# *News articles summarization*
# Machine Learning for Natural Language Processing 2022

**Gaspard Michel**
ENSAE, 3A
gaspard.michel@ensae.fr

**Dounia Chekembou**
ENSAE, MS Data Science
dounia.chekembou@ensae.fr

## Abstract

In this project, we present one extractive method and two abstractive methods for text summarization. These methods are applied on pairs of news articles and their human written summaries. Their performances are compared based on ROUGE scores, the gold standard metric for summarization tasks.

## 1 Problem Framing

Summarization is the act of reducing the size of a text by keeping only the most important information. The problem we address in this project is the synthesizing of diverse articles. We are using the CNN/DailyMail dataset which contains 287,000 training pairs of news article and their human written summaries. The validation set contains around 13,000 pairs and the test set around 11,000. The dataset was presented in (Nallapati et al., 2016) that used a Sequence 2 Sequence (Seq2Seq) architectures built on Recurrent Neural Networks for this summarization task. Our goal is to first understand how a classic baseline method can perform, and then compare it to a Seq2Seq Transformer model (Vaswani et al., 2017) trained from scratch. We decided to use a Transformer model as it has shown state-of-the-art performances for this task[1] .We also compare it to a pre-trained BART model (Lewis et al., 2019), finetuned on the CNN/DM dataset. Performances are evaluated quantitatively using ROUGE scores (Lin, 2004) that calculates overlapping n-grams between predicted and true summaries. The chosen baseline model produces extractive summaries: a particular set of sentences available in the articles are chosen to form a summary. Conversely, the other architectures are abstractive: they generate directly a summary by just using the encoded version of the summary. We also investigate qualitatively models predictions by looking at predicted against reference summaries.

## 2 Data cleaning and exploration

The only data processing made was to replace line breaks with dots on both articles and summaries. On average, training and validation articles contain around 803 words spanned in 38.57 sentences. Summaries are shorter and contain on average 57 words spanned in 3.84 sentences. This dataset contains very long documents that must be summarized in around 10 times less sentences and more than 10 times less words. Histograms and more elements on the number of words and sentences can be found in Appendix A.

## 3 Methodology

### 3.1 Baseline: Lead-3

We use lead-3 as a baseline method. Introduced in (See et al., 2017), this very simple method uses only the first three sentences of the articles as a summary. It demonstrates really strong performances (as measured by the ROUGE score) such that RNN-based abstractive models were not able to outperform it. (See et al., 2017) observed that news articles are structured and tend to contain most important information at the start, explaining the strength of lead-3.

### 3.2 Seq2Seq Transformer

Our Sequence 2 Sequence model builds on Transformers (Vaswani et al., 2017). It takes as input sequences of words (converted to token ids) and various token masks and returns a decoded sequence containing the predicted summaries. The tokens masks play the role of an auto regressive decoding strategy: at each point in time, the decoder is only allowed to attend to all positions in

[1] A benchmark is available here.

the encoder and only the previous positions in the decoder. This preserves the model to attend to future tokens that it needs to predict. We use a smaller architecture than (Vaswani et al., 2017) in order to have manageable computation times: we use only a stack of $N = 2$ encoders and decoders. The size of the embeddings is set to $d_{model} = 512$, the number of attention heads to $h = 8$ and the dimension of the position-wise fully connected layer to $d_{ff} = 2048$ as in the original paper. The final model contains $173,521,028$ trainable parameters. Besides, input articles are truncated to 800 tokens and input summaries to 100 words. The maximum articles vocabulary size is set to 150000 words and 80000 for the summaries. We use a slightly different learning rate scheme than the original paper: we also use an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ but the learning rate increases linearly to $d_{model}^{-0.5}$ for the first 10000 warmup steps, and then decreases proportionally to the inverse square root of the step number. The loss function is the Cross Entropy (between predicted and true tokens). We iterate through the training data for 12 epochs by batch of size 16. At each epochs, the training set is shuffled. We compute the validation loss at the end of each epochs and stop the training if it stops decreasing for one epoch. Each epoch took approximately 1h40 to run on Google Colab. At decoding time, we use greedy search to build predict summaries. We also implemented beam search but it increases inference time to such an extent that we could not use it (to match the report deadline).

