Text Summarization applied to French news articles
Machine Learning for Natural Language Processing 2022

Giuseppe Deni
giuseppe.deni
@ensae.fr

Béline Aubergeon
beline.aubergeon
@ensae.fr

Solène Blasco Lopez
solene.blascolopez
@ensae.fr

Abstract

This project will focus on the text summarization task applied to the summarization of French written documents. Formally, our goal will be to implement and compare different approaches to produce a summary from a given text. We adopted extractive (TextRank algorithm) and abstractive (mainly BARTHez model), trained and tested on OrangeSum dataset, with both quantitative (using ROUGE metric) and qualitative experiments.

Our code is available in a github repository [1].

1 Problem Framing

1.1 Pre-processing protocol

The pre-processing is detailed in the notebook attached to this report. The main protocol was to lower our texts, remove special characters, and remove french stopwords.

1.2 Extractive approach - TextRank

As a ”baseline”, we implemented a ”naive” method to address the summarization task : the TextRank algorithm (Mihalcea and Tarau, 2004), an unsupervised ”Extractive Approach”. The idea is to detect and extract the most revelant sentences of the text as a summary.

We first need to define a similarity measure ϕ between two sentences. During our experiments, our approach was to numerize our sentences, using pre-trained embeddings such as FastText ones, and to use a cosine similarity, defined as :

$$\forall u, v \in \mathbb{R}^d : \phi(u, v) := \frac{< u; v >}{\|u\| \|v\|}$$

We then iterate the following scoring procedure until convergence, starting from uniform scores :

$$\psi(S_i) = (1-\gamma) + \gamma \sum_{j \neq i} \left(\frac{\phi(S_i, S_j)}{\sum_{k \neq j} \phi(S_k, S_j)} \right) \psi(S_j)$$

with $\psi(S_i)$ denoting the score of sentence i and $\gamma \in (0, 1)$ to tune.
Our summary is then constructed by extracting from the text the 2-th sentences with highest score.

1.3 Abstractive approaches

We also studied ”Abstractive approaches”, for which the idea is to leverage deep learning models to build supervised generative models that will create a summary by generating new sentences.

The architecture of abstractive models relies upon the encoder-decoder structure: the encoder understands the inputs and represent them in a latent (hidden) space, and the decoder is feed by the encoder and reverse its mapping. The role of the encoder is to encode the whole input text, taking into account the context, while our decoder has here a feed-forward structure so that is can generate word by word a summary.

For the purpose at hand, we started by implementing from scratch a mode, using a two-layer bidirectional RNN with LSTMs in the encoder and a decoder which is a seq2seq model with two layers, each one with an LSTM, enriched with a Bahdanau attention on the target data.

The above model being very time-demanding in the training phase, we have also used used a BARTHez pre-trained to be able to do better comparisons. BARTHez (Eddine et al., 2020) is a encoder-decoder model for French texts which uses the attention mechanism of transformers, known to better encode context. The main idea behind this model is also to use a bidirectional encoder (similar as BART/BERT architectures), and a left-to-right auto-regressive decoder.

[1] https://github.com/gdeni89/NLP-summarization-of-French-written-documents

2 Experiments Protocol

2.1 Description of OrangeSum dataset

This project mainly focuses on OrangeSum dataset, introduced together with BARTHez abstractive model in (Eddine et al., 2020). It is composed of pairs of French articles and summaries, scrapped from ”Orange Actu” website.

The descriptive statistics made on the training set of OrangeSum are presented in appendix A.

- As seen in figure 1, articles and summaries are relatively shorts (up to 600 and 60 words). The number of words per sentences are similar between articles and summaries, which may been a plus for extractive models.

- In both figure 4 and 3, we can see that the politic and society topics are omnipresent. Indeed, among most frequent words we have : ”president”, ”gouvernement”, ”ministre”, ... We may think that it would be a source of hardness, above all for abstractive supervised models, as they are ”subjective” and ”oriented”/”partial” articles.

