Automatic summarization of legal documents using NLP

Hackathon - Airlingo TeX

Norgile BONOU - Camila KRIKA - Clémentine LE SECH - Laura RAVOI

19th - 25th February 2024

AIRBUS

One pager - Airlingo Tex

Introduction | Datas | Algorithms | Results | Conclusion

Introduction

What ?
7 days for automatic summarization of legal documents using NLP

👥 Who ?
AirlingoTex team:

- Norgile Bonou
  norgile.n.bonou@airbus.com
- Camila Krika
  camila.krika@airbus.com
- Clémentine Le Sech
  clementine.le-sech@airbus.com
- Laura Ravoi
  laura.ravoi.external@airbus.com

Datas

Datas used:
Concatenation of 2 datasets:
- Open source dataset
- Internal dataset

> 📁Input (.json):
859 rows and 3 columns
"Original_text": text to summarize
"Reference_summary": text summarized
"Uid" : unique ID

Preprocessing:
For each "original_text" & "reference_summary":
Tokenisation
- Padding
- Attention mask

Using of dataloaders

Splitting:
Train: 80% - Validation: 20%

Algorithms

Algorithms tested:
Extractive methods : TextRank
Abstractive methods:
- BERT
- T5-Base
- Flan-T5
- T5-Small

✅Final choice:
T5-Base
+ Use of PEFT (Parameter-Efficient Fine-Tuning) for reducing computational costs

Time/Ressources:
- Kaggle/Google Collab
- GPU T4x2

Results

Evaluation: ROUGE
- ROUGE-1
- ROUGE-2
- ROUGE-L

Scores:
- **ROUGE-1: 0,73**
- **ROUGE-2: 0,68**
- **ROUGE-L: 0,72**
- **ROUGELsum: 0,73**

> 📁Output (.json):
3 columns: "original_text" , "predicted_summary", "uid"

Conclusion

Limits:
We tried to use human feedback but we had some issue regarding computing capacity.

To improve our model:
- Using of human feedback
- Increase computing capacity in order to train T5 Large

Datas used:

Concatenation of **2** datasets:

- Open source dataset
- Internal dataset

> Input (.json):

859 rows and **3** columns : “original_text” , “reference_summary”, “uid”

Description of the final dataset used:

	index	count	unique	top	freq
0	original_text	859	856	The Seller reserves the right to change the su...	2
1	reference_summary	859	701	terms may be changed any time at their discret...	11
2	uid	859	859	train_sum01	1

> On average, the original text contains 78 words across the entire dataset.

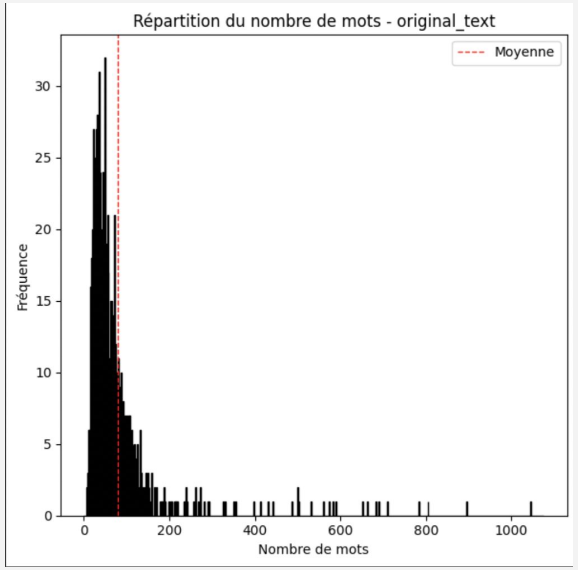
> On average, the summarized text contains 22 words across the entire dataset.

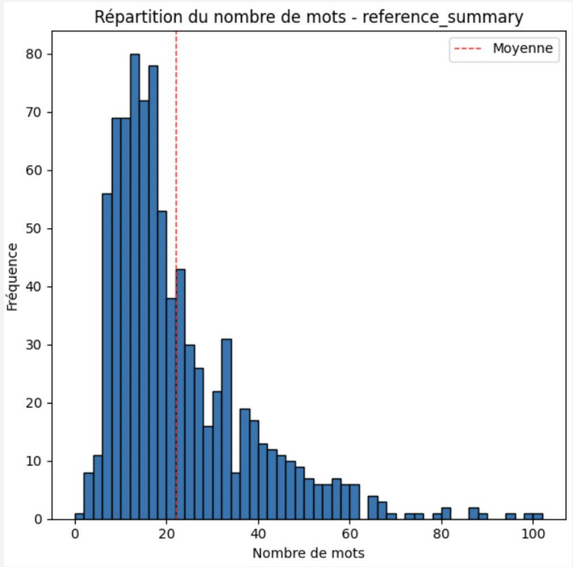
Data splitting	TRAIN (80%)	VALIDATION (20%)
Shape (nrow, ncol)	(687,3)	(172,3)

```
Pour la colonne 'reference_summary':
Moyenne du nombre de mots: 22.08498253783469
Nombre minimum de mots: 1
Nombre maximum de mots: 102

-----------------------------------------------------

Pour la colonne 'original_text':
Moyenne du nombre de mots: 78.74388824214202
Nombre minimum de mots: 7
Nombre maximum de mots: 1077
```





