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

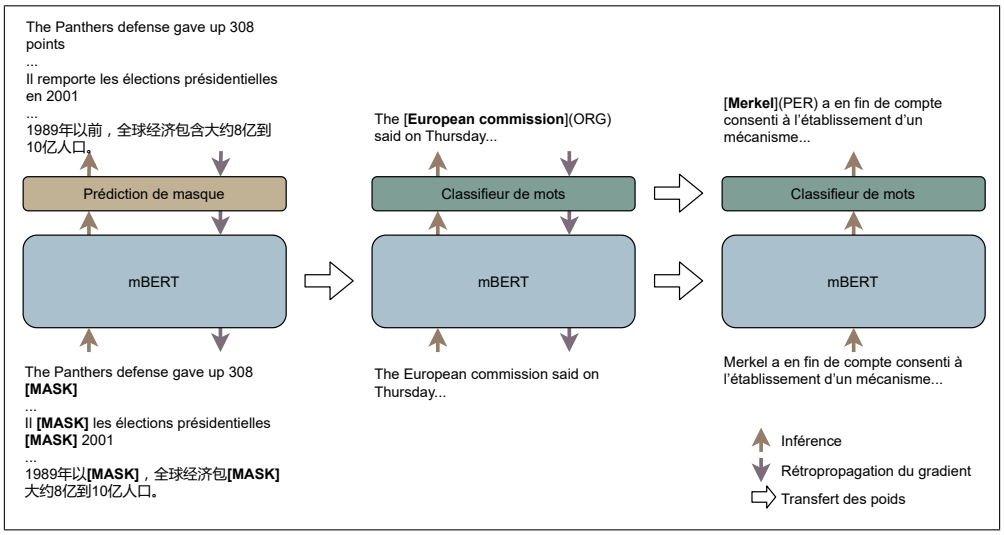


FIG. 1: Apprentissage par transfert multilingue réalisé par Pires et al. (2019).

deux phrases se suivent. Pires et al. (2019) montrent que mBERT peut être affiné (fine-tuned) sur une tâche supervisée en anglais (Fig. 1, centre), et évalué sur cette même tâche dans une autre langue (Fig. 1, droite). Cette démarche s'appelle l'apprentissage par transfert multilingue et donne de bons résultats alors que le modèle n'a jamais été explicitement encouragé à aligner ses représentations des différentes langues.

Nous cherchons à évaluer la qualité de la représentation multilingue, non pas au travers d'une évaluation indirecte du modèle via une évaluation sur une tâche donné, mais plus directement en comparant des représentations de mots en contexte. Notre contribution est double : (1) nous proposons une méthode de construction d'un corpus de paires de mots traduits en contexte, à l'aide d'un ensemble de phrases traduites et d'un dictionnaire bilingue ; (2) nous utilisons ce jeu de données pour montrer que les représentations construites par mBERT des

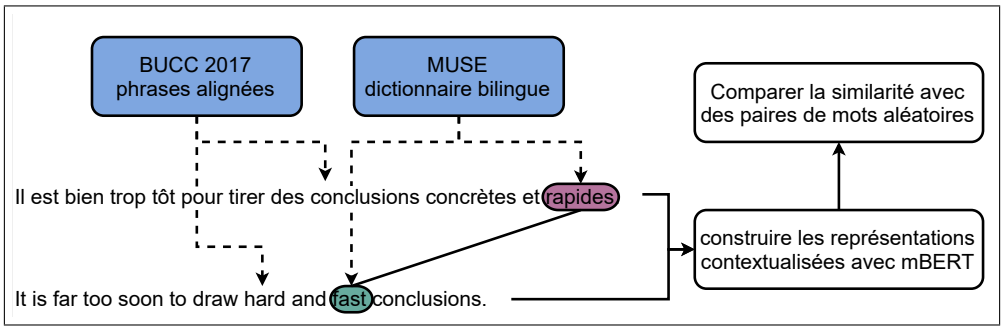


FIG. 2: Notre méthode extrait des paires de mots traduits en contexte à l'aide de phrases traduites et d'un dictionaire bilingue.

mots en contexte traduits dans différentes langues sont suffisamment proches pour pouvoir être distinguées des représentations d'autres mots.

