Projet Interpromo 2021

Groupe 8 : Innovation

Résumés d'articles

Cheffe : Flora Estermann
Cheffe adjointe : Katia Belaid
Chef qualité : Tanguy Dirat

Samba Seye
Célya Marcélo
Théo Saccareau
Damien Sonneville
Mélina Audiger
Jordi Mora Fernandez

SOMMAIRE

1. Introduction ……...……....1

2. État de l'art ……2

2.1. Amélioration des résumés à partir des commentaires utilisateurs……2
2.2. Mise à jour des embeddings……...4

3. Résultats ……..…5

4. Discussion ……..……5

1. Introduction

La partie NLP de notre projet s'est concentrée sur la mise à profit des feedbacks des utilisateurs pour produire de la valeur ajoutée, notamment en récupérant le contenu textuel et des infos sur le comportement des utilisateurs à travers leurs interactions avec le moteur de recherche. De nouvelles idées d'application peuvent alors émerger, notamment sur le moyen d'intégrer les commentaires utilisateurs à la construction de résumés abstractifs, ou bien d'améliorer les embeddings obtenus en pondérant les mots cités fréquemment dans les requêtes pour que les résultats affichés soient plus pertinents aux yeux des utilisateurs par exemple.

2. État de l'art

2.1. Amélioration des résumés à partir des commentaires utilisateurs

Les méthodes du résumé : par abstraction ou par extraction.

En général, les algorithmes de résumé sont soit extractifs, soit abstractifs.

Les algorithmes d'extraction forment des résumés en identifiant et en collant ensemble les sections pertinentes du texte d'entrée. Ainsi, ils ne dépendent que de l'extraction des phrases du texte original. Pour une telle raison, les méthodes d'extraction produisent naturellement des résumés grammaticaux et nécessitent relativement peu d'analyse linguistique. Cependant, les résumés produits peuvent manquer de cohérence ou contenir des phrases répétitives.

Les méthodes de résumé abstractives imitent, jusqu'à un certain degré, le processus naturel accompli par l'homme pour résumer un document. Par conséquent, elles produisent des résumés plus similaires aux résumés manuels. Ce processus peut être décrit par deux étapes majeures : la compréhension du texte source et la génération du résumé, souvent en apprenant au modèle l'utilisation de paraphrases ou fusion et compression de phrases.

A travers notre problématique évoquée dans l'introduction, on souhaite améliorer les contenus des résumés d'articles produits en y intégrant des éventuels commentaires utilisateurs. Pour cela, on suppose que les données dont nous disposons sont une concaténation du contenu de l'article et des commentaires utilisateurs. Comme nous n'avons aucun commentaire utilisateur à notre disposition, on peut les simuler en partant du principe que les deux premiers tiers d'un article correspondent au contenu et que le dernier tiers est constitué des commentaires utilisateurs.

Ainsi, pour produire un résumé de cette concaténation d'informations textuelles, il semblerait que les méthodes abstractives soit plus adaptées que celles d'extraction qui pourraient être biaisées par le fait que la partie commentaire utilisateur ne soit pas écrite de la même façon que le contenu de l'article (pas dans notre situation où on simule ces commentaires mais dans la situation à terme où ils seront vraiment écrits par un utilisateur).

T5 : le nouveau modèle de transformateur de Google

Concrètement, pour résumer un texte donné de manière abstractive, nous avons utilisé la librairie ***Huggingface's Transformers***. Le T5 est donc l'un des modèles pré-entrainés fournit par cette libraire. Il est très utilisé dans plusieurs tâches de NLP (*Natural Language Processing*), il obtient des résultats de pointe pour la recherche de résumé mais aussi pour la traduction automatique ou encore les questions-réponses.

Son utilisation est très simple, pour commencer, il faut effectuer un **prétraitement** du texte. Ce prétraitement est particulièrement succinct : il suffit de supprimer les espaces en trop et de supprimer les sauts de lignes.

Ensuite, comme le T5 est un modèle codeur-décodeur, l'étape suivante consiste à **encoder** le texte prétraité, c'est-à-dire, convertir le texte (chaîne de caractère) en une séquence de nombres entiers en utilisant un tokenizer (analyseur lexical), lui aussi déjà pré-entraîné.

L'étape clé du modèle est bien évidemment la **construction du modèle construisant le résumé abstractif**. Il est possible de jouer avec une petite dizaine de paramètres, pour déterminer les valeurs optimales nous nous sommes aidés de la documentation et d'exemples disponibles sur le Web.

Enfin, la dernière étape consiste bien évidemment à **décoder**, c'est-à-dire à retransformer la nouvelle liste d'entiers retournée à l'étape précédente en un texte compréhensible (chaîne de caractères) correspondant au résumé.