2.2 Experiment protocol

For the experiments, we mainly concentrated on a subpart of OrangeSum validation set.

To quantitatively compare our models, we used ROUGE [2] metrics (Lin, 2004). We concentrated on ROUGE-N ($N \in \{1, 2\}$) corresponding to the overlap of N-grams between two summaries.

We also used a pre-trained topic classification model (see details in our notebook), to compute the proportion of summaries for which the main topic attributed coincides with the one attributed to the source article.

As qualitative experiments, we compared by ourself different articles, and also challenged our models on a completely different type of text : an extract of a scientific research article.

3 Results[3]

3.1 Quantitative results

Our quantitative results are summed up in tables 2 and 3. As expected, the pre-trained BARTHez, which is more complex, supervised and better trained, achieve better ROUGE scores than the extractive model. However, the performances are similar with our experiments on the main topic designed by hand, which was expected as OrangeSum is mainly about politic and society subjects.

3.2 Qualitative results

Some examples of generated summaries are presented in figures 5, 6 and 7. TextRank extractive algorithm has the inconvenient to be dependent on the ”quality” of the article, as when there is a lot of citations we end up with a ”patchwork” non concise summary, with pronouns out of context, ...

Moreover, BARTHez performs extraordinary well, even on totally different articles, surprisingly well understanding the scientific article (fig 7). In the first example (fig 5), it is also able to put a citation in the summary, correctly assign to its protagonist and included in the sentence, which shows a high level of understanding of the context and language modelling. Its summaries are also concise and factual, but maybe too small and suppressing too much context and protagonists (figure 6).

4 Discussion/Conclusion

To tackle the summarization task, we implemented and compared various approaches, from unsupervised extractive models to supervised abstractive models using deep learning methods.

Unsupervised extractive methods obtained already satisfying results, conditionned to the ”quality” of the articles’ sentences. Using pre-trained embeddings avoid having to learn the language, and the generated summaries have well-formed sentences, but may look like incoherent patchworks... On the contrary, supervised abstractive methods leads to better generated summaries, above all using the Attention Mechanism from Transformers. However, training such a model from scratch is really hard, and requires a wide corpus to model a language.

Last but not least, we can see some specific words such as ”coronavirus” or ”Macron” appear among the most frequent words of OrangeSum (fig 4) The language modelling used in our models being fixed, our models may become obsolet. To avoid this usual source of hardness in NLP and deal with new words and concepts, we may use pre-existing wide vocabularies or pre-train our models on wide datasets, but we should even think of a way to update them without fully re-training.

[2] stands for ”Recall-Oriented Understudy for Gisting Evaluation”

[3] We didn’t detail the performance of our abstractive model made from scratch, as such a model requires too much training time to correctly generate human-like sentences.

References

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. page 10.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text.

Moussa Kamal Eddine, Antoine J.-P. Tixier, and Michalis Vazirgiannis. 2020. Barthez: a skilled pretrained french sequence-to-sequence model. *CoRR*, abs/2010.12321.

Appendix

A Descriptive Statistics on OrangeSum training dataset

Note that all the following statistics were done on the training dataset of OrangeSum dataset, using the simpler tokenizers (`word_tokenize` and `sent_tokenize`), and removing french stopwords.

Dataset	Number of article/summary pairs
Training set	21401
Test set	1500
Validation set	1500

Table 1: Structure of OrangeSum dataset : with number of pairs article-summary per dataset)