2 Travaux connexes

Des travaux ont montré que mBERT possède des capacités de généralisation multilingues suprenantes (Pires et al., 2019; Wu et Dredze, 2019), mais ne démontrent pas que les représentations des mots des différentes langues qu'il construit sont alignées. Wu et Dredze (2019) montrent au contraire que ces embeddings pourraient comporter des composantes spécifiques à la langue. Des travaux se sont penchés sur la similarité de représentations de phrases traduites (Pires et al., 2019; Singh et al., 2019) avec des conclusions différentes. Zhao et al. (2020) se sont penchés, comme nous, sur la similarité de mots traduits en contexte, mais observent, contrairement à nous, que la similarité de paires de mots traduits est souvent proche de celle de paires aléatoires.

D'autres modèles similaires à mBERT existent. XLM (Lample et Conneau, 2019) utilise un corpus parallèle pour son pré-entraînement et mBART (Liu et al., 2020) identifie chaque langue par un symbole spécifique. Nous nous focalisons sur mBERT qui n'utilise pas de corpus parallèle et n'a pas connaissance du langage du texte en entrée.

3 Méthode

La méthode proposée (voir Fig. 2) consiste à extraire, dans des phrases traduites, des paires de mots issues d'un dictionnaire bilingue et à comparer la similarité des représentations contextualisées construites par mBERT pour ces paires avec des paires extraites aléatoirement.

3.1 Construction des paires de mots

Pour construire les paires de mots traduits en contexte, notre méthode s'appuie sur : (1) un jeu de données contenant des paires de phrases traduites : BUCC 2017 ; (2) des dictionnaires bilingues : MUSE (Conneau et al., 2017).

Les dictionnaires bilingues contiennent des paires de mots traduits. Dans chaque paire de phrases issue du corpus de traduction, les paires de mots issues du dictionnaire présentes dans les phrases sont extraites pour obtenir des paires de mots traduits et accompagnés de leur contexte. On obtient ainsi une forme de contextualisation du dictionnaire bilingue. Les représentations contextualisées de mots des différentes langues pourront être directement comparées, au lieu de construire des représentations des phrases. Pour éviter toute ambiguité, une paire de mots alignés est retenue uniquement s'il n'y a qu'un seul candidat à la traduction de chaque mot de la paire.

Les paires de phrases traduites sont extraites de BUCC 2017, qui n'est pas un corpus de traduction à proprement parler, mais un corpus de fouille de bi-textes. Il comporte une liste de phrases pour chaque langue dont il faut extraire les paires de phrases qui sont des traductions. La différence avec un corpus parallèle est que certaines phrases n'ont pas de traduction, ce qui permet de disposer d'un ensemble de phrases plus large pour extraire des paires aléatoires.

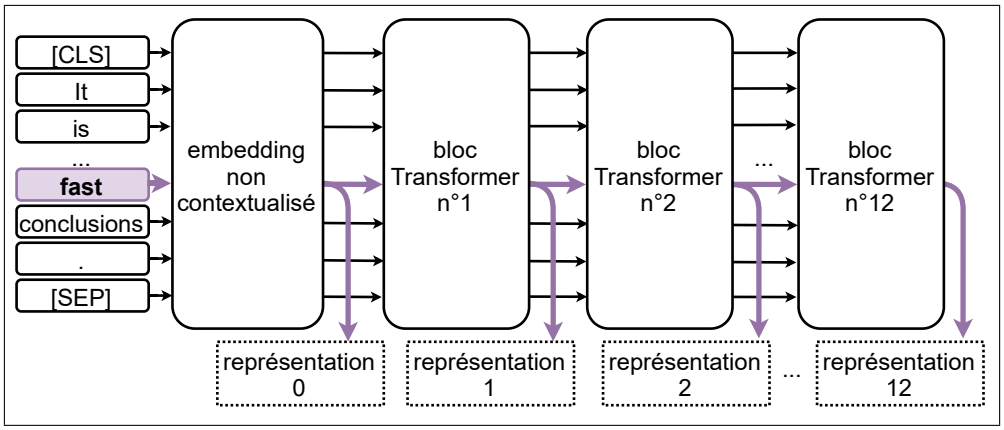


FIG. 3: Schéma de l'extraction des représentations du mot "fast" construites par mBERT.

