Ingénieur de recherche/Post-Doc confirmé en Traitement Automatique des Langues :

Mise au point d'un assistant virtuel d'enseignement

Type de contrat: CDD 1 an

Informations générales :

Lieu de travail : CNRS/LISN, rue du Belvédère, campus universitaire Université Paris-Saclay

Type de contrat : CDD

Durée du contrat : 12 mois

Date d'embauche prévue : A partir de Juin 2022

Quotité de travail : Temps complet

Rémunération : Selon profil

Niveau d'études souhaité : Docteur/Ingénieur

Expériences souhaitées : travaux dans le domaine du traitement automatique du langage (*TAL*)

Contexte

Nous recherchons un ingénieur de recherche/post-doctorant expérimenté pour travailler au sein du laboratoire LISN (laboratoire mixte CNRS-Université Paris-Saclay), avec des chercheurs spécialisés en Traitement Automatique des Langues (TAL). Cette recherche s'inscrit dans le cadre du programme de maturation de 18 mois entre l'entreprise Professorbob.ai, leader dans l'apprentissage adaptatif, la SATT Paris Saclay (**S**ociété d'**A**ccélération du **T**ransfert de **T**echnologies) et le CNRS. Les postes sont localisés dans les locaux du CNRS LISN (Saclay, 91).

Il s'agit de travailler sur un projet d'assistant virtuel d'enseignement dédié à l'éducation et la formation,

qui fait l'objet d'une collaboration entre le laboratoire et l'entreprise qui travaille sur la mise au point de Professorbob.ai (https://professorbob.ai/)

Cet assistant devra être en mesure d'aider des étudiants dans leurs apprentissages :

- En répondant à des questions en rapport avec les sujets des cours
- En proposant des outils pour l'ancrage de connaissances
- En personnalisant l'apprentissage via des méthodes "d'apprentissage adaptatif".

La création de l'assistant virtuel requiert des connaissances et une maîtrise techniques avancées sur les modèles et problématiques en traitement du langage naturel. Plus spécifiquement, nous nous intéresserons aux problématiques de génération de textes, de recherche d'informations, d'évaluation du langage et de transfert de domaine.

Les récentes avancées en matière de traitement de la langue nous permettent d'envisager la construction d'un tel système, en particulier grâce aux approches neuronales pour la génération de questions ou la recherche d'informations. Malheureusement, si les modèles les plus performants permettent d'obtenir des résultats satisfaisants en langue anglaise, peu de modèles pré-existent pour la langue française. Aussi, même s'il existe des corpus publiquement accessibles pour la tâche de génération de questions, ces corpus ne correspondent que partiellement aux types de question souhaités pour un assistant de cours. Pour pallier ce manque de données, nous travaillons à la mise en place d'un corpus de questions de cours en langue française d'ici à la date du début de contrat.

Les principales problématiques étudiées dans le poste proposé porteront sur la génération de questions et de réponses.

Génération de questions. Avec les données préalablement récoltées, l'objectif est de créer/générer des questions via des modèles de génération de questions basés sur des approches neuronales. Pour mesurer la qualité des questions générées par le modèle, il est souhaitable de considérer des métriques innovantes. En effet, les approches basées sur les n-grammes sont encore aujourd'hui largement dominantes[1]. Néanmoins, récemment, plusieurs approches ont été mises à jour : des approches ne comparant non pas des mots mais des représentations vectorielles contextualisées de sous-mots (Bert score)[2]; des approches proposant d'utiliser un contexte pour vérifier la pertinence des textes générés[1][3, 4] . Notons, par ailleurs, que ces approches obtiennent encore aujourd'hui de moins bons résultats qu'une évaluation humaine.

Génération et sélection de réponses. Une deuxième étape importante du projet concerne la sélection de réponse considérant une requête utilisateur. Cette recherche d'informations correspond donc à l'amélioration et la spécialisation de moteurs de recherche. Cette approche peut être à la fois envisagée via l'utilisation de moteurs de recherche basés sur des fréquences de mots ou des approches via des réseaux de neurones. Une fois les documents pertinents retrouvés, il est nécessaire de produire une réponse, soit par extraction de contenu (réponse extractive) et/ou par génération (réponse abstractive) [5]. Dans le cas de la génération de réponses, il est alors nécessaire de vérifier la pertinence et la véracité des faits retranscrits par le modèle, ainsi, des travaux sur l'évaluation de la génération sont à

1Q² et QuestEval utilisent la génération de questions pour vérifier l'exactitude de faits retranscrits entre un contexte et un texte généré. Ces approches ne sont donc pas exploitables en l'état pour l'évaluation de questions générées.

envisager [2, 6, 7].

