Vers un TAL plus vert et durable : entraîner un modèle neuronal de génération de texte avec peu de données et de calculs

Encadrement : florian.boudin@univ-nantes.fr
Lieu du TER : LS2N (Site de la FST)

Contexte

La quantité de calcul requise pour entraîner, évaluer et déployer des modèles de Traitement Automatique des Langues (TAL) par apprentissage profond a considérablement augmentée au cours des dernières années. Il en résulte une empreinte carbone étonnamment élevée pour les expériences en TAL, et un coût financier pour réaliser ces expériences de plus en plus lourd pour les universitaires, les étudiants et les chercheurs. Ce constat est la conséquence directe du comportement de la communauté TAL qui valorise principalement les résultats « état-de-l'art », par opposition à la comparaison de méthodes avec des quantités variables de données, de paramètres et de temps d'entraînement.

Objectifs

L'objectif de ce Travail d'Étude et de Recherche (TER) est de mener une série d'expériences visant à comparer les performances d'un modèle neuronal de génération automatique de mots-clés en faisant varier la quantité de données d'entraînement de quelques milliers à plusieurs millions d'exemples. Plus précisément, il s'agira d'étudier le comportement d'un modèle par apprentissage séquence-à-séquence [1], dont une implémentation est disponible dans la bibliothèque opennmt (https://opennmt.net/), sur différents ensembles de données de tailles croissantes. Au terme de ce travail, nous ambitionnons de répondre aux questions suivantes : quelles ressources minimales sont nécessaires pour entraîner un modèle de génération automatique de mots-clés ? quel est le coût environnemental de l'entraînement de tel modèle ? comment réduire ce coût sans baisser significativement les performances du modèle ?

Dans le cas où les résultats obtenus dans le cadre du TER sont convaincants, l'écriture et la soumission d'un article de recherche dans un atelier ou conférence scientifique sera envisagée.

Profil des candidat.e.s

Un/deux personnes sont attendues, avec pour chacune une appétence pour le TAL, la recherche d'information et/ou l'apprentissage automatique. Des connaissances en programmation Python et en langage de scripting sont un plus.

Bibliographie

1. Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, Yu Chi. Deep Keyphrase Generation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. https://aclanthology.org/P17-1054