BATRICE Toufic & KHALIFA Ottavio **ENSAE 3ème année**

Application de BART à la génération de résumés de textes légaux

Rapport de projet de NLP

Benjamin Muller **2020/2021**

1 Introduction

Dans ce projet, on propose une méthode de génération automatique de résumé de document. Il existe deux approches dans cette discipline :

— Méthodes extractives : on extrait directement les phrases les plus importantes du document à résumer.

— Méthodes abstractives : on génère un texte totalement original, censé synthétiser les informations présentes dans le document à résumer.

Les deux approches présentent un certain nombre d'avantages et d'inconvénients. Nous allons ici proposer de les mélanger pour répondre à notre cas d'usage.

Le jeu de données considéré ici est : *Legal Case Reports Data Set*.Il s'agit d'un corpus d'environ 4000 rapports émis par la justice australienne. Chaque rapport est accompagné d'un résumé réalisé par un humain, appelé "catchphrases", et censé servir de standard pour des tâches de résumé de texte.

2 Exploration des données et méthodologie

On commence par observer les résumés de texte de référence du jeu de données. Ceux-ci ne sont pas formés de phrases, mais plutôt de groupes nominaux contenant des mots-clés importants.

Exemple : claim of direct and indirect discrimination in relation to independent travel criteria imposed by airline respondent. criteria require disabled passengers who cannot comply to fly with carer. applicants seek maximum costs order under o 62a r 1.

L'important sera donc de restituer un maximum de ces mots-clés dans le résumé généré par le modèle. L'évaluation du modèle peut donc se faire sur des critères relativement simples. Nous avons alors pensé à la métrique ROUGE présentée dans l'article [Lin04] qui propose une évaluation facilement interprétable des résumés de texte (détails dans le Notebook).

On s'intéresse maintenant à la taille des documents que l'on étudie. On remarque tout de suite que beaucoup d'entre eux ont une taille supérieure à 100000 tokens, ce qui est énorme. Après avoir retiré tous les cas extrêmes on obtient la distribution suivante :

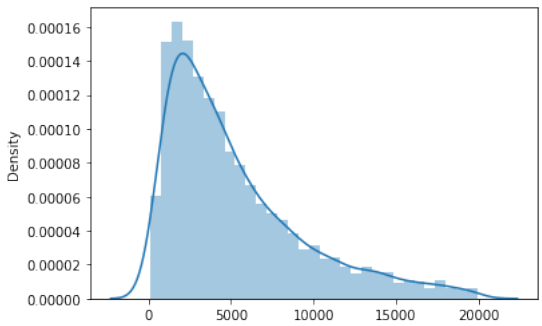

FIGURE 1 – Distribution du nombre de mots dans le dataset

Une grande partie des documents semblent comporter plus de 2000 mots, ce qui les rend trop grands pour être directement mis en entrée d'un Transformer usuel comme GPT2 ou BERT, dont les longueurs maximales possibles en arguments n'excèdent pas 1024. Il est donc primordial d'extraire des informations importantes des documents avant de les passer dans un transformer. Pour cela, on propose d'utiliser en premier une méthode extractive, puis de résumer les phrases extraites de façon abstractive à l'aide d'un transformer.

3 Détails du modèle

Pour la partie extractive de notre modèle, nous proposons d'utiliser l'embedding Glove, puis nous réalisons un k-means dans l'espace latent. Nous récupérons ensuite les phrases les plus proches des centroïdes. Cette méthode est inspirée de l'article [MPe18]. Pour la partie abstractive, nous nous sommes arrêtés sur BART, dont une version a été pré-entraînée par Facebook pour la tâche précise de génération de résumés de texte.

4 Evaluation

4.1 Analyse qualitative

D'un point de vue qualitatif, les résultats (visibles dans le Notebook) ne sont pas très concluants. En effet, la méthode extractive semble bien extraire des phrases importantes, et souvent longues, et lire ce qu'elle renvoie permet souvent de se faire une idée de l'article. En revanche, le transformer renvoie une version abrégée de ce résumé extractif souvent peu différente d'un point de vue sémantique, et ne voulant souvent rien dire. Le modèle de BART que nous utilisons est pré-entraîné pour cette tâche, mais un fine-tuning semble indispensable pour cette tâche. Nous en revenons donc au même problème : nous ne pouvons pas mettre les données en entrée d'un transformer car les documents excèdent largement la longueur maximale qu'ils peuvent prendre.

Malgré cela, on remarque la présence d'un certain nombre de mots-clés dans le résumé obtenu.

4.2 Analyse quantitative

Pour l'analyse quantitative, comme annoncé, nous avons utilisé la métrique ROUGE. Voici les résultats que l'on obtient en moyenne sur 100 échantillons quelconques, en utilisant seulement la méthode extractive (notations détaillées dans le Notebook) :

$$F = 0.18, P = 0.11, R = 0.42$$

On obtient un F1 score peu élevé en raison d'une précision P peu élevée. On pouvait s'y attendre, puisque les résumés de référence sont de simples groupes nominaux, plus proches de l'énumération de mots-clés que de phrases complètes, contrairement au résumé que nous avons généré. Notre résumé est donc fatalement plus long, et contient davantage de prépositions et de ponctuation. En revanche, le rappel R obtenu est assez haut, ce qui prouve que cette méthode nous a permis d'extraire une bonne partie des mots-clés importants du document.

En ajoutant le transformer, le rappel baisse mais la précision augmente un petit peu. On peut l'imputer au fait que le résumé diminue encore. Le résumé abstractif nous a permis d'être plus concis sans perdre une trop grosse part des mots clés.

$$F = 0.18, P = 0.14, R = 0.35$$

Malheureusement, bien que pré-entraîné, le transformer ne propose pas de réelle analyse sémantique des phrases extraites, et ne fait qu'en raccourcir quelques unes. Un fine-tuning aurait été bienvenu, mais il ne nous aurait pas aidés à obtenir des résumés sous forme de phrases construites au vu de la qualité des résumés de référence.

5 Conclusion

Nous avons ici mis en évidence la pertinence de l'analyse extractive pour la détection de mots-clés. L'analyse abstractive réalisée sur les phrases extraites n'a en revanche pas été concluante. Une éventuelle suite à ce projet pourrait être d'entraîner un modèle de transformer non pas sur ce dataset, mais sur des textes légaux en général, pour qu'il apprenne leurs spécificités. Nous aurions aussi pu tester d'autres modèles de transformers tels que GPT2.

Références

[Lin04] Chin-Yew Lin. Rouge : A package for automatic evaluation of summaries. 01 2004.

[MPe18] H.R Divakar M.R Prathima1 et. Automatic extractive text summarization using k-means clustering. 2018.