Désambiguïsation des mots polysémiques de la ville dans des romans de science-fiction

Sami Guembour[1], Catherine Dominguès[2]

[1]LASTIG, Univ Gustave Eiffel, ENSG, IGN, France – sami.guembour@ign.fr

[2]LASTIG, Univ Gustave Eiffel, ENSG, IGN, France – catherine.domingues@ign.fr

Abstract

This work is part of a project on the representation of the city of the future in a corpus of science fiction novels. The city's words were identified using an existing terminology resource. Some of these words are polysemous or can be used in contexts other than urban description. This article proposes a method to disambiguate them. It uses the CamemBERT language model. Classifiers were fine-tuned to determine the employment context: urban/non-urban. The evaluations on two different corpora show very significant efficiency, which reflects a possibility of distinction in the use of these words.

Keywords: NLP, disambiguation, polysemy, CamemBERT, embedding vector, Fine-Tuning, classification, city, science fiction, corpus.

Résumé

Ce travail s'insère dans le contexte d'un projet sur la représentation de ville du futur dans un corpus de romans de science-fiction. Les mots de la ville ont été identifiés grâce à une ressource terminologique existante. Certains de ces mots sont polysémiques ou peuvent être utilisés dans d'autres contextes que la description urbaine. Cet article propose une méthode pour les désambiguïser. Elle utilise le modèle de langue CamemBERT. Des classifieurs sont affinés (fine-tuning) pour déterminer le contexte d'emploi : urbain/non urbain. Les évaluations sur deux corpus différents montrent une efficacité très importante, ce qui traduit une possibilité de distinction dans l'emploi de ces mots.

Mots clés : TAL, désambiguïsation, polysémie, CamemBERT, vecteur de plongement, affinement, classification, ville, science-fiction, corpus.