مدرسة علوم المعلومات
ⵜⵉⵎⵔⴰⵜ ⵏ ⵜⵓⵙⵙⵏⵉⵡⵉⵏ ⵏ ⵜⵣⵔⴼⵜ
ECOLE DES SCIENCES
DE L'INFORMATION

Plateforme de Résumé Automatique d'Articles d'Al Jazeera

Réalisé par :
Ibtissam LABYADY
Sokhna Mai WANE
Mohamed CISSE

Encadré par:
Najima DAOUDI
Ghizlane BOURAHOUAT

Le plan

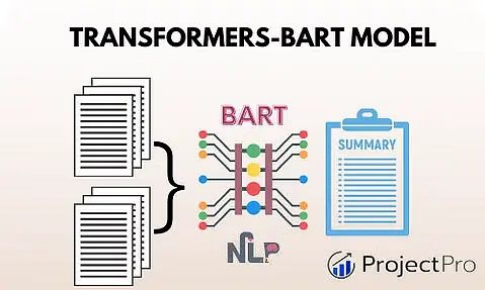
1. Introduction
2. Conception du projet
3. Visite notre plateforme interactive
4. Web scraping
5. Le modèle BART
6. Le modèle PEUGASUS
7. Le Modèle T5
8. Le Fine Tuning du modèle T5
9. Le modèle le plus performant
10. Conclusion

Introduction

Définition :

La summarization de texte (résumé automatique de texte) est le processus de distillation des informations les plus importantes d'un texte afin de produire une version abrégée pour une tâche spécifique et un utilisateur donné.

Un résumé automatique de texte est une version condensée d'un document textuel, obtenu au moyen de techniques informatiques. La forme la plus connue et la plus visible des condensés de textes est le résumé, représentation abrégée et exacte du contenu d'un document.

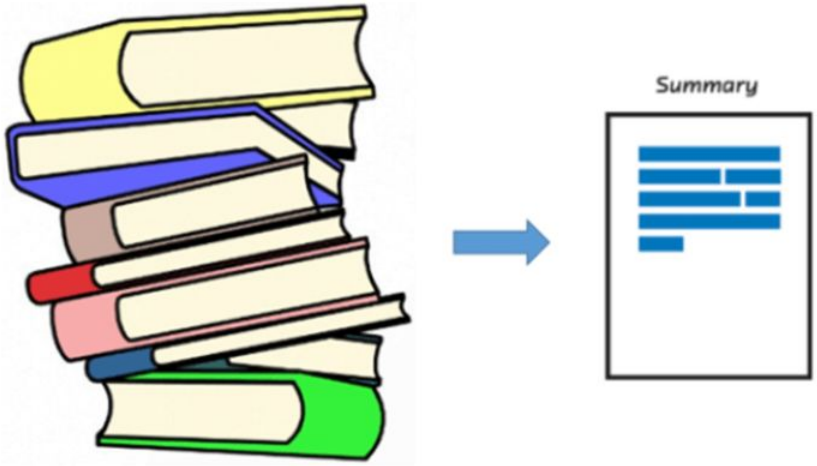


Text Summurization

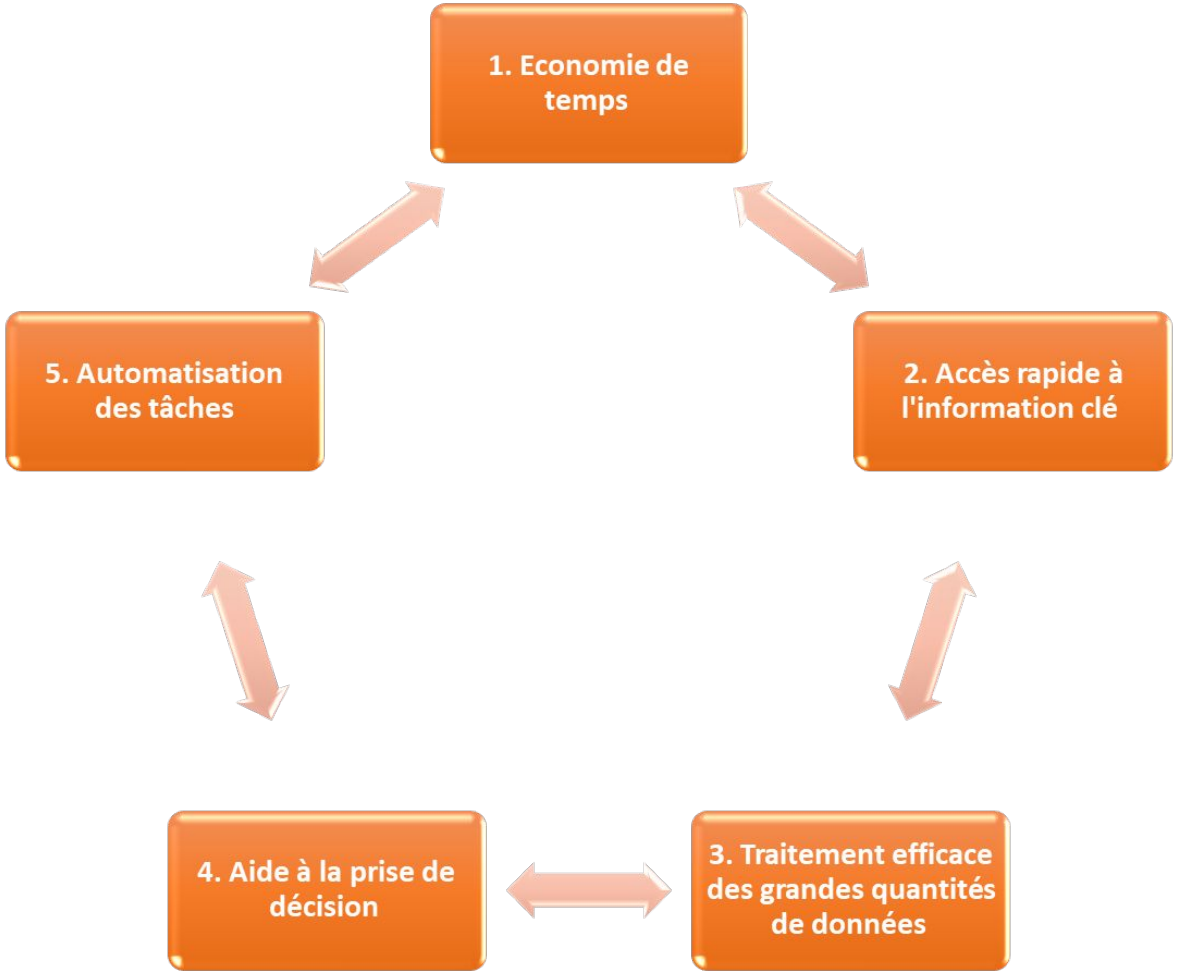
La résumé crée une version plus courte d'un document ou d'un article qui capture toutes les informations importantes.

La résumé peut être :

- Extrative : extraire les informations les plus pertinentes d'un document.
- Abstractive : générer un nouveau texte qui capture les informations les plus pertinentes.



