Apprentissage et évaluation de plongements lexicaux sur un corpus SNCF en langue spécialisée

Nicolas Dugué*, Nathalie Camelin*, Luce Lefeuvre**, Xining Li*, Coralie Reutenauer**, Cyndel Vaudapiviz**

*Le Mans Université, LIUM, EA 4023,
Laboratoire d'Informatique de l'Université du Mans
**SNCF INNOVATION & RECHERCHE
1-3, avenue Francois Mitterrand,93210 SAINT-DENIS

Résumé. Au sein du groupe SNCF, le programme PRISME d'excellence sécurité intègre une démarche de simplification de l'accès à l'information et de la production de contenus dans la documentation métier. Dans ce contexte, nous avons mis en œuvre des traitements sur un corpus de référentiels métiers SNCF afin de guider l'utilisateur dans sa recherche documentaire. Les travaux présentés visent à évaluer l'usage des plongements lexicaux pour générer des représentations sémantiques denses sur lesquelles se baseront des méthodes de deep learning pour structurer le corpus SNCF. Le protocole mis en place consiste en l'évaluation empirique des voisinages de mots par des experts. Dans cette étude, nous montrons les difficultés d'apprentissage et d'évaluation inhérentes à ce type de corpus avec de nombreux mots soit très spécifiques, soit polysémiques, rendant la construction d'un espace de représentations robuste difficile.

1 Introduction

Au sein du groupe SNCF, la documentation métier est aujourd'hui en pleine mutation, avec des métiers qui se digitalisent, plus mobiles et marqués par de nouveaux modes de consommation de l'information. Dans le cadre du programme PRISME de transformation en matière de sécurité ferroviaire, SNCF cherche à simplifier l'accès à l'information et la production de contenus dans la documentation métier. L'évaluation de nouveaux systèmes intelligents d'accès aux contenus, d'aide à l'interprétation et à la saisie participe à cette démarche.

Dans ce contexte, nous avons mis en œuvre des traitements sur un corpus de référentiels métiers SNCF pour répondre à deux objectifs. Le premier est celui de guider l'utilisateur dans sa recherche documentaire à travers la structuration des résultats de recherche, en regroupant en thématiques (classification *non supervisée*) les documents retournés en réponse à une requête. Le second objectif est d'aider les rédacteurs à qualifier automatiquement de nouveaux documents selon des thèmes définis dans une arborescence construite par les experts métiers (classification *supervisée*). Les enjeux dans notre contexte d'application industrielle sont de disposer de systèmes performants (réponse rapide et pertinente), maîtrisés (résultats interprétables et explicables) et adaptés aux besoins des utilisateurs de ces systèmes (pertinence des