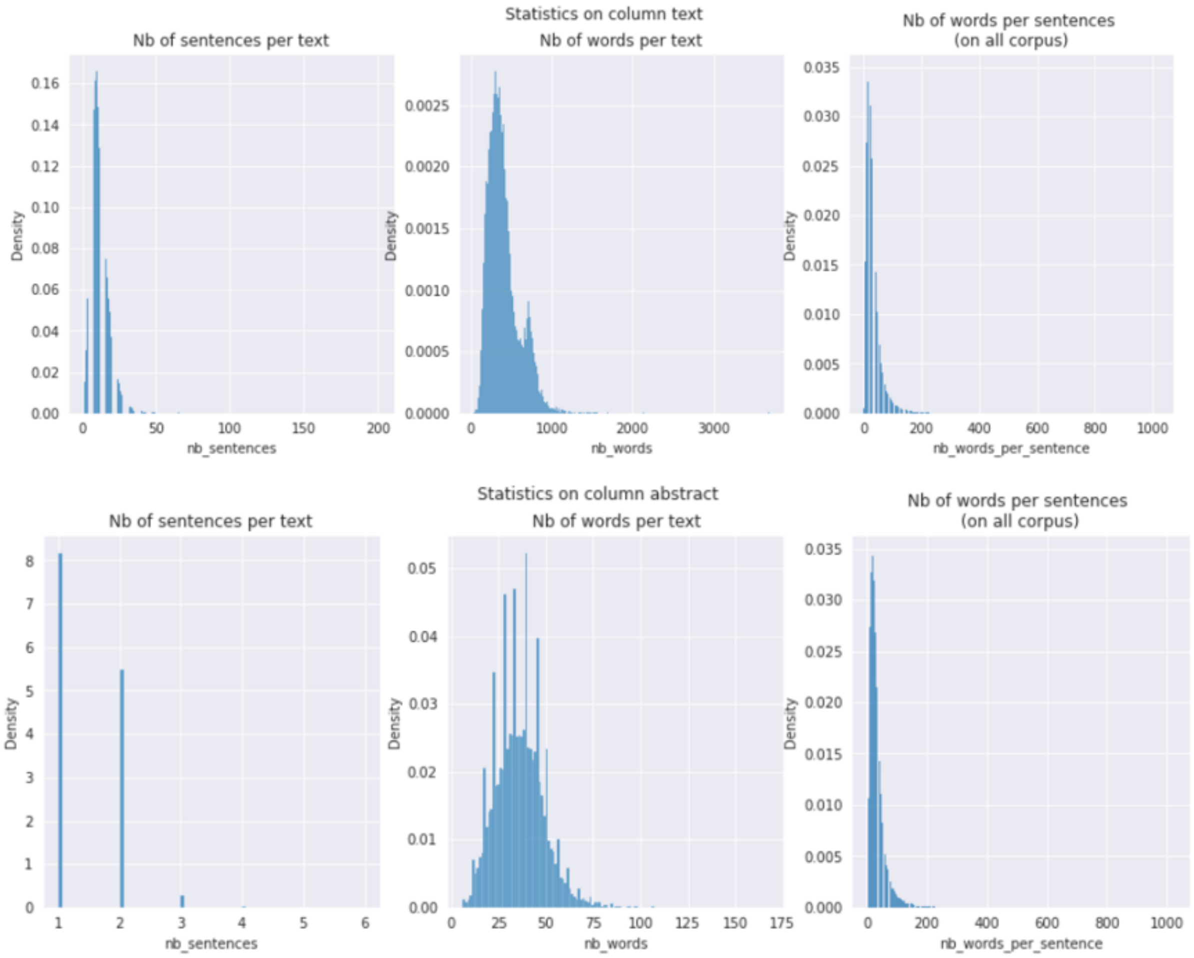


Figure 1: Statistics on text's structure (number of sentences, words and words per sentences)