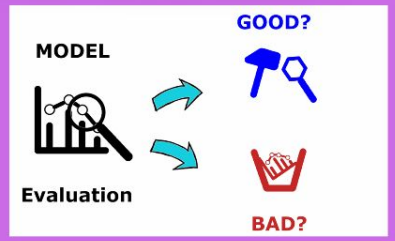
Introduction : Avantages de text summarisation



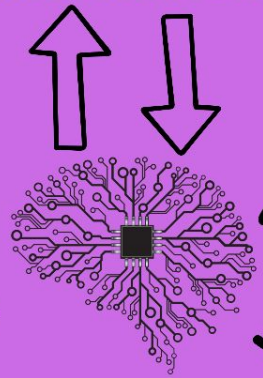
Conception de la Plateforme de Résumé Automatique d'Articles d'Al Jazeera

WEB SCRAPING

PREPROCESSING DATA



EVALUATE MODEL



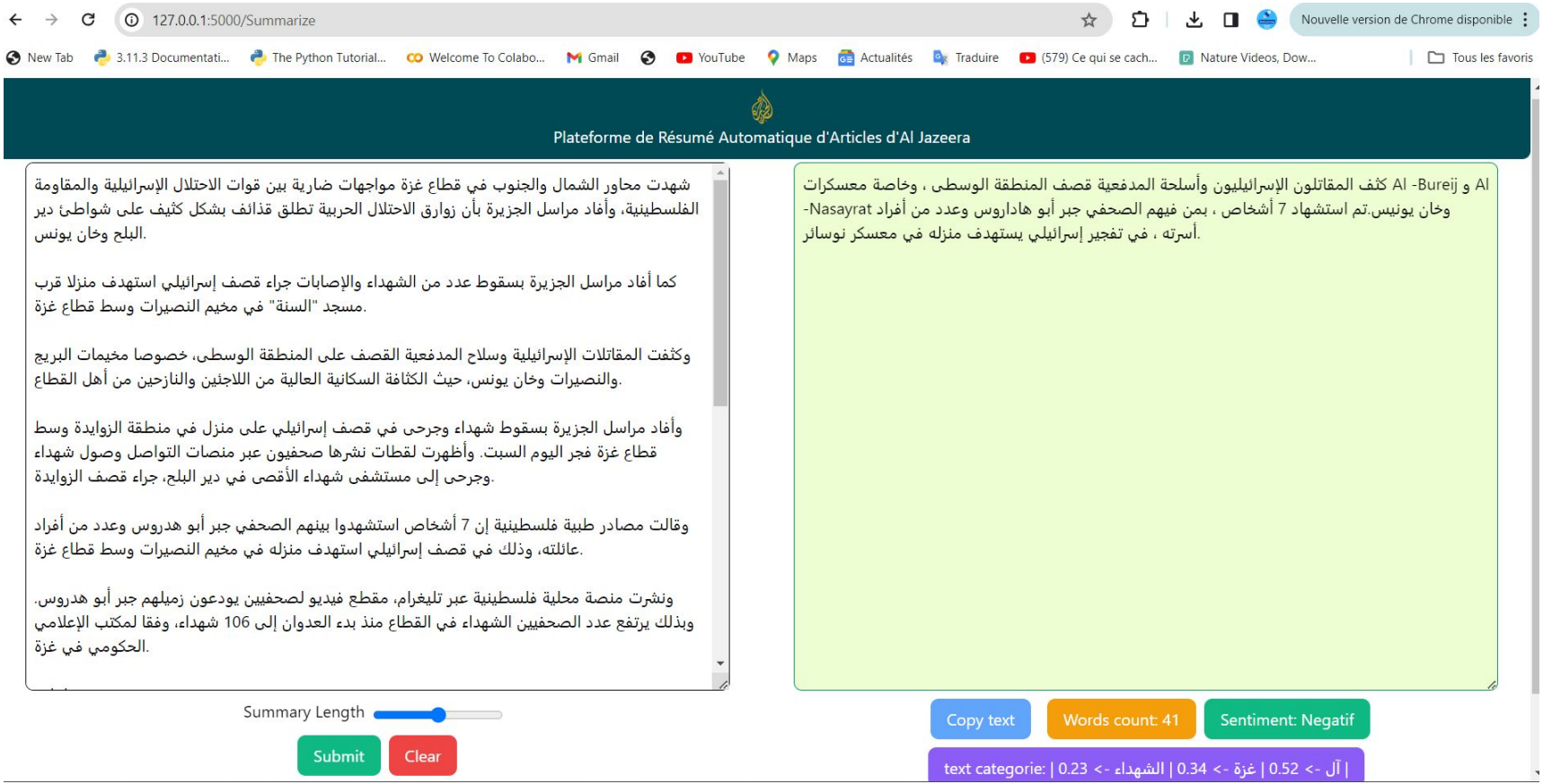
CREATE MODEL

Final User

Interactif website

Created by mohamed CISSE

Implementation : visite de notre plateforme



Web Scraping Aljazeera (https://www.aljazeera.com/)

Dans ce processus de collecte de données, l'objectif était d'extraire des informations du site web Al Jazeera. On extrait **"url" (URL de l'article), "titre" (titre de l'article), "catégorie" (catégorie), "texte" (texte de l'article).** Le DataFrame compte 100 entrées et fournit une collection complète de données pour une analyse ou un traitement ultérieur.

[18]:
```
# Charger le DataFrame contenant les URLs
df = pd.read_csv("/kaggle/working/Aljazeera_dataset.csv")
df.head(50)
```

[18]:

	id	url	title	category	summary	texte
0	1	https://www.aljazeera.com//news/liveblog/2023/...	'Unlawful': UN aid chief decries Israeli a...	news	UNRWA says Israeli forces fired on aid con...	UNRWA says Israeli forces fired on aid convoy ...
1	2	https://www.aljazeera.com//gallery/2023/12/29/...	Photos: Palestinians perform Friday praye...	news	Israel restricts Palestinians' access to ...	In Pictures Israeli authorities barred Palesti...
2	3	https://www.aljazeera.com//news/2023/12/29/arg...	Argentina announces that it will not join B...	news	The move is the latest shift in economic an...	The move is the latest shift in economic and f...
3	4	https://www.aljazeera.com//program/newsfeed/20...	Russia launches "most massive aerial att...	news	Russia launches one of its largest nights of...	An overnight barrage of Russian missiles and d...
4	5	https://www.aljazeera.com//news/2023/12/29/sou...	South Africa files case at ICJ accusing Isr...	news	Israel, which has been accused of meting ou...	Israel, which has been accused of meting out c...

