Stage Master 2 / ingénieur

LIMSI-CNRS / AP-HP

Titre : Adaptation d'un système d'apprentissage neuronal à de nouveaux domaines

Mots-clés : apprentissage automatique, traitement automatique des langues, réseaux de neurones, adaptation au domaine, domaine médical, analyse de dossiers patients

Lieu : LIMSI (Orsay), AP-HP (Paris, campus Picpus)

Durée : 4 à 6 mois

Date de début : printemps 2018

Contexte

Le parcours de soin d'un patient dans un hôpital est documenté par des données numériques et structurées (résultats d'analyse, prescription de médicaments, etc.) mais également par un grand nombre de documents textuels rédigés par le personnel soignant : comptes-rendus d'hospitalisation, comptes-rendus d'opérations chirurgicales, lettres entre médecins, etc.

Être capable d'extraire de l'information pertinente de ces documents textuels pour enrichir les connaissances sur le patient et son itinéraire (par exemple, l'histoire de sa maladie, ses antédécents, ceux de sa famille, ses facteurs de risque) permet d'accumuler des données pertinentes sur les parcours de soin. Ces données peuvent par la suite être utilisées dans toutes sortes d'études visant à mieux adapter la prise en charge aux spécificités de chaque patient.

Une des approches populaires pour l'extraction d'information dans les textes consiste à constituer des corpus annotés à la main par des experts et à mettre en oeuvre des outils d'apprentissage automatique. C'est cette piste qui est suivie au LIMSI avec l'élaboration d'un système à base de réseaux de neurones [1][2][3] et l'utilisation d'un corpus annoté en français [4].

Deux difficultés se présentent alors :

- d'une part, l'annotation manuelle est longue et coûteuse, et donc nécessairement faite en quantité limitée.
- d'autre part, les comptes-rendus médicaux utilisent un vocabulaire et une structure propre à chaque domaine (cancérologie, endocrinologie, gastro-entérologie, etc.) et, dans une moindre mesure, à chaque service ou hôpital. Il est impossible à l'heure actuelle d'envisager l'annotation de données de chaque domaine en quantité suffisante pour les modèles d'apprentissage.

Dans le but de réaliser des campagnes d'annotation aussi pertinentes et ciblées que possible, nous souhaitons donc quantifier précisément les besoins et les capacités de nos systèmes à s'adapter à des domaines nouveaux ou faiblement couverts par les annotations manuelles.

Travail attendu

Le ou la stagiaire recruté(e) devra prendre en main les corpus et les systèmes existants en interne. Ces systèmes permettent d'annoter des entités de différents types (procédures, symptômes, maladies, médicaments, etc.) dans les comptes-rendus médicaux. Il réalisera des études sur les différents points suivants :

- quantité des données annotées nécessaires pour obtenir des résultats satisfaisants
- configuration optimale et/ou changements nécessaires aux modèles pour garantir une adaptation efficace à un domaine nouveau comportant peu ou pas de données annotées
- comparaison avec d'autres approches (application de dictionnaires, systèmes à bases de règles)

Compétences souhaitées

Nous recherchons un(e) étudiant(e) ayant des compétences solides en programmation et en apprentissage automatique, intéressé(e) par le traitement de contenu en langage naturel et par une application médicale.

Les compétences en programmation ne sont cependant pas le seul critère, et la personne retenue devra également faire preuve de créativité et d'esprit d'analyse.

Les candidatures doivent comporter :

- Une lettre de motivation
- Un relevé de notes récent
- Les noms et coordonnées de deux personnes référentes
- Un curriculum citae (CV)

Contacts

nicolas.paris@aphp.fr (AP-HP)

aurelie.neveol@limsi.fr (LIMSI-CNRS)

xavier.tannier@upmc.fr (UPMC, LIMICS)

Références

[1]: https://github.com/jtourille/yaset

[2]: Julien Tourille, Olivier Ferret, Xavier Tannier, Aurélie Névéol. Neural Architecture for Temporal Relation Extraction: A Bi-LSTM Approach for Detecting Narrative Containers. in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

[3]: Julien Tourille, Olivier Ferret, Xavier Tannier, Aurélie Névéol. LIMSI-COT at SemEval-2017 Task 12: Neural Architecture for Temporal Information Extraction from Clinical Narratives. in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017).

[4]: Campillos L, Deléger L, Grouin C, Hamon T, Ligozat AL, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLoT). Lang Resources & Evaluation. Springer, Berlin Heidelberg, Germany. 2017:1-31