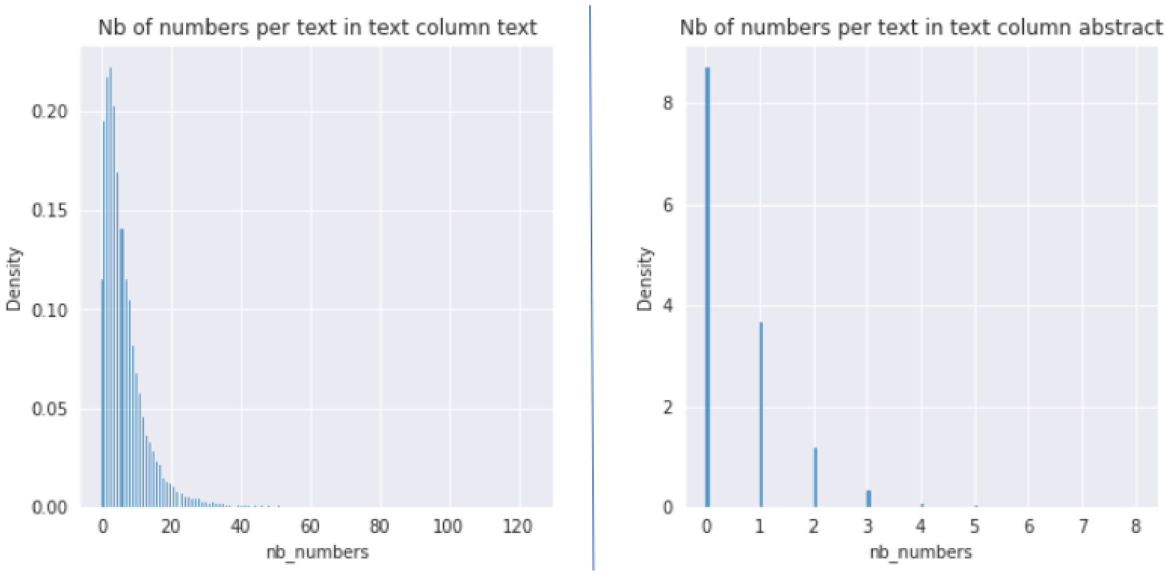


Figure 2: Statistics on the presence of numbers within texts

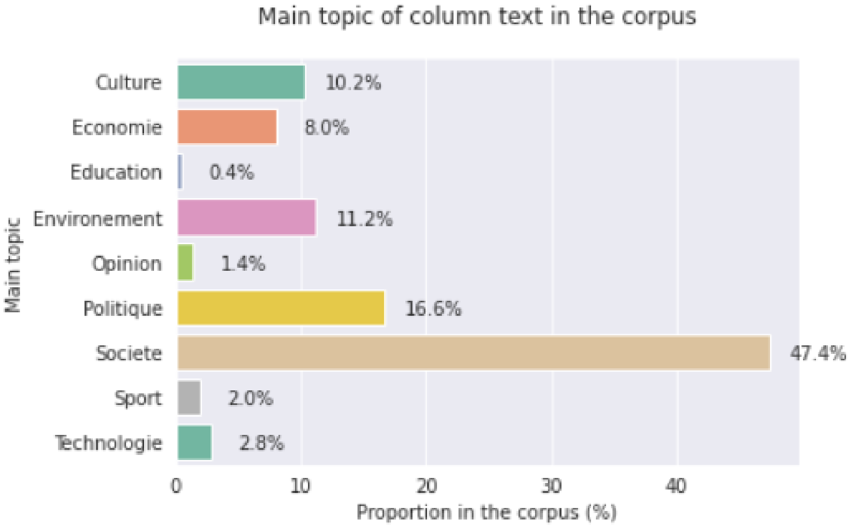
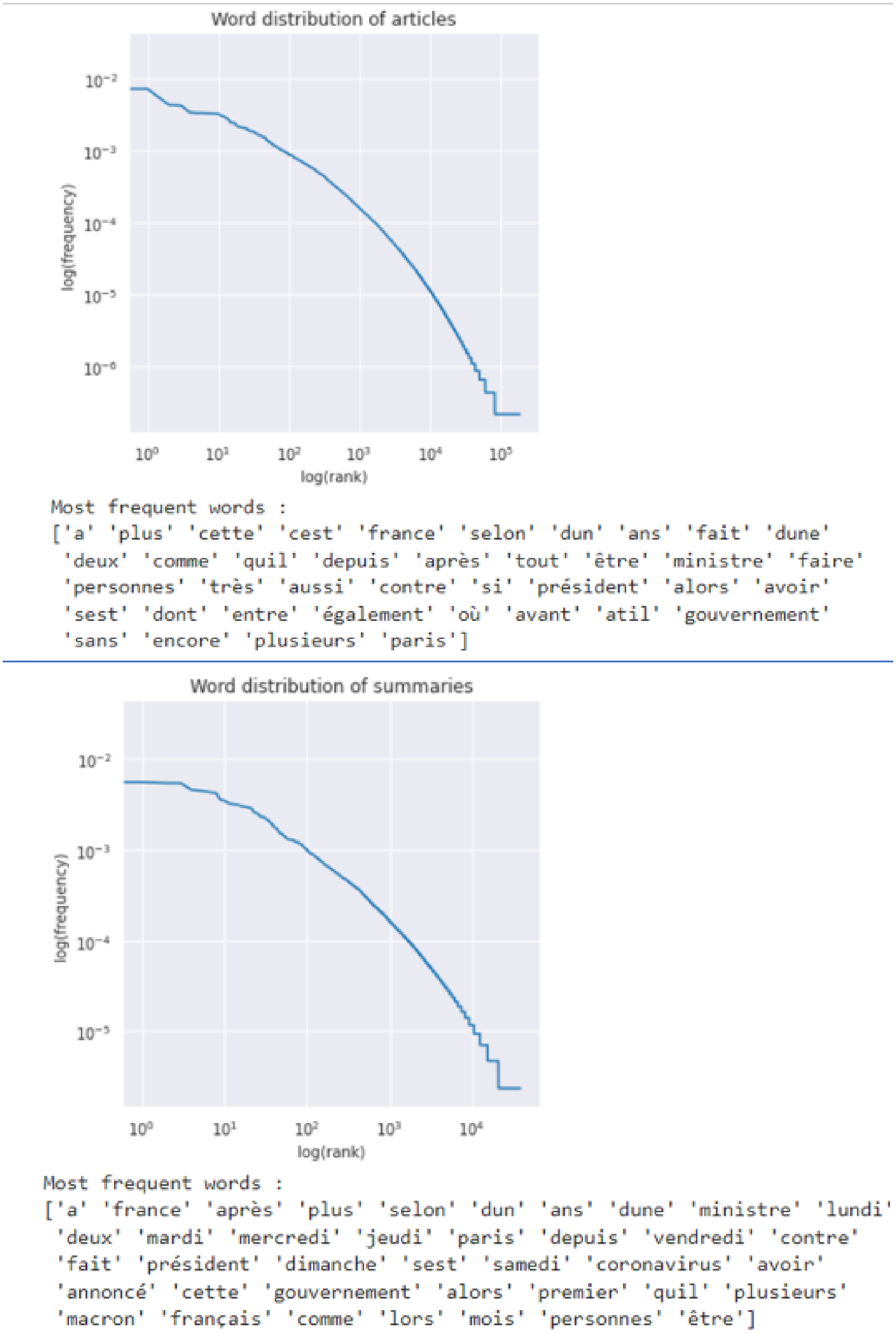


Figure 3: Main topic obtained using pre-trained `flaubert-mlsum-topic-classification` model, only on the 500 first articles of the dataset



```
Most frequent words :
['a' 'plus' 'cette' 'cest' 'france' 'selon' 'dun' 'ans' 'fait' 'dune'
 'deux' 'comme' 'quil' 'depuis' 'après' 'tout' 'être' 'ministre' 'faire'
 'personnes' 'très' 'aussi' 'contre' 'si' 'président' 'alors' 'avoir'
 'sest' 'dont' 'entre' 'également' 'où' 'avant' 'atil' 'gouvernement'
 'sans' 'encore' 'plusieurs' 'paris']
```

```
Most frequent words :
['a' 'france' 'après' 'plus' 'selon' 'dun' 'ans' 'dune' 'ministre' 'lundi'
 'deux' 'mardi' 'mercredi' 'jeudi' 'paris' 'depuis' 'vendredi' 'contre'
 'fait' 'président' 'dimanche' 'sest' 'samedi' 'coronavirus' 'avoir'
 'annoncé' 'cette' 'gouvernement' 'alors' 'premier' 'quil' 'plusieurs'
 'macron' 'français' 'comme' 'lors' 'mois' 'personnes' 'être']
```

Figure 4: Word distributions and most frequent words

B Results

The main part of our results was obtained on a subpart of the validation set of OrangeSum dataset (on the 200 first articles).