Web Scrapping Aljazeera (https://www.aljazeera.net/)

Dans ce processus de collecte de données, l'objectif était d'extraire des informations du site web Al Jazeera. On extrait **"url" (URL de l'article), "titre" (titre de l'article), "categorie" (catégorie), "texte" (texte de l'article).** Le DataFrame compte 350 entrées et fournit une collection complète de données pour une analyse ou un traitement ultérieur.

```
# Charger le DataFrame contenant les URLs
df = pd.read_csv("/kaggle/working/Aljazeera_dataset.csv")
df.head(50)
```

	id	url	title	category	summary	texte
0	1	https://www.aljazeera.net/news/2023/12/29/%d8%...	حزب الله يقصف تجهيزات ومعدّات تجسّس إسرائيلية ...	news	قال حزب الله اللبناني في بيانات منفصلة، إن حنا...	أعلن حزب الله اللبناني، الجمعة، استهدافه تجهيز...
1	2	https://www.aljazeera.net/news/2023/12/29/%d8%...	إصابة 4 مستوطنين بعملية دهس قرب الخليل	news	أفادت القناة 12 الإسرائيلية بإصابة 4 مستوطنين ...	أفادت القناة 12 الإسرائيلية بإصابة 4 مستوطنين ...
2	3	https://www.aljazeera.net/news/2023/12/29/%d8%...	بعد اختراق الاحتلال كاميرات المراقبة.. حزب الل...	news	دعا حزب الله سكان البلدات الحدودية جنوب لبنان ...	دعا حزب الله سكان البلدات الحدودية جنوب لبنان ...
3	4	https://www.aljazeera.net/news/2023/12/29/%d9%...	ولاية "مين": ترامب غير مؤهل لمنصب الرئيس	news	في أزمة جديدة، قضت ولاية "مين" الأميركية بعدم ...	قضت ولاية مين الأميركية -أمس الخميس- بعدم أهلي...
4	5	https://www.aljazeera.net/news/2023/12/29/%d8%...	الاحتلال يطلق النار على قافلة مساعدات في غزة	news	قالت وكالة الأونروا اليوم إن جنودا إسرائيليين ...	قالت وكالة غوث وتشغيل اللاجئين الفلسطينيين (أو...
5	6	https://www.aljazeera.net/news/presstour/2023/	كاتب إسرائيلي: لا أحد يستطيع	news	قال الكاتب الإسرائيلي عدعون ليفي	قال الكاتب الإسرائيلي عدعون ليفي

Modèle BART :
Bidirectional + Auto-Regressif

Définition :

Le modèle BART est un modèle séquence-à-séquence entraîné en tant qu'autoencodeur de débruitage, capable de réaliser des tâches telles que la traduction automatique, la réponse aux questions, la summarization de texte, la classification de séquences, et d'autres applications spécifiques après un finetuning sur des jeux de données adaptés.

Modèle BART : BartForConditionalGeneration

La nature bidirectionnelle et auto-encodeur de **BERT** est...

+ idéal pour les tâches en aval (par exemple, classification) qui nécessitent des informations sur l'ensemble

séquence

- pas si bon pour les tâches de génération où le mot généré ne devrait dépendre que de

mots générés précédemment

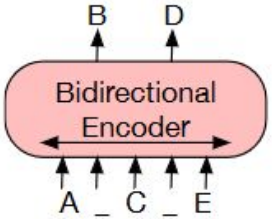
L'approche unidirectionnelle et autorégressive de **GPT** est...

+ bon pour la génération de texte

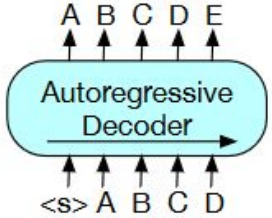
- pas si bon pour les tâches qui nécessitent des informations sur toute la séquence, par exemple la classification

BART est le meilleur des deux mondes

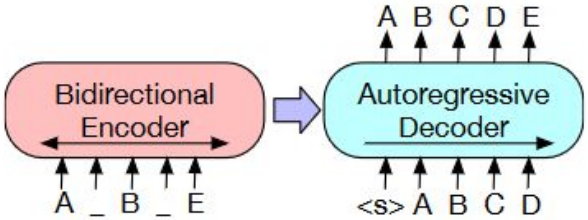
Modèle BART : BartForConditionalGeneration



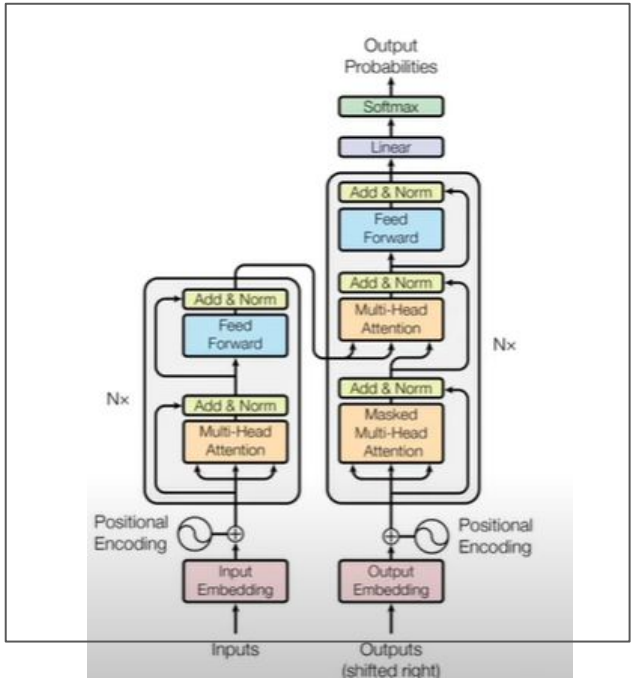
(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

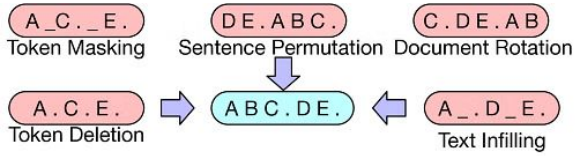


(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.





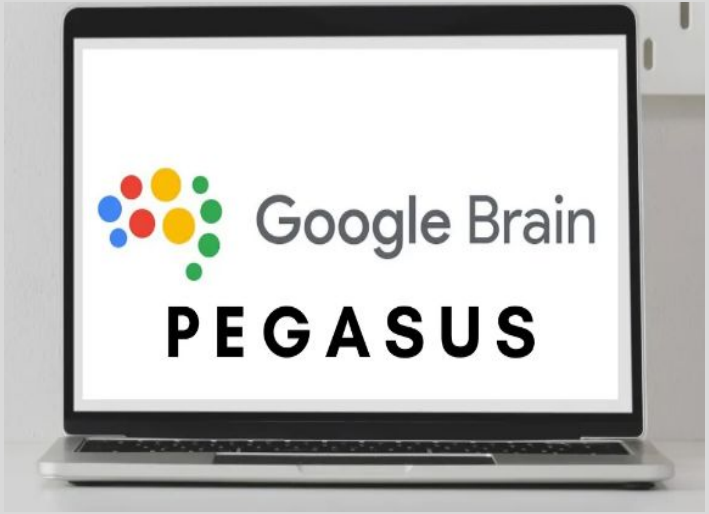
Modèle BART : Evaluation

```python
# Calculer la similarité entre les résumés réels et prédits (par exemple, ROUGE score)
from rouge import Rouge

rouge = Rouge()
scores = rouge.get_scores(df['predicted_summary'], df['summary'], avg=True)

print("Scores ROUGE :")
print(scores)
```

```
Scores ROUGE :
{'rouge-1': {'r': 0.45664227469297125, 'p': 0.16943289191500277, 'f': 0.24526417777333276}
```

Les mesures **ROUGE-1, ROUGE-2 et ROUGE-L** indiquent que le résumé généré par le modèle BART a une performance modérée, avec des scores de rappel, de précision et de f1-score variables pour les unigrammes, les bigrammes et les unités lexicales par rapport au résumé de référence.

