université virtuelle Burkina Faso

Master Fouilles de Données et Intelligence Artificielle

Module : **Natural Language Processing (NLP)**

Sujet : **Résumé de Texte**

ETUDIANTS: Ali BARRO, Boubacar KOANDA

ENSEIGNANT: Dr Abdoul Kader KABORE

Table des matières

I. Introduction 2

II. Les étapes principales du projet 3

III. Les résultats obtenus et leur analyse 4

IV. Limites identifiées 7

V. Pistes d'amélioration 8

VI. Conclusion 9

I. Introduction

Dans le cadre du Master en Fouilles de Données et Intelligence Artificielle à l'Université Virtuelle du BURKINA FASO, un cours le Traitement Langage Naturel (Natural Langage Processing ou NLP en anglais). Ce cours a été dispense par Dr Abdoul Kader KABORE. A l'issu du cours théorique et quelques travaux dirigés, un exercice pratique a été proposé par l'enseignant afin de mettre en pratiques les différentes techniques vues dans le cours. L'objectif de ce projet est de d'évaluer les performances de deux approches de résumé automatique à savoir **la méthode extractive** qui consiste en une sélection de phrases clés directement issues du texte source. La deuxième méthodes est basée sur une **approche abstractive** dont la génération de résumés reformulés est basé sur des architectures de type Transformer.

II. Les étapes principales du projet

Pour mener à bien cette de résumé de texte à partir des méthodes basées sur les modèles de l'Intelligence Artificielle, nous avons adopté les étapes suivantes :

- **La préparation des données** : dans ce exercice, le jeu de données *CNN/Daily Mail*. C'est l'un des datasets les plus populaires pour le résumé automatique de texte. Il a été conçu pour des tâches de traitement du langage naturel (NLP), principalement pour l'entraînement et l'évaluation de modèles de génération de résumés (80 % pour l'entraînement, 10 % pour la validation, et 10 % pour le test). Ces données on été nettoyées et cela a consisté en une suppression des espaces multiples et des caractères spéciaux.
- **Modèle extractif**:
  - Modèle utilisé: BERT.
  - Méthodologie: extraction des phrases clés via un scoring basé sur les embeddings des phrases.
- **Modèle abstrait**:
  - Modèles utilisés: BART et T5.
  - Entraînement pour générer des résumés cohérents et informatifs à partir des textes sources.

III. Les résultats obtenus et leur analyse

A l'issu des différentes techniques ci-dessous appliquées, nous avons obtenu un certain nombre de résultats.

- Pour le modèle **modèle extractif**, les valeurs des métriques se présentent comme suites (confert figure 1 ci-dessous):
    - ROUGE-1: **0.3338**
    - ROUGE-2: **0.0787**
    - ROUGE-L: **0.2516**
    - BLEU: **2.6523**
- Au niveau de la **méthode abstrait**, nous avons:
    - ROUGE-1: **0.3838** (+15 % vs extractif)
    - ROUGE-2: **0.1733** (+120 % vs extractif)
    - ROUGE-L: **0.2636** (+4 % vs extractif)
    - BLEU: **2.6523** (identique au modèle extractif).

Nous avons tester ces deux méthodes de résumé de texte sur un autre dataset à savoir Le **jeu de données SAMSum** (figure 2 ci-dessous). C'est un dataset conçu pour le **résumé automatique des dialogues**. Contrairement à des datasets comme CNN/Daily Mail, qui sont basés sur des articles d'actualité, SAMSum se concentre sur des conversations sous forme de dialogues écrits, comme ceux trouvés dans des applications de messagerie instantanée. Les résultats suivants ont été obtenus sur ce jeu de données:

- ROUGE-1 : **0.0847**
- ROUGE-2 : **0.0000**
- ROUGE-L : **0.0595**
- BLEU : **2.6523**.

L’analyse de ses différentes métriques montre qu’en terme de comparaison de méthodes, l'approche **abstraite** surpasse l'extractive en termes de capture de séquences significatives (ROUGE-2) et de cohérence globale (ROUGE-L). Nous notons aussi que Le modèle extractif reste limité dans sa capacité à reformuler les idées, mais garantit une fidélité au texte original.

