Apprentissage automatique et TAL : la question du vocabulaire

Alexandre Allauzen (LIMSI)

LL exceptionnel du lundi 26 mars 2018

Ces dernières décennies, les modèles d'apprentissage automatique ont enrichi les perspectives de recherche en traitement automatique des langues. Néanmoins si la plupart des modèles sont considérés comme "universels" dans leur conception, la diversité des langues implique une réalité bien différente.

Par exemple, les définitions du mot et du vocabulaire sont cruciales pour de nombreuses applications (e.g. reconnaissance automatique de la parole, traduction). Selon les langues et leurs processus morphologiques, la dimension des vocabulaires et la notion de mots diffèrent grandement et altèrent la pertinence des modèles d'apprentissage considérés pourtant comme état de l'art.

Cet exposé aborde cette question avec deux applications : un modèle neuronal sans vocabulaire pour l'étiquetage morphosyntaxique de l'allemand ; approche bayésienne non-paramétrique pour la segmentation morphologique non-supervisée.