B.1 Quantitative results

Model	ROUGE-1	ROUGE-2	ROUGE-3
TextRank	0.1852	0.0414	0.1527
BARTHez	0.2487	0.0856	0.2032

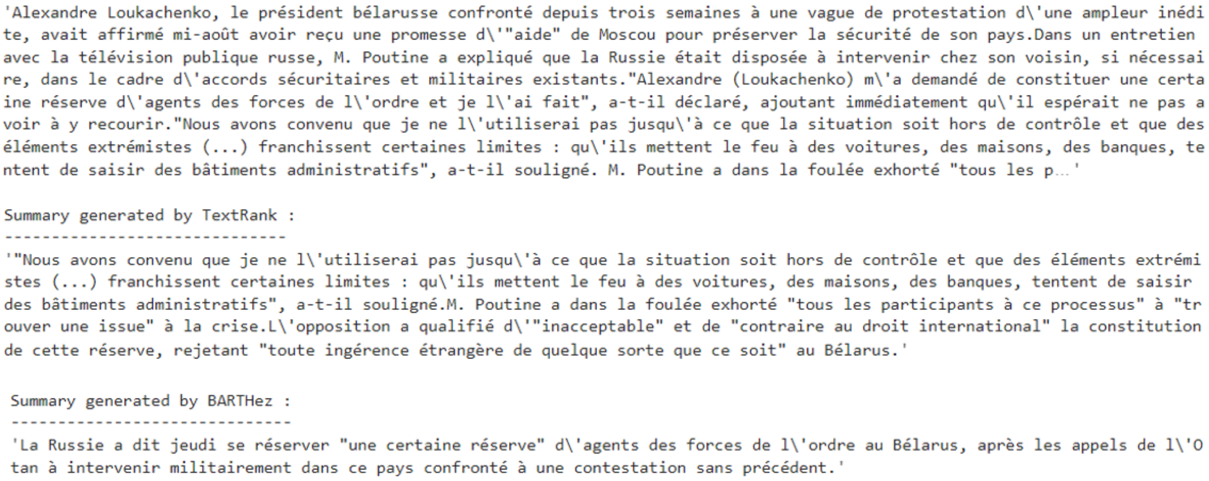
Table 2: ROUGE f1 scores obtained with the different models (on the 200 first articles of the validation set)

Model	Proportion of sumamries-article with same main topic
TextRank	69,50 %
BARTHez	66,50 %

Table 3: Proportion of generated summaries that were assigned the same main topic as the source article by `flaubert-mlsum-topic-classification` (on the 200 first articles of the validation set)

B.2 Qualitative results

Here, we present 3 examples of summaries generated on different articles, the first one from the validation set of OrangeSum, the last one on a completely different text.