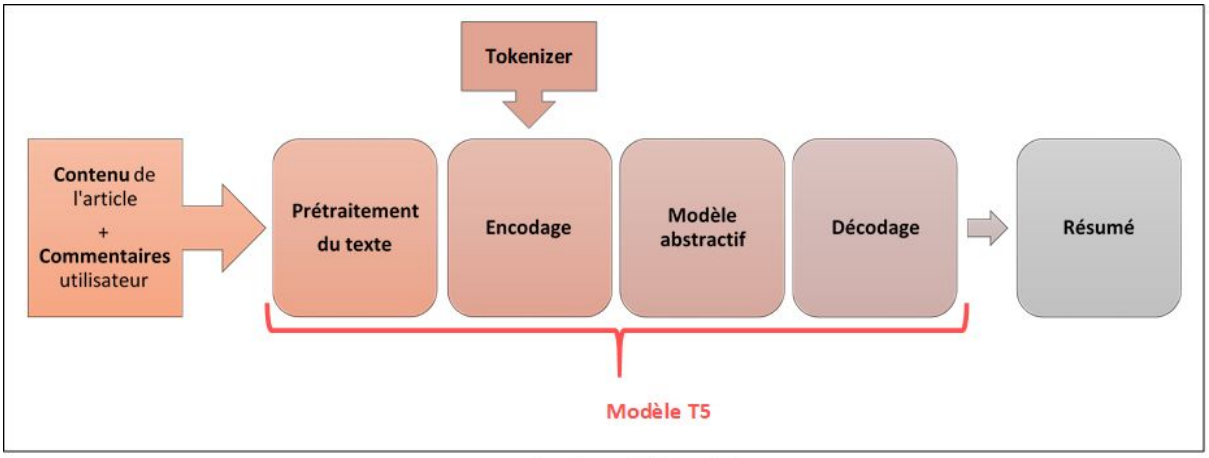

Figure 1 - Résumé simplifié du modèle T5

Paramètres utilisés dans notre étude :

{ *num_beams : 4,*
no_repeat_ngram_size : 2,
min_length : 30,
max_length : 100,
early_stopping : True
}

2.2 Mise à jour des embeddings

L'idée consiste à utiliser l'information textuelle contenue dans les requêtes utilisateur pour mettre à jour les différentes représentations de texte utilisées pour construire nos modèles et ainsi rendre les résultats de notre système plus significatifs aux yeux des utilisateurs.

Par exemple, en augmentant le poids des mots qui reviennent régulièrement lors de sessions de recherche, certaines phrases pourraient être davantage mises en avant lors de la construction de résumés, et ces mêmes résumés présentés par la suite à l'utilisateur lui sembleraient peut-être plus pertinents vis-à-vis de sa requête.

La technique utilisée est relativement simple :
On initialise un dictionnaire ayant pour entrée les mots présents dans le vocabulaire de la représentation de texte choisie, puis on incrémente de 1 la valeur de chaque entrée lorsque le mot apparaît dans une nouvelle requête.
En convolant les valeurs de ce dictionnaire comme un filtre (normalisé au préalable) sur les embeddings sélectionnés pour mettre à jour les mots concernés, on obtient ainsi de nouveaux poids pour les mots récurrents dans les requêtes.

3. Résultats

Application du T5 :

Le T5 donne de bons résultats pour résumer un texte en anglais. Il est très utile pour regrouper un événement avec une date, une citation, une personne ou un lieu. On a aussi essayé de l'appliquer sur des textes en français et les résultats étaient tout aussi concluants. Voici un exemple avec un court article :

"Sans surprise, la Chambre des représentants a voté ***mercredi 13 janvier*** *l'impeachment (mise en accusation) de Donald Trump pour incitation à l'émeute, une semaine après l'irruption violente de ses partisans dans le bâtiment du Capitole à Washington, où siègent les deux chambres du parlement des États-Unis.* ***222 ont voté pour la résolution, 197 contre, tous républicains même si le chef du groupe parlementaire de droite, Kevin McCarthy, a reconnu que Trump n'était « pas exempt de responsabilités »****. C'est une première, aucun autre hôte de la Maison-Blanche avant lui n'ayant subi deux impeachments. Mis en accusation une première fois en décembre 2019 dans le scandale ukrainien, Trump avait été acquitté à l'issue de son procès devant le Sénat majoritairement républicain. Il devrait de nouveau échapper à la destitution."*

Dans cet exemple, le modèle a su regrouper les informations ensemble, dont la date, pour donner le résultat suivant :

"Mercredi 13 janvier, 222 ont voté pour la résolution, 197 contre, tous républicains même si le chef du groupe parlementaire de droite, Kevin McCarthy, a reconnu que Trump n'était « pas exempt de responsabilités »."

Bien sûr, le modèle est encore plus efficace sur un long article et arrive à regrouper des informations venant de divers endroits, divers phrases et de les combiner avec une logique à la hauteur d'un être humain.

4. Discussion

Il existe plusieurs modèles pour les résumés abstractifs. Certains très complexes et sûrement plus efficaces. Mais vu le temps imparti, se lancer dans de tels modèles aurait demandé beaucoup plus d'efforts en termes d'implémentation.