PEGASUS：Pre-training with Extracted Gap-sentences for Abstractive Summarization



PEGASUS : Pre-training with Extracted Gap-sentences for Abstractive Summarization

PEGASUS est une approche novatrice pour les modèles d'encodeur-décodeur basés sur Transformer, spécifiquement conçue pour la génération de résumés extractifs.

Dans le contexte de PEGASUS , le masquage des parties moins essentielles du texte se fait en utilisant des techniques de traitement de langage naturel.

Ce processus permet au modèle de se focaliser sur les éléments clés du texte lors de la génération du résumé, en lui donnant la capacité de distinguer et de prioriser les informations essentielles par rapport aux détails moins importants.

Application

L'extraction des articles du site Al Jazeera a été réalisée en utilisant BeautifulSoup pour le scraping des données à partir de la page web. Le processus comprenait la récupération des catégories, des liens vers les articles, ainsi que le titre, le contenu et le résumé de chaque article.

```
[ ]     writer.writerows(data)

print("Extraction terminée. Les articles ont été enregistrés dans 'aljazeera_articles_scrap.csv'.")
```

```
Extraction terminée. Les articles ont été enregistrés dans 'aljazeera_articles_scrap.csv'.
```

```
[ ] # Charger les données depuis le fichier CSV
df = pd.read_csv("/content/aljazeera_articles_scrap.csv")

# Afficher les cinq premières lignes du dataframe
df.head()
```

	Title	Content	Category	Link	summary
0	Israel restricts Palestini-ans' access to ...	Israel restricts Palestinians' access to ...	News	https://www.aljazeera.com/news/	Israel restricts Palestinians' access to ...
1	Israel restricts Palestini-ans' access to ...	Israel restricts Palestinians' access to ...	Middle East	https://www.aljazeera.com/middle-east/	Israel restricts Palestinians' access to ...
2	Overthrow of leaders in Niger and Gabon has ...	Overthrow of leaders in Niger and Gabon has ...	Africa	https://www.aljazeera.com/africa/	Overthrow of leaders in Niger and Gabon has ...

Ensuite, Pegasus a été utilisé pour générer des résumés automatiques pour chaque article.

```
# Charger le modèle Pegasus et le tokenizer associé
model_Pegasus = 'google/pegasus-xsum'
tokenizer = PegasusTokenizer.from_pretrained(model_Pegasus)
model = PegasusForConditionalGeneration.from_pretrained(model_Pegasus)
```

Après avoir obtenu ces résumés automatiques à l'aide de Pegasus, la métrique ROUGE a été employée pour évaluer la qualité et l'efficacité de ces résumés générés

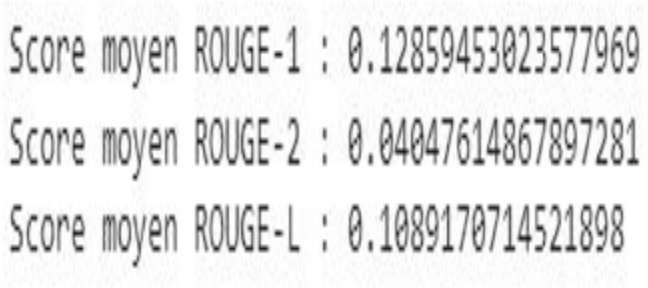
```
[ ]
# Appliquer la fonction generate_summary à chaque ligne de la colonne "Content"
df['pred_Summary'] = df['Content'].apply(generate_summary)
df
```

	Title	Content	Category	Link	summary	pred_Summary
0	Israel restricts Palestinians' access to ...	israel restricts palestinians' access ala...	News	https://www.aljazeera.com/news/	Israel restricts Palestinians' access to ...	israel intensifies ground offensive cent...
	Israel re-					

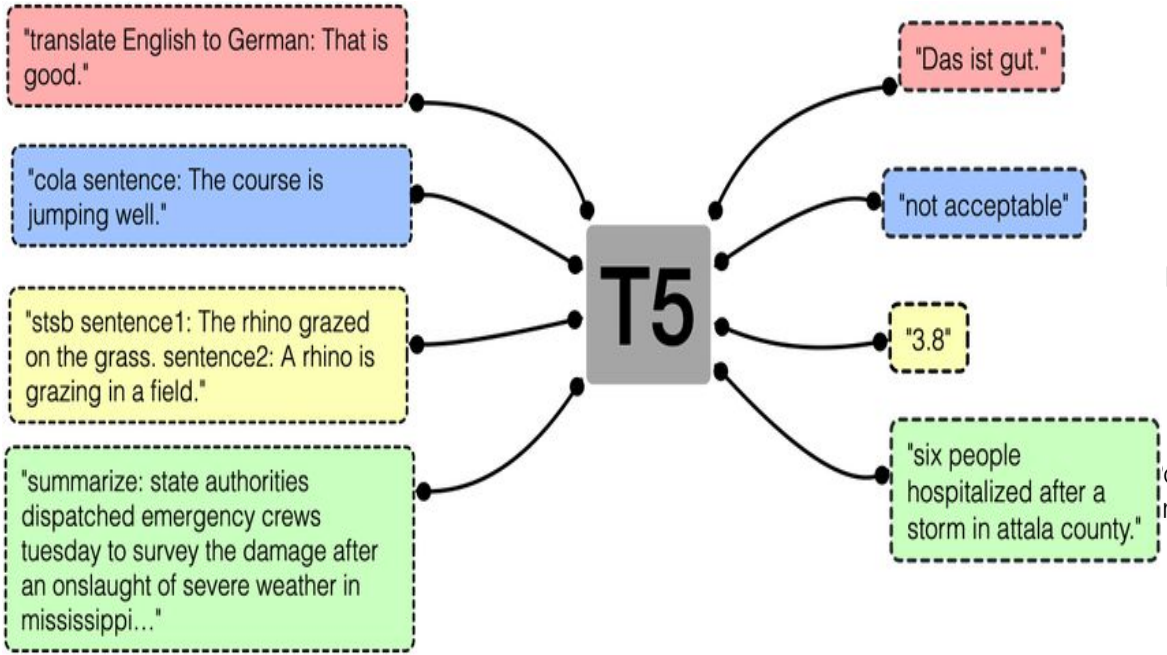
Les scores ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) ont été calculés pour mesurer la similarité entre les résumés générés automatiquement et les résumés de référence,