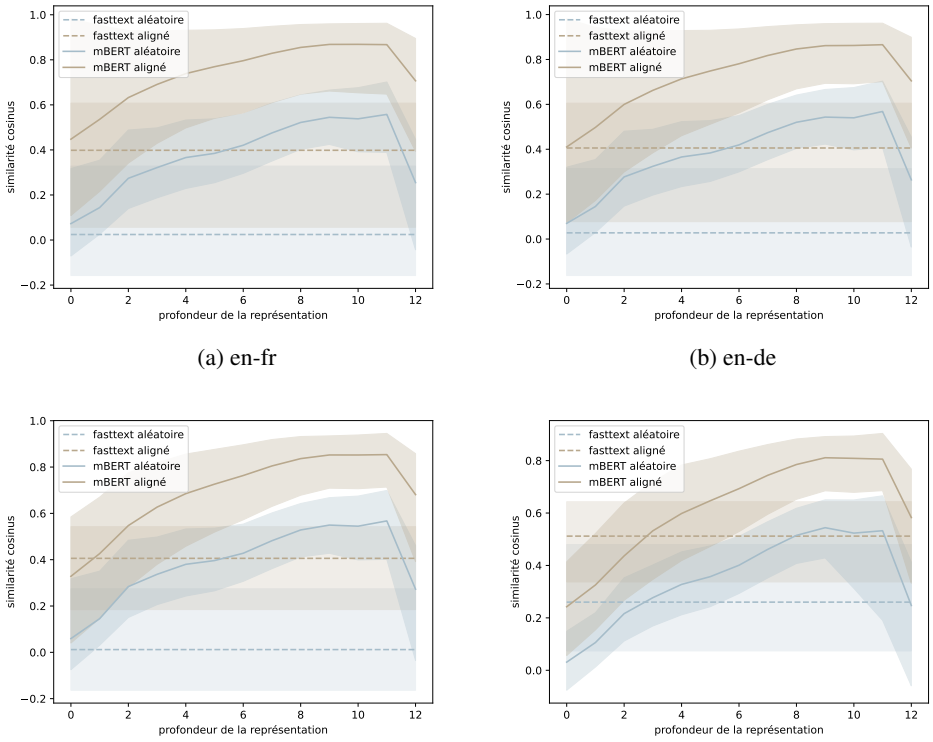
Le corpus de fouille de bi-textes BUCC 2017 permet d'extraire des paires de phrases pour quatre paires de langues : anglais-français, anglais-allemand, anglais-russe et anglais-mandarin. Le jeu de données contient 9 086 (en-fr), 9 580 (en-de), 14 435 (en-ru) et 1 899 (en-zh) paires de phrases traduites. Avec notre méthode d'extraction des paires de mots en contexte, nous obtenons pour chaque paires de langues, les nombres de paires de mots suivants : 57 826 (en-fr), 61 802 (en-de), 57 116 (en-ru) et 3 550 (en-zh).

3.2 Construction des représentations contextualisées

Une fois les paires de mots contextualisés extraites, il faut calculer la similarité de leurs représentations par mBERT et la comparer avec celles de paires de mots aléatoirement extraites du corpus de traduction.

Le modèle BERT (Devlin et al., 2018), et sa variante multilingue mBERT, peuvent donner lieu à différentes représentations des mots. Comme indiqué Fig. 3, l'entrée du modèle est un texte subdivisé en sous-mots et entouré de deux symboles artificiels [CLS] et [SEP]. mBERT associe ensuite à chaque sous-mot un embedding non-contextualisé.

Les blocs Transformer prennent une séquence de représentations en entrée et produisent une autre séquence de représentations en sortie. Chaque bloc Transformer contient la même architecture décrite par Vaswani et al. (2017) : une couche d'auto-attention suivie d'un perceptron multi-couches avec des connexions résiduelles et normalisation. mBERT comporte 12 blocs Transformer dont la sortie de chacun constitue une représentation étudiée dans nos expériences, en plus de la sortie de l'embedding non-contextualisé. Dans le cas où un mot d'une paire est constitué de plusieurs sous-mots, on considère la moyenne des représentations des sous-mots comme représentation du mot.



