PSTLAN: projet 2020

Benoit Favre

6 décembre 2020

1 Contexte

Depuis le début de la crise sanitaire liée à la COVID-19, le nombre d'articles scientifiques sur le sujet a explosé. Il est à environ 2500 articles par semaine depuis le mois d'avril [1]. Les médecins, qu'ils soient chercheurs ou praticiens, n'arrivent pas à suivre et "digérer" la masse d'information qu'ils pourraient exploiter pour créer de nouveaux traitements. Un groupe de médecins d'AMU et de l'université de Grenoble ont créé un site, `https://bibliovid.org/`, sur lequel ils postent des évaluation manuelles de l'intérêt des articles publiés. Comme c'est fait manuellement, ils ne peuvent couvrir l'ensemble des articles qui sortent, mais seulement une petite sélection. Votre objectif pour ce projet est de proposer une méthode automatique pour leur faire gagner du temps et potentiellement leur permettre de traiter plus d'articles.

Les problématiques à envisager sont multiples :

- Comment récupérer les textes et métadonnées associées à ces articles ? Dans le domaine médical, il existe plusieurs sites qui référencent les publications au fur et à mesure de leur sortie : PubMed [2], BioRxiv [3], Semantic Scholar [4], etc. Ces sites proposent des API pour récupérer les données à partir d'un identifiant d'article.
- Quelles tâches peut-on automatiser pour faciliter la vie des médecins ? Le site bibliovid contient pour chaque article une catégorie (Diagnostique, Thérapeutique...), plusieurs spécialités (Gériatrie, Dermatologie...), des commentaires textuels (Résultats, Méthode...), un niveau de preuve (Fort, Faible...) et des métadonnées. Chacun de ces éléments peut être prédit automatiquement à partir de l'article. Les tâches de génération de texte sont plus difficiles car elles nécessitent de comprendre et écrire du texte. Les tâches de prédiction de catégories peuvent être envisagées pour ce projet.
- Comment récupérer la supervision pour les tâches envisagées ? On peut récupérer le contenu du site bibliovid et en extraire les articles et étiquettes associées. Un scrappeur a été développé par l'équipe TALEP.

Comme point de départ, vous pouvez utiliser l'extracteur de données dont les sources sont disponibles [5]. L'extraction contient trois fichiers :

- bibliovid.json : des données extraites du site bibliovid, en particulier les étiquettes associées aux articles ainsi que les champs textuels renseignés. En plus, ont été ajoutés pour un sous-ensemble les identifiants pubmed des articles et les résumés associés.
- litcovid.json : des données extraites du site litcovid, en particulier les étiquettes des huit catégories annotées semi-automatiquement pour 77 000 articles (comme expliqué dans leur FAQ [6]).
- cord19-metadata.json : des métadonnées sur environ 300 000 articles médicaux reliés à l'épidémie (titre, résumé, etc). Les données ont été mises à disposition de la communauté par l'Allen

1. Voir `https://www.ncbi.nlm.nih.gov/research/coronavirus/`
2. `https://pubmed.ncbi.nlm.nih.gov/`
3. `https://www.biorxiv.org/`
4. `https://www.semanticscholar.org/`
5. `https://gitlab.lis-lab.fr/bibliovid/scrappers`. Dump récent ici : `https://pageperso.lis-lab.fr/benoit.favre/covid19-data/20201206/`
6. `https://www.ncbi.nlm.nih.gov/research/coronavirus/faq`

AI institute. On peut aussi télécharger une version contenant le texte des articles prétraités `https://www.semanticscholar.org/cord19/download`. Il n'y a pas d'annotations associées.

2 Travail attendu

Il s'agit de proposer une méthode de catégorisation automatique des articles médicaux selon les classes de bibliovid pour faciliter le travail des médecins.

En particulier, les problèmes que vous allez rencontrer sont de l'ordre suivant :

— Quel architecture de réseaux de neuones, quel classifieur donnent les meilleures performances pour cette tâche ?
— Comment prendre en compte les spécificité du domaine médical ?
— Comment tirer parti de l'apprentissage de représentation pour palier au peu de supervision disponible sur bibliovid ?

Le travail attendu est de produire un système fonctionnel, ainsi qu'un rapport expliquant la méthode ayant permis de construire ce système. En particulier, le rapport devra justifier les choix scientifiques à l'aide d'expériences comparant les performances des différentes approches mises en compétition. La métrique à considérer sera le F-score.

Des axes intéressants optionnels de travail sont : un système capable de traiter "en flux" les articles après leur publication sur un des sites les recensant ; une visualisation attrayante pour les médecins.

3 Modalités

Le travail se fait en équipes de 4 à 6 étudiants. Chaque étudiant devra identifier spécifiquement sa contribution originale et son rôle dans l'équipe. Le rapport devra contenir un lien vers un git attestant du code source produit. Il devra faire une vingtaine de pages maximum. Une soutenance le 6 janvier 2021, lors du cours de PSTALN, permettra de présenter les résultats principaux obtenus par l'équipe.

Les critères d'évaluation sont les suivants :

Critère	**Mieux**	**À revoir**
Cadre expérimental	Conditions de test réalistes et rigoureuses	Test sur l'entraînement
Nombre et diversité des approches	Réimplémentation de l'état de l'art	Les trois baselines vues en cours
Explication scientifique des performances	Argumentation étayée par des expériences censées ou des preuves théoriques	Choix arbitraires
Qualité du code	Le code est structuré et réutilisable	le code est obscur ou copié depuis une source publique
Travail d'équipe	Contribution équilibrée et identifiée des membres de l'équipe	Chacun a travaillé dans son coin
Qualité du rapport	Bonne structure, sans fautes, synthétique mais complet	Rapport incompréhensible

4 Calendrier

— Remise du sujet : 7/12/2020
— Composition des équipes : 11/12/2020
— Remise du rapport : 04/01/2021
— Soutenance : 06/01/2021 lors du cours de PSTALN (20 minutes par équipe + questions)