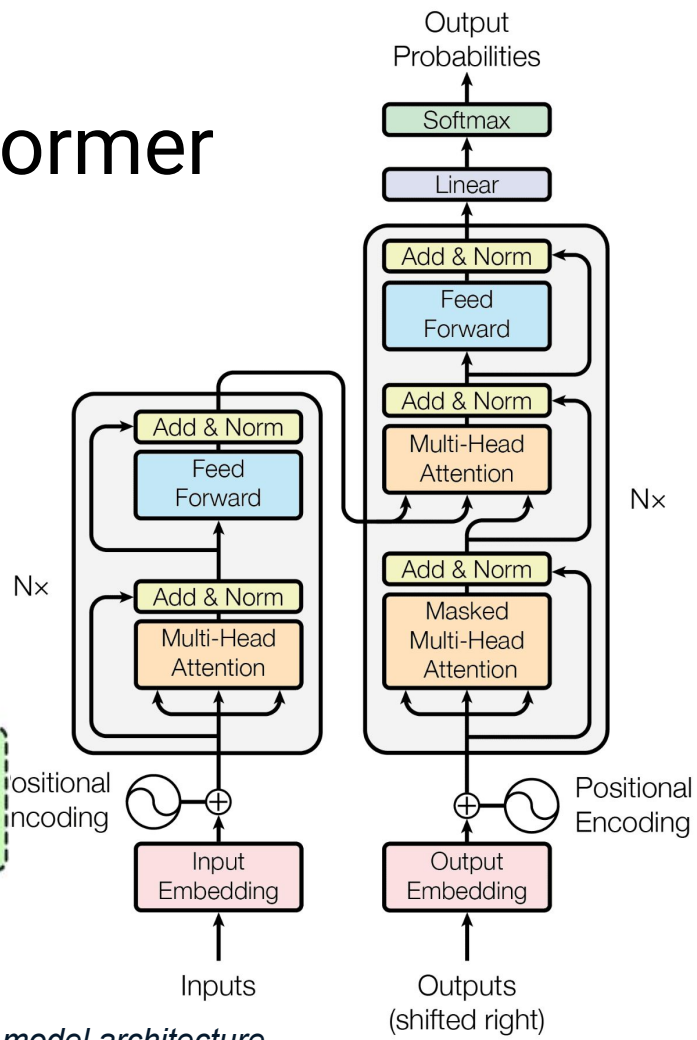
Les résultats de l'évaluation ROUGE ont fourni des scores moyens pour chaque métrique. Par exemple :

- ROUGE-1 : Environ 12.86% de similarité entre les mots individuels des résumés générés et ceux de référence.
- ROUGE-2 : Environ 4.05% de similarité entre les paires de mots consécutifs.
- ROUGE-L : Environ 10.89% de similarité en tenant compte de la plus longue séquence de mots commune.



T5: Text-To-Text Transfer Transformer





The Transformer model architecture

Partie code avec le modele t5_base api

```python
# Set the device to GPU if available, otherwise use CPU
import torch
from transformers import AutoModelWithLMHead, AutoTokenizer
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

# Load the tokenizer and model
tokenizer = AutoTokenizer.from_pretrained("mrm8488/t5-base-finetuned-summarize-news")
model = AutoModelWithLMHead.from_pretrained("mrm8488/t5-base-finetuned-summarize-news")
model = model.to(device)
max_length=150
# Define the function to generate summaries
def generate_summary(article):
    input_ids = tokenizer(article, return_tensors="pt", truncation=True, padding=True,add_special_tokens=True).input_ids.to(device)
    generated = model.generate(input_ids=input_ids, num_beams=2, max_length=max_length,  repetition_penalty=2.5, length_penalty=1.0, earl
    output = tokenizer.decode(generated, skip_special_tokens=True, clean_up_tokenization_spaces=True)
    return output

# Apply the generate_summary function to the "texte" column
df["resume_genere"] = df["texte"].apply(generate_summary)

# Save the DataFrame with ROUGE scores to a CSV file
df.to_csv("t5_resume_scores.csv", index=False)
```

Resume Genere

```
memory usage: 16.1+ KB
```

```
df.head()
```

[14]:

	id	url	title	category	summary	texte	resume_genere	rouge_scores
0	1	https://www.aljazeera.com//news/liveblog/2023/...	'Unlawful': UN aid chief decries Israeli a...	news	UNRWA says Israeli forces fired on aid con...	UNRWA says Israeli forces fired on aid convoy ...	Israeli forces fired on aid convoy travelling ...	[{'rouge-1': {'r': 0.5, 'p': 0.210526315789473...
1	2	https://www.aljazeera.com//gallery/2023/12/29/...	Photos: Palestinians perform Friday praye...	news	Israel restricts Palestini-ans' access to ...	In Pictures Israeli authorities barred Palesti...	Israeli forces barred Palestinians from enteri...	[{'rouge-1': {'r': 0.46153846153846156, 'p': 0...
2	3	https://www.aljazeera.com//news/2023/12/29/arg...	Argentina announces that it will not join B...	news	The move is the latest shift in economic an...	The move is the latest shift in economic and f...	Argentina has announced that it will not join ...	[{'rouge-1': {'r': 0.5555555555555556, 'p': 0....
3	4	https://www.aljazeera.com//program/newsfeed/20...	Russia launches "most massive aerial att...	news	Russia launches one of its largest nights of...	An overnight barrage of Russian missiles and d...	122 missiles and dozens of drones were involve...	[{'rouge-1': {'r': 0.29411764705882354, 'p': 0...
4	5	https://www.aljazeera.com//news/2023/12/29/sou...	South Africa files case at ICJ accusing Isr...	news	Israel, which has been accused of meting ou...	Israel, which has been accused of meting out c...	South Africa has filed a case against Israel a...	[{'rouge-1': {'r': 0.4444444444444444, 'p': 0....

Les résultats de l'évaluation ROUGE

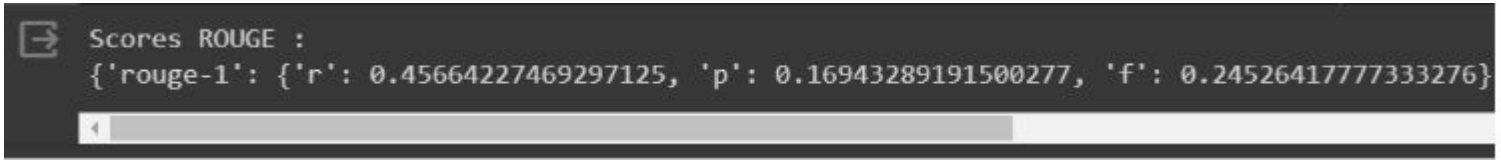
Les résultats de l'évaluation ROUGE ont fourni des scores moyens pour chaque métrique. Par exemple :

- ROUGE-1 : Environ 17.10% de similarité entre les mots individuels des résumés générés et ceux de référence.
- ROUGE-2 : Environ 5.55% de similarité entre les paires de mots consécutifs.
- ROUGE-L : Environ 15.9% de similarité en tenant compte de la plus longue séquence de mots commune.

