Titre du stage : Annotation sémantique de documents multi/cross lingues par apprentissage frugal non supervisé.

Stage de recherche A4/Master 1 ou 2

Contexte général

Les données textuelles envahissent nos quotidiens personnels et professionnels. La recherche de documents pertinents répondant à des besoins métier devient une tâche très fastidieuse et nécessite un investissement en termes d'effort humain à annoter ces documents pour pouvoir les exploiter correctement. L'annotation sémantique de documents multimodaux est un sujet de recherche brûlant que nous proposons d'attaquer selon l'angle des résumés extractifs. Nous supposons qu'une annotation ou une étiquette n'est pas suffisante pour indexer sémantiquement un document. En revanche, un résumé peut représenter collectivement les informations les plus importantes ou les plus pertinentes du contenu d'origine. Par conséquent, toute opération qui émane de l'exploration des documents d'origines telles que la classification, la recherche, la segmentation, ou encore la catégorisation des documents peut être effectuée sur la base du résumé dès lors que ce dernier soit fidèle à l'information d'origine. Ce niveau de fidélité peut être évalué par le biais de différentes métriques qui seront sélectionnées d'une manière automatique en fonction de la tâche.

Les objectifs du stage

L'évaluation des systèmes de traitement automatique de la langue a toujours été un défi majeur pour les chercheurs. En effet ces tâches reposant sur des compétences abstraites de haut niveau, avant d'être difficile à réaliser elles sont tout simplement difficile à évaluer.
Par exemple, afin d'évaluer un simple système de résumé automatique de texte, il est nécessaire de demander à des experts de créer des résumés à la main. Cependant, contrairement à une tâche d'annotation d'images certes fastidieuse mais simple, dans le cas du résumé, l'expert doit comprendre finement les documents sources afin d'en générer une synthèse fidèle. Une fois ces résumés de référence obtenus, il est nécessaire de développer une méthodologie afin de pouvoir évaluer la qualité des résumés générés automatiquement.

La métrique la plus utilisée ROUGE2(Lin, 2004) va simplement compter le nombre de bi-grammes commun entre le résumé de référence et le résumé automatique. Plus un résumé aura de bigrammes communs avec le résumé de référence plus le système sera considéré comme performant. Avec l'essor de l'apprentissage profond, ces métriques ont été améliorées par exemple avec le BERTScore(Zhang & Al, 2020) qui permet de comparer les phrases au niveau vectoriel et ainsi identifier des phrases sémantiquement proches même si elles diffèrent complètement syntaxiquement (par l'usage de synonyme par exemple). Certaines méthodes vont encore plus loin en faisant complètement abstraction de toute annotation de référence. C'est le cas de la métrique BARTScore (Weizhe & Al, 2021). Celles-ci ont été testées dans différentes applications et pour différentes tâches. Dans ce travail, nous visons leur exploitation pour une tâche d'extraction de résumés à partir de documents thématiques. Deux contextes applicatifs seront étudiés dans le cadre de ce stage. Ce travail portera sur le tourisme et s'intéresse particulièrement à l'analyse des sentiments des visiteurs basés sur les données collectées à partir de hotel.com, TripAdvisor, Booking, etc.

Les hypothèses suivantes que nous souhaitons explorer dans ce stage sont comme suit :

1) Ces métriques permettent de construire des résumés extractifs synthétiques pertinents et porteurs de sens. Cependant le cadre méthodologique de ces métriques ne permet pas d'expliquer le processus d'extraction. Or si nous cherchons à annoter sémantiquement les documents via les résumés, il est nécessaire de tracer la pertinence des mots/ phrases.

2) Via les métriques, la quantification de l'hallucination des LLM sera étudiée.

3) Ces métriques indépendantes de toute annotation peuvent être adaptées au cas de résumés génératifs de documents textuels et amorcer une boucle automatique d'amélioration de ces modèles.

4) Ces métriques peuvent être étendues aux cas de résumés génératifs textuels d'images en se basant sur les prompts et sur les commentaires associés aux images.

En perspective, ce travail pourra être envisagé pour l'étendre au résumé d'images en s'inspirant de ces mêmes métriques

Les compétences attendues
Le stagiaire sera en mesure de comprendre l'état de l'art récent sur les métriques d'évaluation et l'apprentissage frugal. De proposer une implémentation modulaire de ces métriques et de les tester sur des jeux de données de benchmark pour se comparer mais également de construire un jeu de données images et textes pour tester les limites de nos hypothèses.

Durée du stage
Le stage durera 6 mois. Il est co-encadré par Nédra Mellouli et Christophe Rodrigues.
Lieu : Campus Cyber, La Défense.

Références
Lin, 2004, {ROUGE}: A Package for Automatic Evaluation of Summaries In Proceedings of the Workshop on Text Summarization Branches Out, WAS.

Zhang & Al, 2020, BERTScore: Evaluating Text Generation with BERT in proceedings of International Conference on Learning Representations, ICLR.

Weizhe & Al, 2021, BARTScore: Evaluating Generated Text as Text Generation in proceedings of Advances in Neural Information Processing Systems, NEURIPS.

Pour Candidater, envoyez votre candidature aux Contacts suivants : Nédra Mellouli et Christophe Rodrigues
Nedra.mellouli@devinci.fr et christophe.rodrigues@devinci.fr

-Votre CV

-Votre lettre de motivation

-Vos derniers bulletins de notes (Master 1 et S9 du master2)