Avis de Soutenance

Madame Thi Hong Hanh TRAN

Spécialité : Informatique et Applications

Soutiendra ses travaux de thèse intitulés

« **Approches neuronales de l'extraction automatique de termes** »

dirigés par Monsieur Antoine DOUCET

Cotutelle avec l'École supérieure internationale de Jozef Stefan (Slovénie)

le jeudi 26 septembre 2024 à 9h30

Lieu : **Jozef Stefan Institute Jamova 39,**
1000 Ljubljana Slovenia
Salle : IPS lecture room

Composition du jury proposé

M. Jatowt ADAM	**University of Innsbruck**
M. Antoine DOUCET	**Université de La Rochelle**
Mme Lefever ELS	**Ghent University**
M. Boudin FLORIAN	**University of Nantes**
M. Marko ROBNIK ŠIKONJA	**University of Ljubljana**
Mme Pollak SENJA	**Jozef Stefan Institute**

Résumé :

L'extraction automatique de terminologie (EAT) est une tâche de traitement automatique du langage naturel (TALN) qui identifie la terminologie spécialisée à partir de corpus spécifiques à un domaine. En réduisant le temps et les efforts nécessaires à l'extraction manuelle des termes, l'EAT est non seulement largement utilisée pour des tâches terminologiques mais contribue également à plusieurs tâches en aval complexes (par exemple, la traduction automatique et la recherche d'informations). Au cours des quarante dernières années, des progrès considérables ont été réalisés pour fournir automatiquement une liste classée des termes candidats à partir de corpus spécialisés ; cependant, les tâches d'EAT restent un problème notoirement difficile. Notre thèse s'est concentrée sur les aspects suivants. En ce qui concerne les scénarios où les données bien annotées sont adéquates pour des paramètres entièrement supervisés, nous avons étudié l'amélioration des approches neuronales en introduisant la tâche comme un problème de classification de jetons (appelé étiquetage de séquence) en utilisant des Transformers comme modèle de base avec une représentation supplémentaire (par exemple, la sémantique des étiquettes) et des couches modifiées (par exemple, mélange d'experts, RNN). De plus, nous avons développé les systèmes actuels en introduisant NOBI, un nouveau régime d'annotation pour mieux capturer les termes imbriqués. En ce qui concerne les scénarios où les données bien annotées des mêmes langues sont limitées, nous avons proposé un apprentissage interlinguistique et multilingue pour mettre en évidence le potentiel du transfert d'apprentissage des langues à ressources riches ou combinées vers des langues moins connues dans les systèmes neuronaux. En ce qui concerne les scénarios où les données bien annotées et les ressources de calcul sont limitées, nous avons proposé un nouveau pipeline utilisant des modèles de langage volumineux (LLM) appelés LlamATE comme prédicteur pour interroger les termes candidats sans aucune étape de réglage fin supplémentaire, en utilisant simplement une démonstration par few-shot avec des étapes d'auto-vérification. Notre étude s'est terminée par les conclusions suivantes. Premièrement, les approches de classification de jetons se sont révélées des méthodes valides et prometteuses pour l'apprentissage entièrement supervisé en extraction de termes. Ces approches surpassent les performances des classificateurs non séquentiels et à séquence binaire, et réduisent les défis de calcul et de stockage mentionnés dans les tests de performance. L'ajout d'une couche MoE au-dessus du modèle neuronal profond (par exemple, (m)DeBERTA) a constamment amélioré les performances par rapport à la configuration de base avec une tête de classification de jetons dense. L'utilisation des régimes d'annotation NOBI a démontré une amélioration visible pour les termes unifoliés et multi-mots du classificateur de jetons entraîné sur le jeu de données dans lequel le nombre de termes imbriqués est suffisamment important. Deuxièmement, avec des données annotées limitées provenant des mêmes langues, nos résultats sur le classificateur de jetons ont mis en évidence l'impact prometteur de l'apprentissage interdomaine multilingue et interlinguistique lors du transfert des connaissances des langues riches vers des langues moins connues. Enfin, LlamATE suggère le potentiel des LLM avec démonstration par few-shot et auto-vérification pour apprendre à partir de quelques exemples dans le même domaine, même sans que le domaine ne soit explicitement indiqué. Il suggère également le potentiel de transfert de connaissances des langues bien couvertes vers des langues moins représentées dans les LLM. Bien qu'ils ne remplacent peut-être pas entièrement les modèles entièrement supervisés, ils peuvent améliorer l'efficacité et la précision en rationalisant le processus de pré-annotation et en accélérant les efforts d'annotation manuelle.