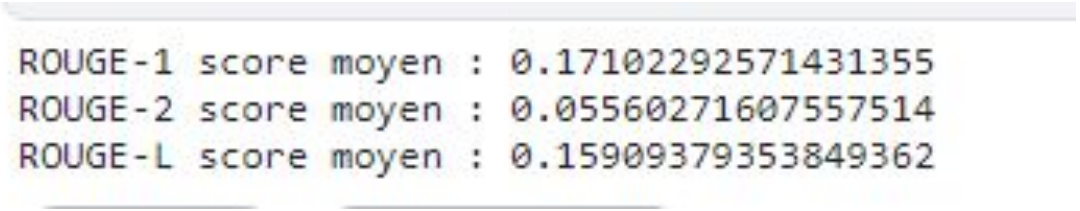
```
ROUGE-1 score moyen : 0.171022925714313555
ROUGE-2 score moyen : 0.055602716075575 14
ROUGE-L score moyen : 0.15909379353849362
```

Le modèle le plus performant (Evaluation)

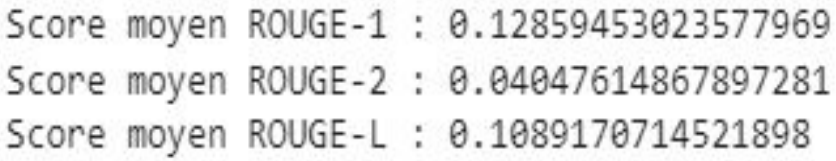
Bert



T5

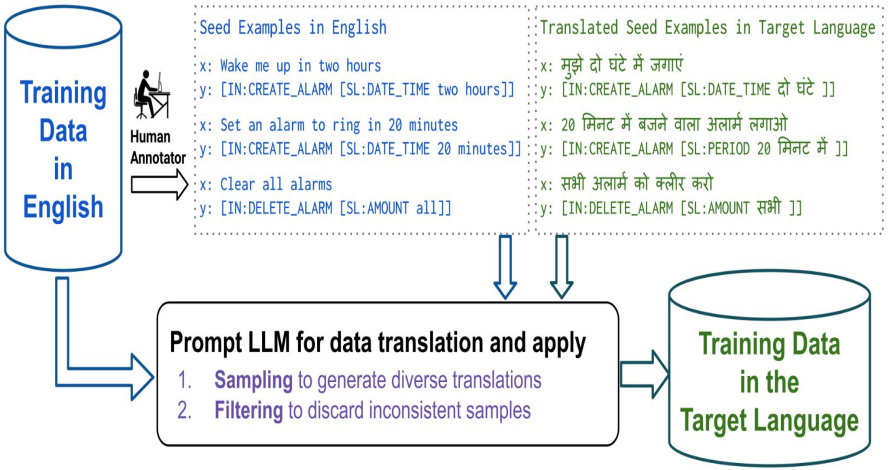


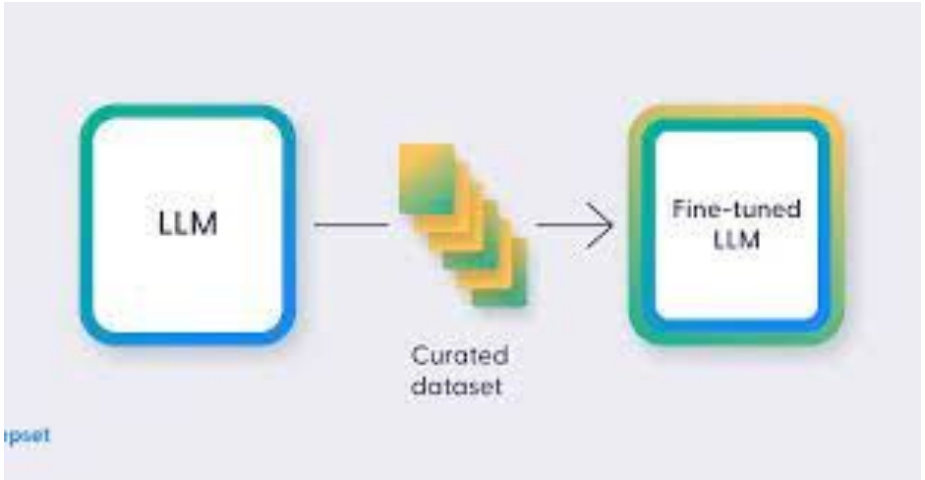
Pegasus



Question

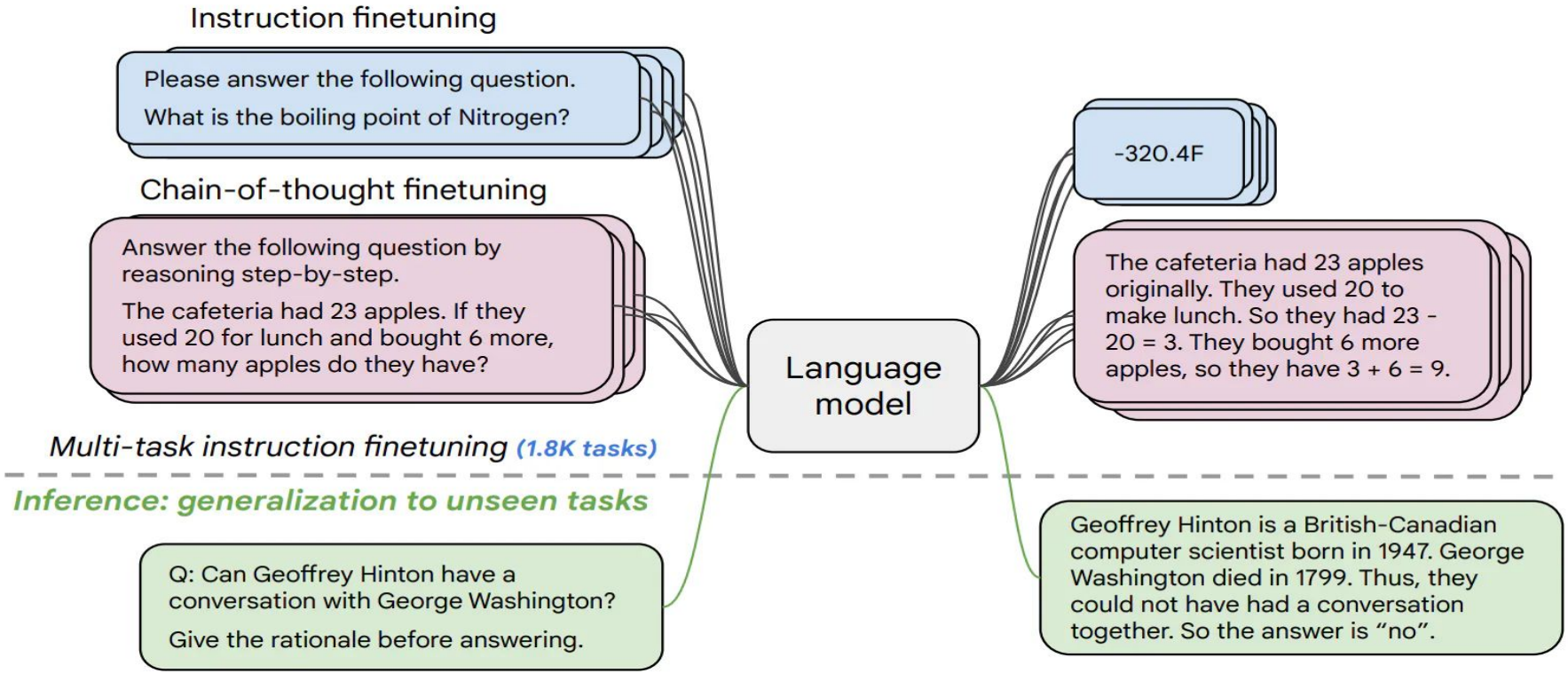
Quelle est la meilleure approche pour générer des résumés dans la langue cible : utiliser la traduction avec un modèle pré-entraîné ou effectuer un fine-tuning sur des données dans la langue cible ?





