due to the complexity of creating an effective preprocessing pipeline, when training a non-pre-trained model, we do not modify the text.

2.2 Data Analysis

We conducted simple data analysis to draw general information from the dataset. Since the data presented well-formatted text and non-topic-specific information, there was no need to engineer specific preprocessing or topic modeling in our pipelines. However, we analyzed the length of the input source text and the output headline text so as to inform the model during the encoding and decoding steps.



(a) Texts

(b) Titles

Figure 1: Distribution of sentence length for (a) input texts and (b) target headlines in the training set.

Figure 1 shows the distribution of sentence lengths for the input corpus and the headlines derived from the training data. Texts are on average 350 tokens long, and headlines are on average 32 tokens long. This information is instrumental to guide the models to expect long sequences, and generate much shorter ones. Furthermore, batching with padding and truncation depends on how short we want our inputs to become.

3 Method(s)

This section describes in detail the methods we implemented. The following three methods will be explained: a self-attention encoder-decoder architecture trained from scratch, a finetuned T5 model, and a finetuned CamemBERT model. The methods are presented in increasing order of performance.

3.1 Self-Attention Encoder-Decoder

The first model we implemented was a self-attention encoder-decoder which was trained from scratch, coded following [7]. The encoder and decoder are GRU models, and the decoder uses a Bahdanau attention mechanism to focus on the encoder's outputs selectively. The data was prepared accordingly, using "<SOS>" and "<EOS>" as start-of-article and end-of-article tokens respectively. As opposed to the pre-trained models, the vocabulary for this model was constructed solely from the challenge's dataset. The input corpus was tokenized and padded or truncated to 512 tokens, since longer input sequences prompted more compute, and most articles are less than 512 tokens long. The model used the negative log-likelihood loss to perform gradient descent. Initially, this was thought to act as a baseline for the other methods, though in practice, the model struggled to converge and produce valuable results.

3.2 Finetuning T5

Since training a fully-supervised model from scratch proved quite challenging, we then focused on tailoring pre-trained models to our task. Pre-training leverages a more extensive understanding of natural language, often translating into better performance on a variety of downstream tasks. T5 [8] is an encoder-decoder pre-trained model on a mixture of supervised and self-supervised tasks. The effectiveness of transfer learning renders T5 a powerful framework for many downstream tasks, including summarization in non-English languages, such as French. To prepare the data, we prepended the prompt "summarize: " to each sample in the corpus to instruct the model of the relevant task. Training the T5 model [8] was then conducted through PyTorch's Lightning [9] module.

3.3 Finetuning CamemBERT and BARThez

Multilingual models like T5, however, have a tendency to perform suboptimally compared to monolingual models, especially for high-resource languages. We therefore compare T5's performance with that of a sequence-to-sequence model based on CamemBERT [2], which is pre-trained solely on French and would intuitively perform at least at par with the former using fewer resources. CamemBERT uses the RoBERTA [10] architecture and was pre-trained using a web-crawled corpus on the Masked Language Modeling (MLM) task. The model uses a CamemBERT backbone for both the encoder and decoder, with shared weights.

Similarly, BARThez [5] is a a French sequence-to-sequence pretrained model based on BART. It is pretrained on the corrupted sentence reconstruction task and is therefore particularly well suited for generative tasks.

3.4 Reinforcement Learning

Taking inspiration from recent works that report improved performances by training language models directly on non-differentiable metrics such as human feedback, textual entailment and ROUGE and BLEU scores [11, 12, 13, 14], along with the fact that the formulation of the challenge significantly favors superior ROUGE-L F1 scores, we attempt to improve the performance of the aforementioned models by further finetuning them using Reinforcement Learning.

Our formulation involves generating headlines, evaluating them against the ground truth and using the computed metrics 4.1 as rewards for the model. The original weights of the model are used as a reference in an additional reward signal, discouraging the model from deviating drastically using the KL divergence between the two. Following [14], we also consider a neural feedback-based reward prioritizing textual entailment between the article and the predicted headline. The final reward is thus the algebraic sum of the following :

$$R_{\text{metric}} = \text{ROUGE-L}_{F1} \tag{1}$$

$$R_{\text{entailment}} = \text{NLI}(x, \hat{y}) \tag{2}$$

Where NLI is a Natural Language Inference model.

The augmented reward is then used to train the model through Proximal Policy Optimization (PPO) [15]. For training with policy gradients, we need an estimation of the value function for each predicted token in the output sequence, and therefore, a value head is added to the model.