Preprocessing

For each “original_text” & “reference_summary”:

1. **Tokenization** : divide the text into words (token)
   a. Padding: is used to ensure that all sequences have the same length,
   b. Attention mask: are used to indicate which tokens should be attended to during processing, taking into account the presence of padding tokens.

> These techniques are essential for effectively handling variable-length sequences in NLP tasks.

2. **Use of data loaders** : load of data

In order to predict automatic summaries of legal documents, we initially conducted a state-of-the-art review on the subject. We identified two types of methods: extractive and abstractive methods for automatic document summarization.

Extractive Method: TextRank…
- Extractive methods generate summaries by selecting and extracting relevant sentences or passages from the source text.

Abstractive Method: Bert, GPT, T5…
- In contrast to extractive methods, abstractive methods generate a summary by creating new sentences that may not necessarily exist in the source text.
- These methods use natural language generation techniques to produce a summary that reflects the overall meaning of the source text, but may be phrased differently or contain additional information compared to the original text.
- Abstractive summaries are often more fluid and concise than extractive summaries, but typically require a deeper understanding of the source text.

We then tested both methods, but ultimately preferred to choose **an abstractive model** as its scores were superior: **T5-BASE**

To achieve this, we opted to enhance our model using the **PEFT** (Parameter-Efficient Fine-Tuning) method. PEFT is a library designed to efficiently adapt large pretrained models to various downstream applications by fine-tuning only a small number of additional parameters, thereby significantly reducing computational costs while maintaining comparable performance to fully fine-tuned models.

Evaluation:
The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score is a set of metrics used to evaluate the quality of automatic summaries by comparing them to reference summaries. ROUGE measures the overlap between the n-grams, words, or phrases in the generated summary and those in the reference summary.

Model tested	Scores	Meaning
✅ T5-Base	**ROUGE-1: 0,73** **ROUGE-2: 0,68** **ROUGE-L: 0,72** **ROUGEL Sum: 0,73**	ROUGE-1 (R-1): Measures the overlap of unigrams (individual words) between the automatic summary and the reference summary. A ROUGE-1 score of 0.73 indicates that 73% of the words in the reference summary are also present in the automatic summary. ROUGE-2 (R-2): Measures the overlap of bigrams (consecutive word pairs) between the automatic summary and the reference summary. A ROUGE-2 score of 0.68 indicates that 68% of consecutive word pairs in the reference summary are also present in the automatic summary. ROUGE-L (RL): Measures the longest common subsequence between the automatic summary and the reference summary. A ROUGE-L score of 0.72 indicates that 72% of the longest common subsequence between the two summaries. ROUGELsum: It's the average of ROUGE-L and ROUGE-1 scores. ROUGELsum is also 0.73.
T5-Small	ROUGE-1 Score Global: 0.2284 ROUGE-2 Score Global: 0.0889 ROUGE-L Score Global: 0.1875 Cosine Similarity Global: 0.2060	ROUGE-1 Score Global: 0.2284 means that the global ROUGE-1 score is 0.2284. This indicates the quality of the automatic summary compared to the reference summary in terms of overlap of individual words. ROUGE-2 Score Global: 0.0889 means that the global ROUGE-2 score is 0.0889. This indicates the quality of the automatic summary compared to the reference summary in terms of overlap of consecutive word pairs. ROUGE-L Score Global: 0.1875 means that the global ROUGE-L score is 0.1875. This indicates the quality of the automatic summary compared to the reference summary in terms of longest common subsequence. Cosine Similarity Global: 0.2060 means that the global cosine similarity is 0.2060. This indicates the overall similarity between the automatic summary and the reference summary in terms of word representation vectors.

Et voici notre application AirLingoTEX résultante : ICI (via streamlit)
https://airlingotex2-sfvp6huxvwfpywaqxj8qvz.streamlit.app/

Thank you

AIRBUS