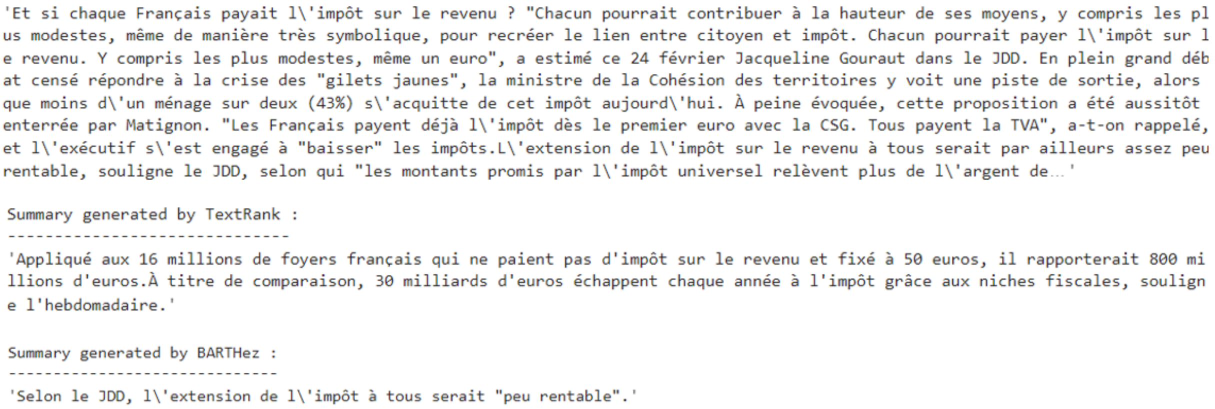


```
'Alexandre Loukachenko, le président bélarusse confronté depuis trois semaines à une vague de protestation d\'une ampleur inédi
te, avait affirmé mi-août avoir reçu une promesse d\'"aide" de Moscou pour préserver la sécurité de son pays.Dans un entretien
avec la télévision publique russe, M. Poutine a expliqué que la Russie était disposée à intervenir chez son voisin, si nécessai
re, dans le cadre d\'accords sécuritaires et militaires existants."Alexandre (Loukachenko) m\'a demandé de constituer une certa
ine réserve d\'agents des forces de l\'ordre et je l\'ai fait", a-t-il déclaré, ajoutant immédiatement qu\'il espérait ne pas a
voir à y recourir."Nous avons convenu que je ne l\'utiliserai pas jusqu\'à ce que la situation soit hors de contrôle et que des
éléments extrémistes (...) franchissent certaines limites : qu\'ils mettent le feu à des voitures, des maisons, des banques, te
ntent de saisir des bâtiments administratifs", a-t-il souligné. M. Poutine a dans la foulée exhorté "tous les p…'

Summary generated by TextRank :
------------------------------
'"Nous avons convenu que je ne l\'utiliserai pas jusqu\'à ce que la situation soit hors de contrôle et que des éléments extrémi
stes (...) franchissent certaines limites : qu\'ils mettent le feu à des voitures, des maisons, des banques, tentent de saisir
des bâtiments administratifs", a-t-il souligné.M. Poutine a dans la foulée exhorté "tous les participants à ce processus" à "tr
ouver une issue" à la crise.L\'opposition a qualifié d\'"inacceptable" et de "contraire au droit international" la constitution
de cette réserve, rejetant "toute ingérence étrangère de quelque sorte que ce soit" au Bélarus.'

 Summary generated by BARTHez :
 ------------------------------
 'La Russie a dit jeudi se réserver "une certaine réserve" d\'agents des forces de l\'ordre au Bélarus, après les appels de l\'O
 tan à intervenir militairement dans ce pays confronté à une contestation sans précédent.'
```

Figure 5: Example 1 - generated summaries for an article from OrangeSum validation set