3.4.1 Entailment Feedback

The textual entailment feedback, as described above, is obtained from an NLI model. To this end, we use a CamemBERT instance pretrained for inference. The model follows the output structure of the XNLI dataset [16]. We treat the text t as the premise and its predicted headline h as the hypothesis for the entailment task, and CamemBERT$_{\text{NLI}}$ models entailment as the likelihood:

$$\text{NLI}(t, h) = p(t \Rightarrow h) \tag{3}$$

The expected behavior of our NLI model is illustrated in Table 1.

4 Experiments

Three main approaches were investigated, as described in the previous sections: training a self-attention encoder-decoder model from scratch, fine-tuning the pre-trained multilingual T5 model, and fine-tuning the French-specific CamemBERT and BARThez models.

Premise p	**Hypothesis** h	**NLI**(p, h)
Le score pour les bleus est élevé	*L'équipe de France a fait un bon match*	0.872
	l'équipe de France va sûrement perdre le match aujourd'hui.	0.015
	La Seine a plus de 14 000 ans.	0.247

Table 1: Entailment scores estimated by our NLI model for the same premise with three different hypotheses. It can be seen that positive entailment receives a high score, while neutral and negative entailments receive low ones, and their relative magnitude indicates that entailment feedback will penalize factual and logical contradictions more than neutral summarizations.

Additionally, a reinforcement learning technique was explored to further improve the fine-tuned models by incorporating reward signals based on the ROUGE-L metric and textual entailment between the source and summary. All experiments were conducted on an Nvidia A100 GPU.

4.1 Metrics

Our evaluation metrics are in line with the challenge. To evaluate our models, we use the Rouge-L F-Score metric:

$$\text{ROUGE-L}_{precision} = \frac{\sum_{s \in S} LCS(s, g)}{\sum_{s \in S} |s|} \tag{4}$$

$$\text{ROUGE-L}_{recall} = \frac{\sum_{s \in S} LCS(s, g)}{\sum_{g \in G} |g|} \tag{5}$$

$$\text{ROUGE-L F-Score} = \frac{2 \cdot \text{ROUGE-L}_{precision} \cdot \text{ROUGE-L}_{recall}}{\text{ROUGE-L}_{precision} + \text{ROUGE-L}_{recall}} \tag{6}$$

where $LCS(s, g)$ represents the Longest Common Subsequence between ground truth headline s and generated headline g. To compare the models, we also used the ROUGE-1 and ROUGE-2 metrics, which compute the F1-Score relating to the number of common unigrams and bigrams between predictions and ground truths. Furthermore, we kept track of the compute required by each model.

4.2 Results

We provide a comparison of the ROUGE-L F1 Score for our models in Table 2. Specifically, for submissions to the challenge, we consider three architectures - T5, CamemBERT Seq2Seq and BARThez. For the first two, we compare performance with and without further training on ROUGE-L and Entailment Feedback.

We see that variants without the additional RL formulation perform better, and the Finetuned CamemBERT attains the highest ROUGE score, marginally beating BARThez, followed by T5.

Method	**ROUGE-L F1**
T5	0.203
CamemBERT	**0.227**
BARThez	0.226
T5 + PPO	0.110
CamemBERT + PPO	0.125

Table 2: ROUGE-L F1 Scores of our methods on the validation set.

4.2.1 Issues with Encoder-Decoder Training

As opposed to the other models, the first encoder-decoder model is trained solely on this dataset. It is therefore possible that the dataset is not large enough for the model to learn to generate valuable headlines. We found

Ground Truth	Generated Headline
Selon France Inter, le ministre de l'Intérieur et l'Élysée ont eu vent des faits concernant Alexandre Benalla dès le 2 mai, soit le lendemain	Le Parisien s'est retrouvé nez <EOS>
Le Premier ministre libanais désigné a annoncé samedi renoncer, faute de consensus, à former un nouveau gouvernement destiné à sauver le pays d'une des pires crises économiques de son histoire près de deux mois après l'explosion au port de Beyrouth.	Le président Donald Trump a fait quatre morts dans le pays. <EOS>
Dans une interview accordée au magazine Point de vue, la princesse raconte que sa fille Gabriella a failli se noyer dans une piscine.	SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS SOS

Table 3: Comparison between ground truth headlines and generated headlines derived from our encoder-decoder model.

that convergence was complicated to achieve, and the model often resorted to ignoring the encoder's outputs and focusing on the previous decoding steps. The resulting headlines were repetitive and uninformed (see Table 3). For these reasons, we do not report ROUGE-L F-Score for our encoder-decoder architecture.

