BERTEPro : Une nouvelle approche de représentation sémantique dans le domaine de l'éducation et de la formation professionnelle

Guillaume Lefebvre*,**, Haytham Elghazel*, Théodore Guillet**, Alexandre Aussem*, Matthieu Sonnati**

*Université Lyon 1, LIRIS, UMR CNRS 5205, F-69622
prenom.nom@liris.cnrs.fr
**Inokufu, France,
https://www.inokufu.com/
prenom.nom@inokufu.com

Résumé. FlauBERT et CamemBERT ont établi une nouvelle performance de pointe pour la compréhension de la langue française. Récemment, SBERT a transformé l'utilisation de BERT, afin de réduire l'effort de calcul des encastrements de phrases, tout en maintenant la précision de BERT. Cependant, ces modèles ont été entraînés sur des textes non spécifiques de la langue française, ce qui ne permet pas une représentation fine des textes de domaines spécifiques, comme le domaine de l'éducation et de la formation professionnelle. Dans cet article, nous présentons BERTEPro, un modèle basé sur FlauBERT, dont l'apprentissage a été étendu sur des textes du domaine de l'éducation et de la formation professionnelle, avant d'être affiné sur des tâches NLI et STS. L'évaluation des performances de BERTEPro sur des tâches STS, ainsi que sur des tâches de classification, ont confirmé que la méthodologie proposée bénéficie d'avantages significatifs par rapport aux autres méthodes de l'état de l'art.

1 Introduction

Le traitement du langage naturel (NLP) (Ranjan et al., 2016) est un domaine de l'apprentissage automatique visant à permettre aux machines d'interpréter et de traiter le langage humain tel qu'il est écrit ou parlé. Contrairement aux langages de programmation dont la syntaxe est formelle et sans ambiguïté, le langage naturel a une structure très variée et la signification d'un mot dépend fortement de son contexte. Pour la recherche de similarité dans un corpus, le problème est donc de pouvoir comparer des textes en tenant compte des subtilités de la langue telles que les *synonymes*, les *ambiguïtés* ou la *syntaxe*. Afin d'utiliser les données en apprentissage automatique, il est nécessaire de les représenter par une abstraction mathématique (vectorisation).

Au milieu du 20ème siècle, l'une des méthodes les plus utilisées, encore utilisée aujourd'hui dans de nombreux moteurs de recherche, est née : TF-IDF (Jones, 1972). Il s'agit d'une méthode de pondération souvent utilisée en recherche d'information et surtout en fouille de textes.