### 3.3 Bidirectional and Auto-Regressive Transformers (BART)

BART builds upon a denoising autoencoder architecture which allows to pre-train Seq2Seq models (Lewis et al., 2019). It demonstrates state-of-the-art performances when fine-tuned for text-generation tasks. Particularly, it outperformed a wide range of previous works on text summarization on the CNN/DM dataset. We produced test summaries using Facebook implementation of fine-tuned BART[2] on the CNN/DM dataset. Inference took around 1h50 on Colab.

## 4 Results

Our full implementation is available here. The performances are calculated using *Recall-Oriented Understudy for Gisting Evaluation* (Lin, 2004) scores, the gold standard for evaluating an automated summarization tool. We use the ROUGE-1, ROUGE-2 and ROUGE-L scores which are widely use in the literature. ROUGE-1 calculates the F-measure of overlapping words between a source and predicted summary. ROUGE-2 calculates the F-measure of overlapping bigrams and ROUGE-L calculates the F-measure of the Longest Common Subsequence between each sentence in the source summary and the predicted summary. It reflects sentence-level word ordering. Results are presented in Table 1. Lead-3 baseline performs extremely well but is still outperformed by BART. This is quite surprising since lead-3 is an extractive method which might be favored by the ROUGE scores. Indeed, lead-3 summaries contain multi-word named entities (such United-State, United-Nations) which are harder to produce by abstractive methods, resulting in a higher number of overlapping n-grams. It indicates that state-of-the-art summarization models are able to produce grammatically correct and useful summaries. In contrast, our seq2seq model fails to find good summaries. We believe that the main reason is underfitting. We had to stop training at 13 epochs to match the report deadline (we were also limited by Colab since the training took more than 20 hours). Training and validation losses were still decreasing, as shown in appendix B. Various predicted summaries are presented in Appendix C.

|  | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** |
|---|---|---|---|
| Lead-3 | 0.401 | 0.175 | 0.363 |
| Seq2Seq | 0.238 | 0.05 | 0.222 |
| BART | 0.440 | 0.21 | 0.408 |

Table 1: ROUGE scores of the different models on the CNN/DM test set.

In this project, we trained a Transformer based Seq2seq model from scratch to summarize news articles. We compared it to a very simple baseline, lead-3, that performs really well. We also compared it to the state-of-the-art, BART, that produces both grammatically and content-wise correct summaries. BART performances show that automatic summarization tasks can largely be handled by current state-of-the-art deep learning models.

[2] Freely available on Huggingface platform.

## References

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

# A Word and sentences counts in the CNN/DM dataset

| | Train + Valid | Test |
|---|---|---|
| # documents | (287113, 13368) | 11490 |
| avg # words in article | 802.84 | 791.39 |
| avg # sent in article | 38.57 | 33.33 |
| avg # words in summary | 56.78 | 59.81 |
| avg # sent in summary | 3.84 | 3.92 |

Table 2: Various CNN/DM elements on train, validation and test set. The average number of words includes punctuation. The number of sentences was calculated using python library NLTK's sentence tokenizer.

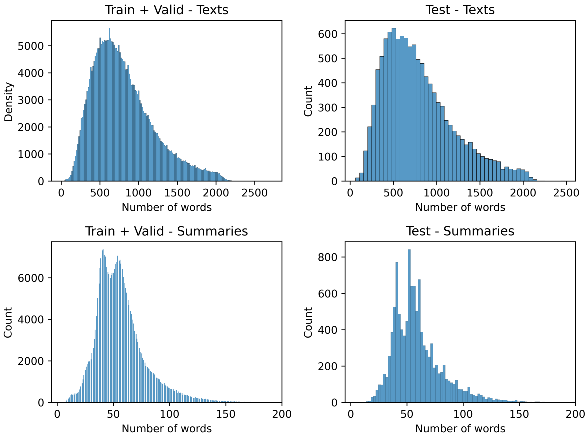

Figure 1: Histogram of word counts in the CNN/DM dataset. Counts were calculated on the concatenation of train and validation dataset and on the test set.

