Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Quand être absent de mBERT n'est que le commencement : Gérer de nouvelles langues à l'aide de modèles de langues multilingues

Benjamin Muller[1, 3] Antonios Anastasopoulos[2] Benoît Sagot[1] Djamé Seddah[1]
(1) Inria, Paris, France
(2) Department of Computer Science, George Mason University, USA
(3) Sorbonne Université, Paris, France
firstname.lastname@inria.fr antonis@gmu.edu

RÉSUMÉ

L'apprentissage par transfert basé sur le pré-entraînement de modèles de langue sur une grande quantité de données brutes est devenu la norme pour obtenir des performances état de l'art en TAL. Cependant, la façon dont cette approche devrait être appliquée pour des langues inconnues, qui ne sont couvertes par aucun modèle de langue multilingue à grande échelle et pour lesquelles seule une petite quantité de données brutes est le plus souvent disponible, n'est pas claire. Dans ce travail, en comparant des modèles multilingues et monolingues, nous montrons que de tels modèles se comportent de multiples façons sur des langues inconnues. Certaines langues bénéficient grandement de l'apprentissage par transfert et se comportent de manière similaire à des langues proches riches en ressource, alors que ce n'est manifestement pas le cas pour d'autres. En nous concentrant sur ces dernières, nous montrons dans ce travail que cet échec du transfert est largement lié à l'impact du script que ces langues utilisent. Nous montrons que la translittération de ces langues améliore considérablement le potentiel des larges modèles de langue neuronaux multilingues pour des tâches en aval. Ce résultat indique une piste prometteuse pour rendre ces modèles massivement multilingues utiles pour de nouveaux ensembles de langues absentes des données d'entraînement.

ABSTRACT

When Being Unseen from mBERT is just the Beginning : Handling New Languages With Multilingual Language Models

Transfer learning based on pretraining language models on a large amount of raw data has become a new norm to reach state-of-the-art performance in NLP. Still, it remains unclear how this approach should be applied for unseen languages that are not covered by any available large-scale multilingual language model and for which only a small amount of raw data is generally available. In this work, by comparing multilingual and monolingual models, we show that such models behave in multiple ways on unseen languages. Some languages greatly benefit from transfer learning and behave similarly to closely related high resource languages whereas others apparently do not. Focusing on the latter, we show that this failure to transfer is largely related to the impact of the script used to write such languages. We show that transliterating those languages significantly improves the potential of large-scale multilingual language models on downstream tasks. This result provides a promising direction towards making these massively multilingual models useful for a new set of unseen languages. [1]

MOTS-CLÉS : modèles de langues multilingues neuronaux, langues peu dotées, translittération,

1. Code available at https://github.com/benjamin-mlr/mbert-unseen-languages.git

Apprentissage par transfert.

KEYWORDS: Neural multilingual language models, under-resourced languages, translitterating, Transfer-based learning.

Travail accepté à NAACL-HLT 2021 (https://aclanthology.org/2021.naacl-main.38.pdf)