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

Référence :
Lieu de travail : LISN, rue du Belvédère, campus universitaire Université Paris-Saclay
Date de publication : juillet 2021
Type de contrat : CDD
Durée du contrat : 12 mois
Date d'embauche prévue : septembre 2021
Quotité de travail : Temps complet
Rémunération :
Niveau d'études souhaité : Doctorat + expérience complémentaire
Expérience souhaitée : expérience d'encadrement

1. Contexte

Nous recherchons un ingénieur de recherche/post-doctorant expérimenté pour travailler au sein du laboratoire LISN du CNRS et de l'Université Paris-Saclay, avec des chercheurs spécialisés en Traitement Automatique des Langues (TAL).

Il s'agit de travailler sur un projet d'assistant virtuel d'enseignement (*chatbot* dédié à l'éducation et la formation), qui fait l'objet d'une collaboration entre le laboratoire et l'entreprise *The AI Institute* qui travaille sur la mise au point de https://professorbob.ai/.
Ce projet est soutenu par la SATT de Paris-Saclay dans le cadre des projets de maturation.

2 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 des quelles 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.
Dans ce contexte, il faut d'abord recueillir une importante quantité de données, afin de constituer un grand nombre d'associations de questions et de réponses, pour ensuite mettre en œuvre des technologies de TAL pour associer la bonne réponse à une question posée.
Dans une première étape, le travail consistera donc à formater les données pour constituer un corpus puis à modéliser comment constituer cet ensemble de paires question/réponse partir **de données** structurées ou non structurées en relation avec la **discipline** en cours d'apprentissage. Les données sont issues de sources diverses : livres universitaires, polycopiés de cours, sites autorisés en lien avec la discipline en cours d'apprentissage par l'apprenant. Il faudra ensuite évaluer quelles sont les modèles les plus adéquats . La difficulté vient également du fait que l'on souhaite ensuite pouvoir généraliser les recherches, et donc minimiser le travail qui sera à refaire quand on veut transférer le système soit à une autre langue, soit à une autre discipline.

3. Compétences Requises

- Connaissance des les outils du TAL :
    - Modèles Deep Learning: connaissances théoriques et manipulation avancée des RNN, Auto-encoders, Transformers, etc..
    - Bibliothèques et frameworks Machine Learning comme NLTK, Spacy, Scikit-learn, Keras, Tensorflow, Pytorch, etc..
- Algorithmique: bonne connaissance des algorithmes classiques sur les textes, arbres, graphe
- Statistiques: connaissances des techniques d'échantillonnage
- Anglais scientifique courant

3. Compétences supplémentaires souhaitables

Moteurs de recherche et traitements textuel: indexation, utilisation d'ElasticSearch ou SolR, formalisation et recherche d'expressions régulières

4. Niveau de Formation

Doctorat avec de l'expérience

Contacts :

anne.vilnat@universite-paris-saclay.fr