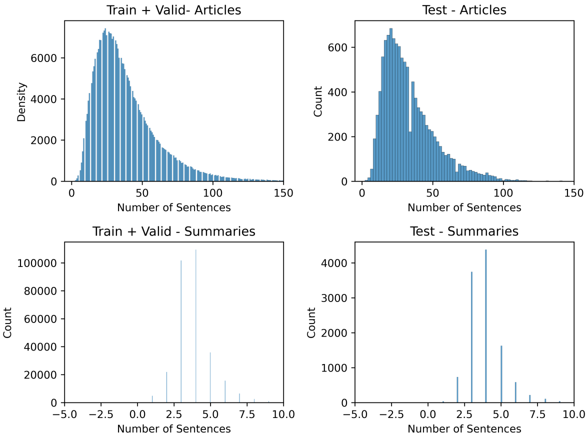

Figure 2: Histogram of sentence counts in the CNN/DM dataset. Counts were calculated on the concatenation of train and validation dataset and on the test set.

# B Training and validation losses

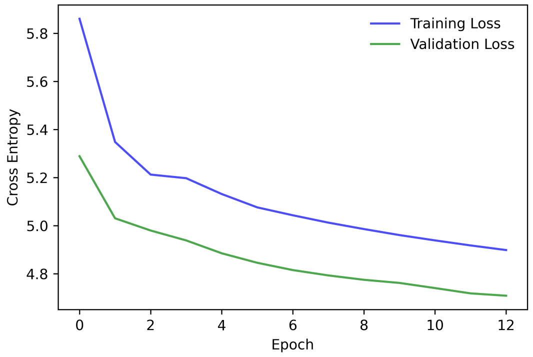

Figure 3: Cross entropy loss of train and validation sets during training. This clearly indicates that our model is underfitting since the losses are still decreasing by quite a lot even at the last epoch.

# C Predicted Summaries

| | |
|---|---|
| Reference article | The build-up for the blockbuster fight between Floyd Mayweather and Manny Pacquiao in Las Vegas on May 2 steps up a gear on Tuesday night when the American holds an open workout for the media. The session will be streamed live across the world and you can watch it here from 12am UK (7pm EDT). |
| Reference | Floyd Mayweather holds an open media workout from 12am UK (7pm EDT). The American takes on Manny Pacquiao in Las Vegas on May 2. Mayweather's training is being streamed live across the world. |
| Lead-3 | The build-up for the blockbuster fight between Floyd Mayweather and Manny Pacquiao in Las Vegas on May 2 steps up a gear on Tuesday night when the American holds an open workout for the media. The session will be streamed live across the world and you can watch it here from 12am UK (7pm EDT). |
| Seq2Seq | Floyd Mayweather will be open to the fight on Tuesday night. The fight will be open in Las Vegas. Mayweather will be open to the fight Floyd Mayweather and Manny Pacquiao. |
| BART | Floyd Mayweather will hold an open workout for the media on Tuesday night. The session will be streamed live across the world and you can watch it here from 12am UK (7pm EDT). Mayweather and Manny Pacquiao will fight in Las Vegas on May 2. |

Table 3: Predicted summaries of the various models for a small size article (63 words spanned over 2 sentences). For small sizes summaries, Lead-3 just take the full article as a summary.