Fine Tuning

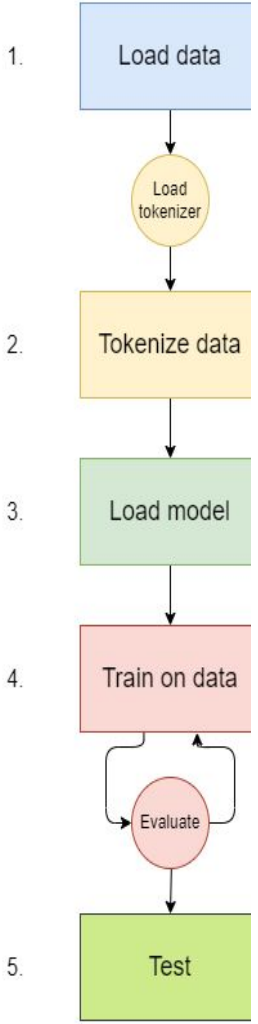
Le processus de fine-tuning permet d'adapter le modèle T5 aux spécificités de la tâche de summarization, améliorant ainsi sa capacité à produire des résumés pertinents et de qualité.



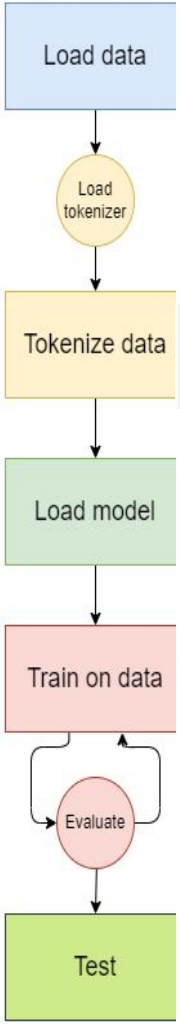
Model training

Quelques étapes importantes du fine-tuning

1. **Sélectionner un modèle pré-entraîné**
2. **Préparer les données d'entraînement en les formatant**
3. **dans un format approprié pour le modèle.**
4. **Charger le modèle pré-entraîné :**
5. **Ajouter de nouvelles couches**
6. **Entraîner le modèle**
7. **Valider le modèle**
8. **Testez le modèle**



Model training




```
MODEL_NAME = "t5-base"
tokenizer = T5Tokenizer.from_pretrained(MODEL_NAME, model_max_length=512)
```

```
other sentence."
```

```
{'input_ids': [[8774, 6, 48, 80, 7142, 55, 1], [100, 19, 430, 7142,
1]], 'attention_mask': [[1, 1, 1, 1, 1, 1, 1], [1, 1, 1, 1, 1, 1]]}
```

```
tokenized_datasets = raw_datasets.map(preprocess_function, batched=True)
```

```
model = AutoModelForSeq2SeqLM.from_pretrained(model_checkpoint)
```

In [27]:

```
from transformers import Seq2SeqTrainer

trainer = Seq2SeqTrainer(
    model=model,
    args=args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["validation"],
    data_collator=data_collator,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics
)
```

We can now finetune our model by just calling the `train` method:

In [28]:

```
trainer.train()
```

```
batch_size = 8
model_name = model_checkpoint.split("/")[-1]
args = Seq2SeqTrainingArguments(
    f"{model_name}-finetuned-xsum",
    evaluation_strategy = "epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    weight_decay=0.01,
    save_total_limit=3,
    num_train_epochs=20,
    predict_with_generate=True,
    fp16=True,
)
```

Evaluation fine tuning with t5-base

ibtissam369/t5-base-finetuned-summarize-news-finetuned-xsum

View run at https://wandb.ai/arab/huggingface/runs/dtz1lwzl

[208/208 03:29, Epoch 8/8]

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	No log	3.061483	19.324800	6.376300	17.250200	17.252600	19.000000
2	No log	2.546520	34.320000	18.953600	32.836400	33.066400	19.000000
3	No log	2.251078	34.606000	18.909800	33.911500	33.942800	19.000000
4	No log	2.054790	36.568000	20.059200	35.550400	35.584500	18.961500
5	No log	1.945011	36.634400	19.543100	35.603400	35.642600	18.961500
6	No log	1.881996	36.183500	19.342900	35.305300	35.401900	18.961500
7	No log	1.841060	36.183500	19.342900	35.305300	35.401900	18.961500
8	No log	1.828572	36.183500	19.342900	35.305300	35.401900	18.961500

Les métriques ROUGE (Rouge1, Rouge2, Rougel, Rougelsum) semblent également montrer une amélioration progressive, ce qui suggère une meilleure qualité des résumés générés.

```
/opt/conda/lib/python3.10/site-packages/transformers/generation/utils.py:1355: UserWarning: Using the model-agnostic default `max_length` (=20) to control the gen
eration length. We recommend setting `max_new_tokens` to control the maximum length of the generation.
  warnings.warn(
32]: TrainOutput(global_step=208, training_loss=2.523646428034856, metrics={'train_runtime': 258.151, 'train_samples_per_second': 6.322, 'train_steps_per_second': 0.80
6, 'total_flos': 1358580296355840.0, 'train_loss': 2.523646428034856, 'epoch': 8.0})
```

Web Scrapping Aljazeera (https://www.aljazeera.net/)

Dans ce processus de collecte de données, l'objectif était d'extraire des informations du site web Al Jazeera. On extrait **"url" (URL de l'article), "titre" (titre de l'article), "categorie" (catégorie), "texte" (texte de l'article).** Le DataFrame compte 350 entrées et fournit une collection complète de données pour une analyse ou un traitement ultérieur.