Activités

Le but global du projet est d'assister un enseignant en l'aidant à répondre à des questions nombreuses et répétitives des apprenants. Il faut donc apprendre à répondre aux questions, en s'appuyant sur des données fiables, fournies par les enseignants. En s'appuyant sur les travaux récents dans le domaine du TAL, on sait qu'il est possible d'améliorer les systèmes classiques et basiques de réponses à des questions. Cependant, les données au sein desquelles les réponses devront être trouvées ne sont pas les données classiques utilisées dans les campagnes d'évaluation, mais des données en lien avec la discipline en cours d'apprentissage.

Il sera dans un premier temps demandé de traiter les données de questions/réponses récoltées lors de la campagne d'annotation. Le travail consistera donc à formater et nettoyer les données disponibles.

Dans un second temps, les travaux porteront sur la génération de questions, mais aussi sur leur évaluation. Pour cela, il faudra évaluer quels sont les modèles et métriques les plus adéquats, mais aussi mettre en place un protocole d'évaluation pour valider les approches proposées. Il faudra par la suite être en mesure de déployer ces approches sur le système.

Enfin, les approches de sélection/génération de réponses seront étudiées et mises en place afin de permettre des améliorations significatives de l'assistant. Notons aussi que les problématiques d'évaluations étudiées pour la génération de questions pourront s'avérer aussi utiles dans cette dernière étape.

Le travail sera fait dans un cadre collaboratif avec 2 autres chercheurs et devra prendre notamment en compte les axes de recherche de l'équipe : transfer learning, continuous learning et IA conversationnelle.

3. Compétences Requises

- Bonne maîtrise des outils du TAL :
    - Modèles Deep Learning: connaissance théorique et manipulation avancée des RNN, Auto-encoders, Transformers (BERT / Roberta / T5,..), etc.. surtout des modèles de Question Answering, Question Generation, etc..
    - Bibliothèques et frameworks Deep Learning/Machine Learning comme Pytorch, Tensorflow, Keras, NLTK, Spacy, Scikit-learn, etc..
- Algorithmique: très bonne connaissance et maîtrise pratique des algorithmes classiques sur les

textes, arbres, graphe - Statistiques: connaissances des techniques d'échantillonnage

- Expérience du développement et du débogage en Python
- Maîtrise de la démarche Data Science : définition des tâches, définition de métriques de performance, veille technologique, analyse de publications scientifiques, implémentation, fine-tuning et évaluation de modèles
- Anglais scientifique courant
- Aptitude à communiquer et à travailler en équipe

3. Compétences supplémentaires souhaitables

- Moteurs de recherche et traitements textuels: indexation, utilisation d'ElasticSearch, Lucène / SolR, formalisation et recherche d'expressions régulières

4. Niveau de Formation

Doctorat ou Master en rapport avec le Deep learning, idéalement avec le traitement en langage naturel

Contact : Anne VILNAT CNRS/LISN (anne.vilnat@universite-paris-saclay.fr) ET chez Professorbob.ai : Samy LAHBABI (Samy.lahbabi@professorbob.ai)

[1] A. Agarwal et A. Lavie, « Meteor, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output », in *Proceedings of the Third Workshop on Statistical Machine Translation, WMT@ACL 2008, Columbus, Ohio, USA, June 19, 2008*, 2008, p. 115-118. [En ligne]. Disponible sur: https://aclanthology.org/W08-0312/

[2] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, et Y. Artzi, « BERTScore: Evaluating Text Generation with BERT », 2020. [En ligne]. Disponible sur: https://openreview.net/forum?id=SkeHuCVFDr

[3] O. Honovich, L. Choshen, R. Aharoni, E. Neeman, I. Szpektor, et O. Abend, « Q\(^\mbox2\): Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering », *CoRR*, vol. abs/2104.08202, 2021, [En ligne]. Disponible sur: https://arxiv.org/abs/2104.08202

[4] T. Scialom *et al.*, « QuestEval: Summarization Asks for Fact-based Evaluation », in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2021, p. 6594-6604. doi: 10.18653/v1/2021.emnlp-main.529.

[5] M. Lewis *et al.*, « BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, juill. 2020, p. 7871-7880. doi: 10.18653/v1/2020.acl-main.703.

[6] W. Yuan, G. Neubig, et P. Liu, « BARTScore: Evaluating Generated Text as Text Generation », in *Advances in Neural Information Processing Systems*, 2021, vol. 34, p.

27263-27277. [En ligne]. Disponible sur: https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf

[7] T. Sellam, D. Das, et A. Parikh, « BLEURT: Learning Robust Metrics for Text Generation », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, juill. 2020, p. 7881-7892. doi: 10.18653/v1/2020.acl-main.704.