| | |
|---|---|
| Reference article | (CNN)A judge this week sentenced a former TSA agent to six months in jail for secretly videotaping a female co-worker while she was in the bathroom, prosecutors said. During the investigation, detectives with the Metro Nashville Police Department in Tennessee also found that the agent, 33-year-old Daniel Boykin, entered the woman's home multiple times, where he took videos, photos and other data. Police found more than 90 videos and 1,500 photos of the victim on Boykin's phone and computer. The victim filed a complaint after seeing images of herself on his phone last year. Boykin plead guilty to unlawful photography, aggravated burglary and violation of the computer act, the Nashville District Attorney's Office said. Police said the incident happened in a TSA-only restroom, and that there was no evidence public restrooms were targeted. A TSA official tells CNN that Boykin worked in an administrative capacity and didn't engage in public security screening. Assistant District Attorney Amy Hunter said this case was one of the worst invasion of privacy cases she's seen. "We are thankful that the sentence includes periodic confinement so that the sentence will hopefully make an impression on this defendant and others," Hunter said in a statement. The judge, Randall Wyatt, on Friday called the invasion of privacy "egregious." His sentence also includes five and a half years of probation, which will include GPS monitoring. Boykin was terminated last year when the investigation began. "TSA holds its employees to the highest ethical standards and has zero tolerance for misconduct in the workplace," TSA's Ross Feinstein said in a statement. |
| Reference | Former TSA agent Daniel Boykin, 33, videotaped his female co-worker in the restroom, authorities say. Authorities say they found 90 videos and 1,500 photos of the victim on Boykin's phone and computer. Boykin worked in an administrative capacity and didn't do public security screenings, TSA official says. |
| Lead-3 | (CNN)A judge this week sentenced a former TSA agent to six months in jail for secretly videotaping a female co-worker while she was in the bathroom, prosecutors said. During the investigation, detectives with the Metro Nashville Police Department in Tennessee also found that the agent, 33-year-old Daniel Boykin, entered the woman's home multiple times, where he took videos, photos and other data. Police found more than 90 videos and 1,500 photos of the victim on Boykin's phone and computer. |
| Seq2Seq | NEW : The judge says the victim was " very seriously ". NEW : The judge says the victim 's privacy is ". NEW : The judge says the worst case is ". NEW : The victim 's privacy is " |
| BART | Daniel Boykin, 33, entered the woman's home multiple times, where he took videos, photos and other data. Police found more than 90 videos and 1,500 photos of the victim on Boykin's phone and computer. The victim filed a complaint after seeing images of herself on his phone last year. |

Table 4: Predicted summary of the various models for a medium size article (307 words spanned over 13 sentences). BART seems to produce a more concise summary with more than enough information.

| | |
|---|---|
| Reference article | Saracens flanker Jacques Burger has been cited for alleged foul play during last Sunday's European Champions Cup quarter-final game against Racing Metro in Paris. European Professional Club Rugby said that Burger is alleged to have struck Racing scrum-half Maxime Machenaud midway through the match at Stade Yves-du-Manoir, which Saracens won 12-11. The complaint was made by match citing commissioner Eugene Ryan. Jacques Burger (right) arrives on the scene as Maxime Machenaud (left) prepare to play a diving pass. Burger then catches Machenaud after the ball has gone but the player was low to the floor. Namibia international Burger, 31, will face a disciplinary hearing on Thursday. Welshman Simon Thomas has been appointed as independent judicial officer. World Rugby's low end punishment for such an offence is a two-week ban, with anything from eight weeks to 52 weeks at the top end. Saracens return to Aviva Premiership action against play-off rivals Leicester on Saturday, before tackling Champions Cup semi-final opponents Clermont Auvergne in Saint Etienne on April 18. Burger gets to grips with Racing Metro full back Brice Dulin during Saracens' dramatic victory. Marcelo Bosch celebrates after his last-gasp kick sent Saracens through to the Champions Cup semi-finals. |
| Reference | Saracens beat Racing Metro 12-11 to advance to Champions Cup semis. Racing were unhappy about a challenge on Maxime Machenaud. Saracens flanker Jacques Burger has been cited for the tackle. |
| Lead-3 | Saracens flanker Jacques Burger has been cited for alleged foul play during last Sunday's European Champions Cup quarter-final game against Racing Metro in Paris. European Professional Club Rugby said that Burger is alleged to have struck Racing scrum-half Maxime Machenaud midway through the match at Stade Yves-du-Manoir, which Saracens won 12-11. The complaint was made by match citing commissioner Eugene Ryan. |
| Seq2Seq | Saracens have been accused of sending off the match against Saracens. Saracens have been accused of sending off the match - fixing. Saracens to the match against Saracens. Saracens have been sent off for two weeks. |
| BART | Jacques Burger is alleged to have struck Racing scrum-half Maxime Machenaud midway through the match at Stade Yves-du-Manoir. The complaint was made by match citing commissioner Eugene Ryan. Burger, 31, will face a disciplinary hearing on Thursday. World Rugby's low end punishment for such an offence is a two-week ban, with anything from eight weeks to 52 weeks at the top end. |

Table 5: Predicted summary of the various models where BART produces poor performances (ROUGE-1: 0.245, ROUGE-2: 0.042, ROUGE-L: 0.224). In comparison, Lead-3 performances are (0.408, 0.229, 0.347) and Seq2Seq performances are (0.182, 0.0, 0.182)