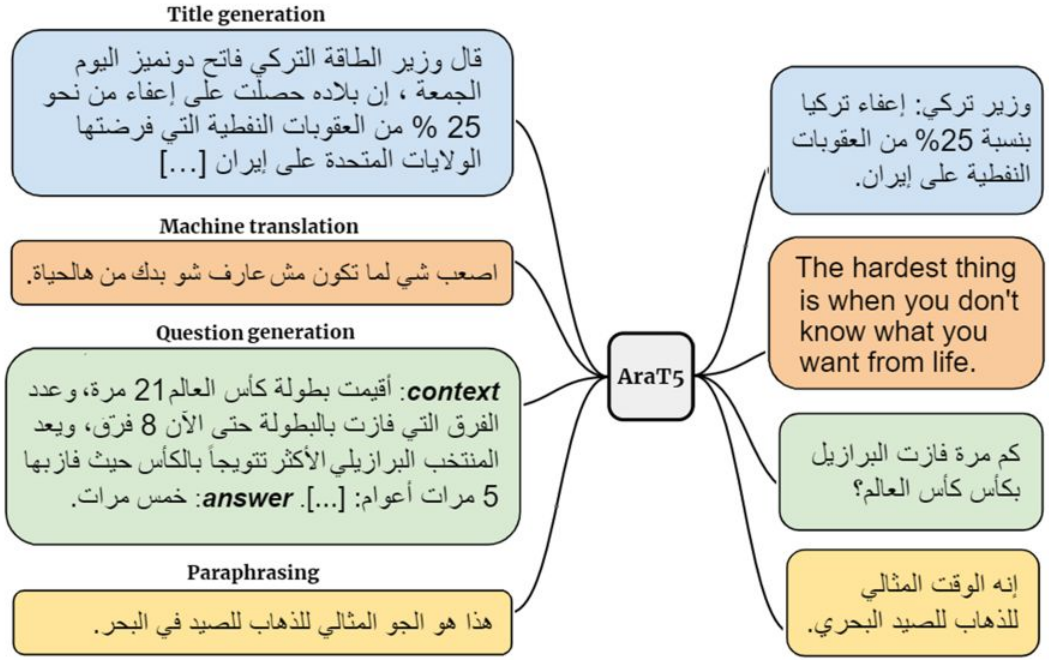
```
# Charger le DataFrame contenant les URLs
df = pd.read_csv("/kaggle/working/Aljazeera_dataset.csv")
df.head(50)
```

	id	url	title	category	summary	texte
0	1	https://www.aljazeera.net/news/2023/12/29/%d8%...	حزب الله يقصف تجهيزات ومعدّات تجسّس إسرائيلية ...	news	قال حزب الله اللبناني في بيانات منفصلة، إن حنا...	أعلن حزب الله اللبناني، الجمعة، استهدافه تجهيز...
1	2	https://www.aljazeera.net/news/2023/12/29/%d8%...	إصابة 4 مستوطنين بعملية دهس قرب الخليل	news	أفادت القناة 12 الإسرائيلية بإصابة 4 مستوطنين ...	أفادت القناة 12 الإسرائيلية بإصابة 4 مستوطنين ...
2	3	https://www.aljazeera.net/news/2023/12/29/%d8%...	بعد اختراق الاحتلال كاميرات المراقبة.. حزب الل...	news	دعا حزب الله سكان البلدات الحدودية جنوب لبنان ...	دعا حزب الله سكان البلدات الحدودية جنوب لبنان ...
3	4	https://www.aljazeera.net/news/2023/12/29/%d9%...	ولاية "مين": ترامب غير مؤهل لمنصب الرئيس	news	في أزمة جديدة، قضت ولاية "مين" الأميركية بعدم ...	قضت ولاية مين الأميركية -أمس الخميس- بعدم أهلي...
4	5	https://www.aljazeera.net/news/2023/12/29/%d8%...	الاحتلال يطلق النار على قافلة مساعدات في غزة	news	قالت وكالة الأونروا اليوم إن جنودا إسرائيليين ...	قالت وكالة غوث وتشغيل اللاجئين الفلسطينيين (أو...
5	6	https://www.aljazeera.net/news/presstour/2023/	كاتب إسرائيلي: لا أحد يستطيع	news	قال الكاتب الإسرائيلي عدعون ليفي	قال الكاتب الإسرائيلي عدعون ليفي

AraT5-base

AraT5: Text-to-Text Transformers for Arabic Language Generation

- AraT5 est une variante du modèle T5 (Text-to-Text Transfer Transformer) spécifiquement entraînée pour le traitement du langage arabe
- .AraT5 se décline en trois versions :
  AraT5MSA - AraT5Tweet - AraT5
- Ces modèles AraT5 ont été développés et entraînés pour effectuer diverses tâches de génération de langage en arabe, telles que la traduction automatique, le résumé de texte, la génération de titres de nouvelles et la génération de questions. Ils ont été évalués à l'aide du benchmark ARGEN.



Evaluation fine tuning with [UBC-NLP/AraT5v2-base-1024] ibtissam369/AraT5v2-base-1024-finetuned-ALjazeera

```
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
```

Tracking run with wandb version 0.16.1

Run data is saved locally in /kaggle/working/wandb/run-20231230_000616-20zb2a3m

Syncing run **pious-dust-9** to Weights & Biases (docs)

View project at https://wandb.ai/arab/huggingface

View run at https://wandb.ai/arab/huggingface/runs/20zb2a3m

[65/65 00:31, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	No log	3.912931	0.000000	0.000000	0.000000	0.000000	10.906200

😔**out of memory**

```
/opt/conda/lib/python3.10/site-packages/transformers/generation/utils.py:1355: UserWarning: Using the model-agnostic default `max_len
eration length. We recommend setting `max_new_tokens` to control the maximum length of the generation.
  warnings.warn(
```

```
]: TrainOutput(global_step=65, training_loss=12.605279071514422, metrics={'train_runtime': 78.2143, 'train_samples_per_second': 3.286, '
1, 'total_flos': 223101562687488.0, 'train_loss': 12.605279071514422, 'epoch': 1.0})
```

+ Code + Markdown

You can now upload the result of the training to the Hub, just execute this instruction:

Conclusion

En conclusion, les RNN, LLMs offrent des avantages significatifs pour les tâches de résumé grâce à leur capacité à capturer les dépendances contextuelles à long terme et à être adaptés à travers le fine-tuning. Cependant, il reste des défis à relever, tels que la génération de résumés cohérents et la gestion des ressources computationnelles. De plus, l'extension de ces modèles à des langues spécifiques comme le darija nécessite des efforts supplémentaires de collecte de données et de formation.

Reference

Nagoudi, E. M. B., Elmadany, A., & Abdul-Mageed, M. (2022). AraT5: Text-to-Text Transformers for Arabic Language Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 628–647). Dublin, Ireland: Association for Computational Linguistics.
repository: https://github.com/UBC-NLP/araT5.

مدرسة علوم المعلومات
ⵜⵉⵎⵔⴰⵜ ⵏⵜⵎⵓⵙⵙⵓⵔⵉⵏ ⵉⵏⴼⵓⵔⵎⴰⵜⵉⵣⵏ
ECOLE DES SCIENCES
DE L'INFORMATION

Merci pour votre attention !