4.2.2 Issues with Learning from ROUGE and Entailment Feedback

As is evident from our reported results, the addition of reinforcement learning step to our overall training pipeline did not improve performance on headline generation. We attribute this to the small size of the dataset (for a language model), due to which the added value head in the decoder does not have sufficient samples to converge. Further, if training is conducted over a large number of steps, then because the value function is entirely arbitrary at the beginning of the PPO training loop, the model might even deteriorate a little due to a 'forgetting' of optimal pre-trained weights over a vast number of poor-performing (low reward-value coherence) episodes.

4.2.3 Evaluation Metrics

Table 4 illustrates the disparity in evaluating a sequence-to-sequence models solely through n-gram matching metrics such as ROUGE and BLEU. Due to their reliance on exact word/phrase matches, these metrics fail to capture semantic equivalence and factual consistency, which are crucial for evaluating the quality of generated headlines. They prioritize lexical similarity over semantic similarity, leading to low scores for headlines that convey the same meaning as the ground truth but through different phrasing.

We posit that metrics like BERTScore [17], which leverage contextual embeddings from pre-trained language models like BERT, can better capture semantic similarity between the generated headline and the ground truth.

5 Conclusions

In this report, we explored various methods for the task of French news article abstractive summarization. We implemented and compared three main approaches: training a self-attention encoder-decoder model from scratch, finetuning the pre-trained multilingual T5 model, and finetuning the French-specific CamemBERT and BARThez models. The encoder-decoder model trained from scratch on the limited dataset struggled to converge and produced poor-quality summaries, indicating the dataset was likely too small to effectively train a model from scratch. In contrast, the pre-trained T5 and CamemBERT models demonstrated significantly better performance, with CamemBERT outperforming BARThez and T5 slightly on the ROUGE-L metric. This highlights the substantial

Headline $\hat{h}$	**NLI**$(t, \hat{h})$	**ROUGE-L**$_{F1}(\hat{h}, h)$
(GT) *Le bateau de croisière, long de 275 m, a percuté un quai lors de son arrivée dans le port de Venise, dimanche 2 juin. Quatre personnes ont été blessées.*	0.898	**1.0**
(T5) *L'accident s'est produit à San Basilio-Zaterre, dans le sud de l'Italie. Quatre personnes ont été blessées, tandis que deux autres ont été transportées à l'hôpital.*	**0.998**	0.327
(CamemBERT) *L'équipage du MSC Opéra a perdu le contrôle du paquebot, à son arrivée dans le port de Venise. Des témoins ont filmé la scène.*	0.983	0.285
(Text t) *Sur les réseaux sociaux, les images sont impressionnantes. Dimanche matin à Venise, l'équipage du MSC Opéra a perdu le contrôle du paquebot, à son arrivée dans le port de la cité des Doges. Le navire, qui peut contenir plus de 2.600 passagers, est venu heurter le quai auquel il voulait s'arrimer...*		

Table 4: Comparison of Entailment scores and ROUGE-L F1 Scores among ground truth and headlines predicted by our finetuned T5 and CamemBERT models. It is evident that even factually consistent headlines receive a poor ROUGE score due to the absence of the exact keywords/phrases as used in the ground truth.

benefit of leveraging large pre-trained models and transfer learning for the summarization task, especially for non-English languages. We further attempted to improve performance using reinforcement learning techniques with reward signals based on the ROUGE-L metric and textual entailment between the source and summary. However, we did not observe gains from this approach, which we attribute primarily to the limited dataset size. Overall, our results demonstrate the powerful capabilities of large pre-trained language models like CamemBERT and BARThez for the French summarization task when finetuned on a downstream dataset.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *CoRR*, abs/1911.03894, 2019.

[3] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french. *CoRR*, abs/1912.05372, 2019.

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

[5] Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*, 2020.

[6] christos.x. News Article Title Generation. https://www.kaggle.com/competitions/inf582-news-articles-title-generation, 2024.

[7] Sean Robertson. NLP From Scratch: Translation with a Sequence to Sequence Network and Attention. https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html, 2024.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[9] William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[11] Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3110–3120, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.

[13] Heewon Jang and Wooju Kim. Reinforced abstractive text summarization with semantic added reward. *IEEE Access*, 9:103804–103810, 2021.

[14] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. Factually consistent summarization via reinforcement learning with textual entailment feedback, 2023.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[16] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations, 2018.

[17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.