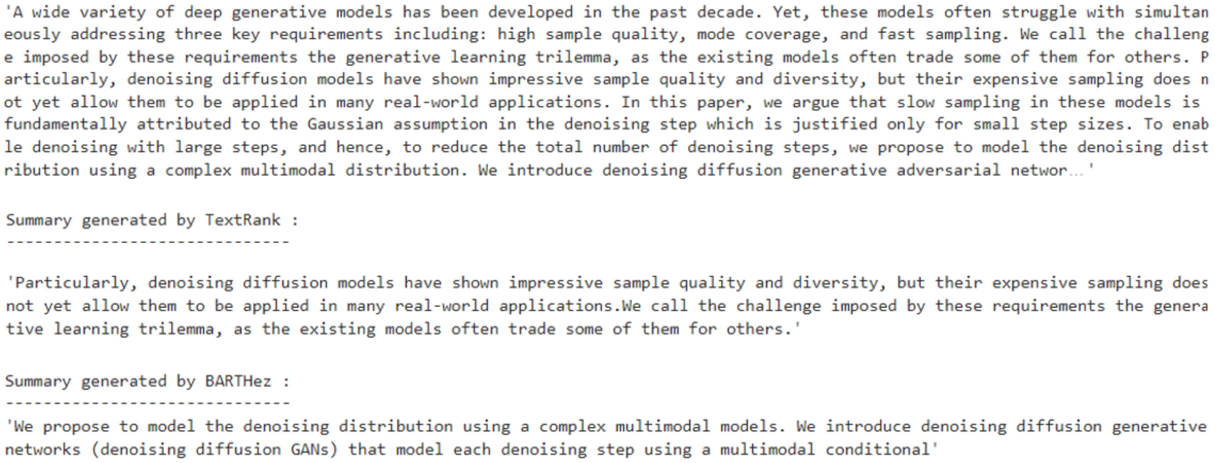


```
'Et si chaque Français payait l\'impôt sur le revenu ? "Chacun pourrait contribuer à la hauteur de ses moyens, y compris les pl
us modestes, même de manière très symbolique, pour recréer le lien entre citoyen et impôt. Chacun pourrait payer l\'impôt sur l
e revenu. Y compris les plus modestes, même un euro", a estimé ce 24 février Jacqueline Gouraut dans le JDD. En plein grand déb
at censé répondre à la crise des "gilets jaunes", la ministre de la Cohésion des territoires y voit une piste de sortie, alors
que moins d\'un ménage sur deux (43%) s\'acquitte de cet impôt aujourd\'hui. À peine évoquée, cette proposition a été aussitôt
enterrée par Matignon. "Les Français payent déjà l\'impôt dès le premier euro avec la CSG. Tous payent la TVA", a-t-on rappelé,
et l\'exécutif s\'est engagé à "baisser" les impôts.L\'extension de l\'impôt sur le revenu à tous serait par ailleurs assez peu
rentable, souligne le JDD, selon qui "les montants promis par l\'impôt universel relèvent plus de l\'argent de...'

Summary generated by TextRank :
------------------------------
'Appliqué aux 16 millions de foyers français qui ne paient pas d'impôt sur le revenu et fixé à 50 euros, il rapporterait 800 mi
llions d'euros.À titre de comparaison, 30 milliards d'euros échappent chaque année à l'impôt grâce aux niches fiscales, soulign
e l'hebdomadaire.'

Summary generated by BARTHez :
-----------------------------
'Selon le JDD, l\'extension de l\'impôt à tous serait "peu rentable".'
```

Figure 6: Example 2 - generated summaries for an article from OrangeSum validation set



```
'A wide variety of deep generative models has been developed in the past decade. Yet, these models often struggle with simultan
eously addressing three key requirements including: high sample quality, mode coverage, and fast sampling. We call the challeng
e imposed by these requirements the generative learning trilemma, as the existing models often trade some of them for others. P
articularly, denoising diffusion models have shown impressive sample quality and diversity, but their expensive sampling does n
ot yet allow them to be applied in many real-world applications. In this paper, we argue that slow sampling in these models is
fundamentally attributed to the Gaussian assumption in the denoising step which is justified only for small step sizes. To enab
le denoising with large steps, and hence, to reduce the total number of denoising steps, we propose to model the denoising dist
ribution using a complex multimodal distribution. We introduce denoising diffusion generative adversarial networ...'

Summary generated by TextRank :
------------------------------

'Particularly, denoising diffusion models have shown impressive sample quality and diversity, but their expensive sampling does
not yet allow them to be applied in many real-world applications.We call the challenge imposed by these requirements the genera
tive learning trilemma, as the existing models often trade some of them for others.'

Summary generated by BARTHez :
------------------------------
'We propose to model the denoising distribution using a complex multimodal models. We introduce denoising diffusion generative
networks (denoising diffusion GANs) that model each denoising step using a multimodal conditional'
```

Figure 7: Example 3 - generated summaries for a long abstract of a research paper