Évaluation des propriétés multilingues d'un embedding contextualisé

Félix Gaschi*,**, Alexandre Joutard**, Parisa Rastin*, Yannick Toussaint*

*LORIA, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy
{nom}.{prenom}@loria.fr,
http://www.loria.fr
**Posos, 55 rue de la Boétie, 75008 Paris
{prenom}@posos.fr
http://posos.co

Résumé. Les modèles d'apprentissage profond comme BERT, un empilement de couches d'attention avec un pré-entraînement non supervisé sur de larges corpus, sont devenus la norme en NLP. mBERT, une version pré-entraînée sur des corpus monolingues dans 104 langues, est ensuite capable d'apprendre une tâche dans une langue et de la généraliser à une autre. Cette capacité de généralisation ouvre la perspective de modèles efficaces dans des langues avec peu de données annotées, mais reste encore largement inexpliquée. Nous proposons une nouvelle méthode fondée sur des mots traduits en contexte plutôt que des phrases pour analyser plus finement la similarité de représentations contextualisées à travers les langues. Nous montrons que les représentations de différentes langues apprises par mBERT sont plus proches pour des couches profondes, et dépassent les modèles spécifiquement entraînés pour être alignés.

1 Introduction

Construire un embedding dans un contexte multilingue est un défi à part entière. Il s'agit de faire en sorte que deux mots sémantiquement similaires aient des représentations proches, qu'ils soient d'une même langue ou non. En fouille de données, un tel embedding permettrait de faire concorder les représentations d'une requête dans une langue et d'un document écrit dans une autre (Artetxe et al., 2019).

Différentes techniques existent pour créer des représentations multilingues (Søgaard et al., 2019). Nous nous intéressons à des embeddings multilingues contextualisés construits par mBERT. "Multilingual BERT" (mBERT) est un modèle de langage (Devlin et al., 2018) pré-entraîné sur un corpus de 104 langues. Il produit un embedding contextualisé multilingue qui semble porter des représentations alignées des différentes langues sans y avoir été explicitement entraîné. En effet, mBERT serait capable de généraliser une tâche apprise dans une langue à une autre langue avec la démarche schématisée en Figure 1. mBERT (Fig. 1, gauche) a été pré-entraîné de manière non supervisée sur deux tâches : (1) la prédiction de mots masqués aléatoirement dans un corpus multilingue et (2) une classification binaire pour déterminer si