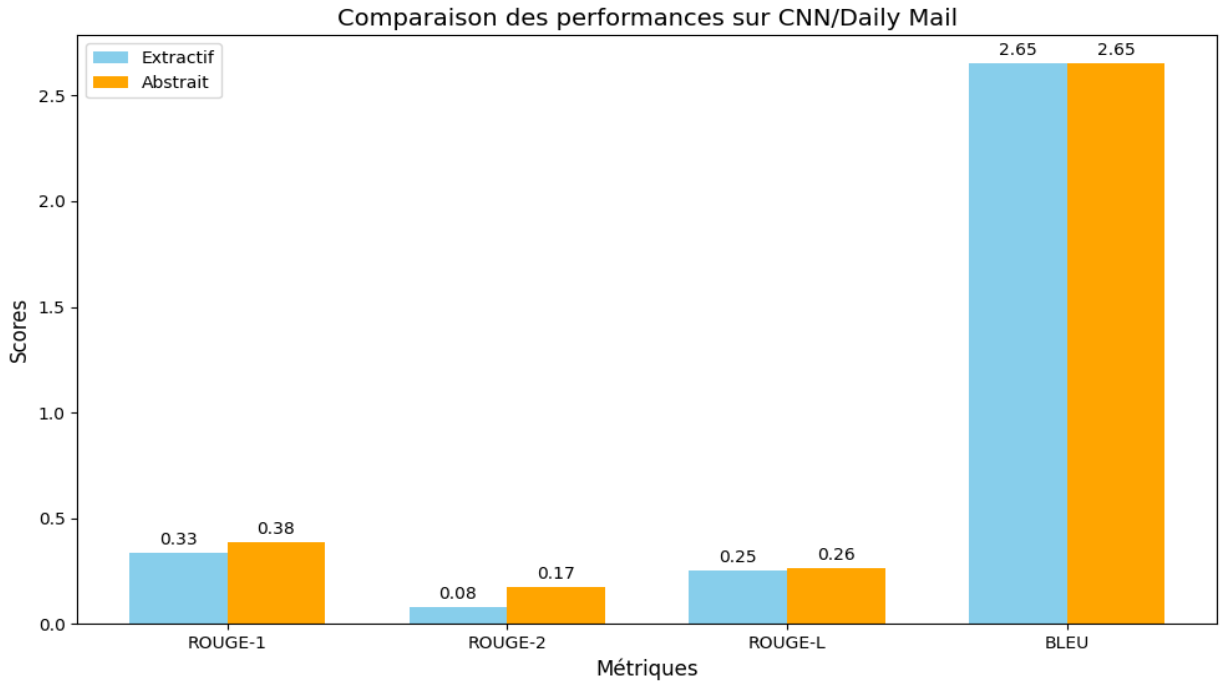

Figure 1: le modèle extractif vs le modèle abstrait sur CNN/Daily Mail

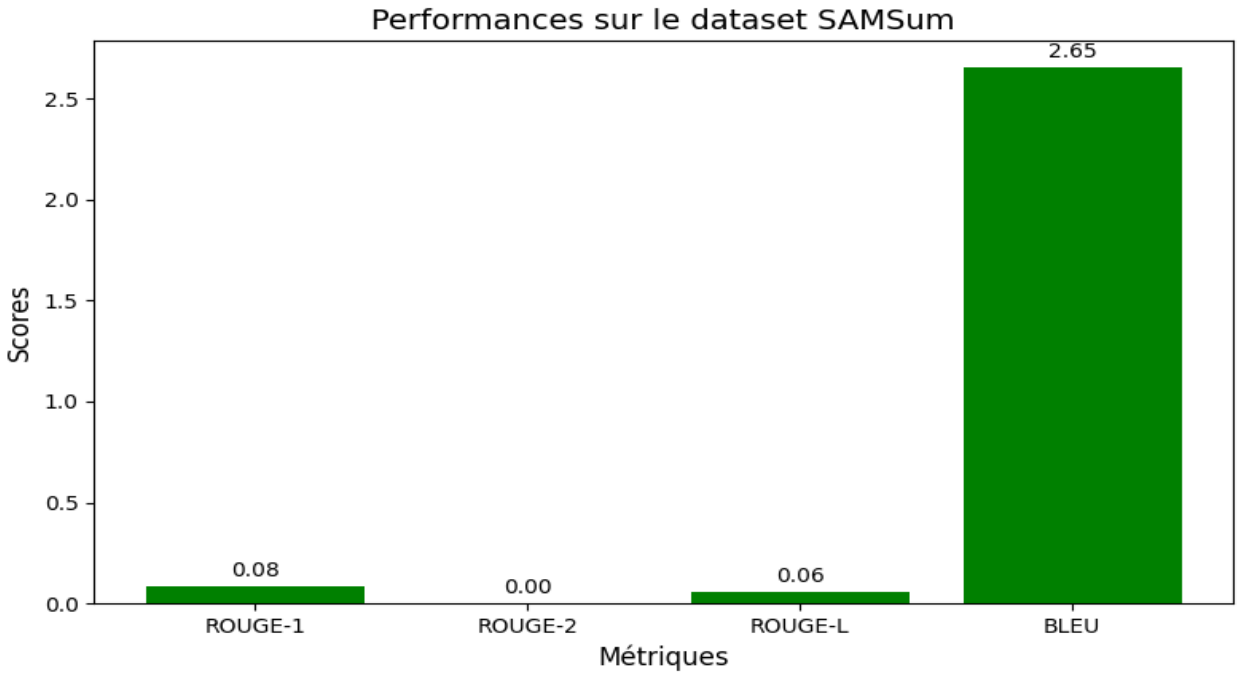

Figure 2: le modèle extractif vs le modèle abstrait sur SAMSum

IV. Limites identifiées

Malgré ces résultats obtenus par les méthodes extractive et abstraite pour la tâche de résumé de texte, nous notons des limites pour celles-ci. Nous avons entre autres:

1. **Modèle extractif** :
    - Répétition des phrases.
    - Sensibilité à la structure des textes.
2. **Modèle abstrait** :
    - Risque de déviation sémantique dans les résumés.
    - Difficultés pour traiter les textes longs.
3. **En général** :
    - Faible capacité de généralisation sur des données externes non vues pendant l'entraînement.

V. Pistes d'amélioration

Comme piste d'amélioration pour ce projet, nous pouvons citer :

- **L'utilisation des techniques avancées** :
  - Régularisation des modèles pour éviter les biais.
  - Ajout de contraintes lors de la génération pour les modèles abstraits.
- **L'entraînement sur des données diversifiées**: qui consiste en une intégration d'ensembles de données contenant des textes complexes ou des dialogues (comme *SAMSum*).
- **La Combinaison des approches**: par la fusion des forces des deux modèles en utilisant un extractif pour garantir la précision, suivi d'un modèle abstrait pour reformuler et contextualiser.

VI. Conclusion

Cet exercice nous a permi de de parcourir des méthodes d'IA dans les taches de résumé de texte. A l'issu du projet, nous notons que **le modèle abstrait** se démarque du **modèle extractif** pour des textes courts et complexes grâce à sa flexibilité, mais les deux approches présentent des atouts complémentaires. Une combinaison des modèles, associée à un fine-tuning sur des données variées, pourrait offrir une solution robuste et généralisable pour le résumé automatique.