Apprentissage Conjoint de Représentations d'Auteurs et de Documents

Antoine Gourru*, Rohit Yadav**,*, Julien Velcin*

*ERIC UR3083, Université de Lyon, Lyon 2, France
{antoine.gourru,julien.velcin}@univ-lyon2.fr,
**Université Jean Monnet, Saint-Etienne, France
rohit.yadav@etu.univ-st-etienne.fr

Résumé. Les modèles de langue les plus récents utilisent des représentations de mots contextualisés à l'aide de Transformers. Ils ont rapidement dépassé les méthodes état de l'art dans de nombreuses tâches de traitement automatique de la langue. Des versions pré-entraînées de ces modèles sont largement utilisées, mais leur spécialisation pour résoudre une tâche spécifique reste une question centrale. Par exemple, ces méthodes ne produisent pas de représentation à l'échelle du document et de l'auteur, mais seulement du mot. Or comme le montrent Reimers et Gurevych (2019), une simple moyenne des plongements de mots ne suffit pas. En utilisant une approche dite du Variational Information Bottleneck, nous développons une architecture simple pour construire des représentations d'auteurs et de documents à partir de modèles pré-entraînés (Devlin et al., 2019). Nous évaluerons de manière quantitative et qualitative notre modèle sur deux jeux de données : un corpus d'articles scientifiques et un d'articles de presse. Notre modèle produit des représentations plus robustes que l'existant, et donne des résultats compétitifs en classification et en identification d'auteurs.

1 Introduction

Depuis les travaux de Mikolov et al. (2013), de nombreux modèles incorporent des représentations continues des mots, appelées plongements, pour résoudre des tâches de traitement automatique de la langue, par exemple de recherche d'information (Kuzi et al., 2016), génération (Brown et al., 2020), et classification (Yang et al., 2015). Grâce aux mécanismes d'attention (Bahdanau et al., 2015; Luong et al., 2015), ces vecteurs sont contextualisés, i.e. un mot possède une représentation dépendante de sa position dans la phrase ainsi que de son contexte. Une nouvelle famille de modèles de langue (Devlin et al., 2019; Brown et al., 2020) basés sur les Transformers (Vaswani et al., 2017), qui incorporent ces mécanismes, a révolutionné le domaine en dépassant l'état de l'art sur un très grand nombre de tâches de traitement automatique des langues (traduction, génération, systèmes de questions-réponses).

De nombreuses problématiques peuvent bénéficier d'une représentation conjointe des documents et des auteurs, comme la recommandation automatique ou l'identification d'auteur. Ces représentations doivent ainsi vivre dans le même espace de façon à pouvoir calculer des