(a) en-fr (b) en-de

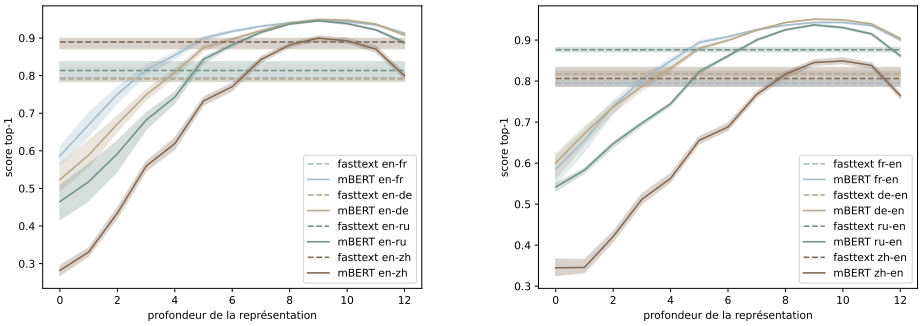
(c) en-ru (d) en-zh

FIG. 4: Similarité entre des paires de mots traduits en contexte (marron) et des paires aléatoires (bleu) pour chaque représentation construite par mBERT (trait plein) et pour la représentation construite par des embedding FastText alignés (pointillés), avec intervalle de confiance à 95%.

4 Expériences et résultats

Dans nos expériences, les représentations construites par mBERT sont comparées avec une autre représentation : des embeddings de mots FastText (Bojanowski et al., 2016) de différentes langues, non-contextualisés, alignés avec une méthode non-supervisée obtenant de bons résultats en traduction mot-à-mot (Joulin et al., 2018).

La similarité cosinus est mesurée pour chacune des paires de mots traduits en contexte ainsi que pour un nombre égal de paires aléatoires (voir Fig. 4). Son évolution en fonction de la profondeur (en trait plein) est comparée à celle de la similarité des représentations par FastText aligné (en pointillé). L'intervalle de confiance à 95% des similarités des paires traduites chevauche peu voire pas du tout celui des paires aléatoires dans certaines couches de mBERT. Ce chevauchement est aussi moins fort que pour FastText aligné. Cependant, il semble difficile d'affirmer que les représentations de mBERT sont mieux alignées. C'est ce qui nous conduit à mener l'expérience qui suit.



(a) Anglais vers une autre langue.

(b) Une autre langue vers l'anglais.

FIG. 5: Précision top-1 de la recherche par plus proche voisin.

Nous réalisons une deuxième expérience de recherche par plus proche voisin de la traduction d'un mot en contexte. Pour chaque langue disponible, 10 000 mots contextualisés sont tirés au hasard dans le corpus. Pour chaque paire de mots traduits en contexte, nous regardons si la représentation de l'un des mots de la paire (dans la langue cible) est plus proche de l'autre (dans le langue source) que celles des 10 000 autres mots contextualisés tirés au hasard dans la même langue cible. Il s'agit finalement d'une sorte de tâche d'induction de dictionnaire pour des mots en contexte. Le critère de similarité utilisé est celui pour lequel a été optimisé l'alignement des embeddings FastText : le critère CSLS (Joulin et al., 2018). Il s'agit d'une similarité cosinus modifiée qui prend en compte la densité autour des représentations. On mesure alors la précision top-1 de ce critère de recherche parmi les 10 000 mots pour chaque paire de mots et pour chaque paire de langue. Nos résultats sont présentés Fig. 5. À droite la langue cible est l'anglais, à gauche c'est l'autre langue de la paire (fr, de, ru ou zh).

Nous observons que l'alignement au sein de mBERT semble meilleur autour de la dixième couche que des premières ou dernières. Autour de cette couche profonde seulement, l'alignement produit par mBERT semble meilleur que celui de FastText aligné. L'alignement semble moins bon sur les trois dernières couches. Lors du pré-entraînement, mBERT cherche à prédire un mot masqué dans une langue donnée. Il faut donc que les représentations des différentes langues soient différenciées pour ne pas remplacer un mot masqué par sa traduction. Il semblerait que la tâche de prédiction de masque, censée être cantonnée à la couche de prédiction de masque, a en réalité contaminé les dernières couches de mBERT.

Zhao et al. (2020) ont effectué une expérience similaire à notre première expérience. Mais ils observent un séparation moins nette entre les distributions des paires traduites et aléatoires. Leur méthode diffère dans la façon d'extraire les paires de mots traduits en contexte des paires de phrases. Là où nous utilisons des dictionnaires bilingues, ils s'appuient sur un outil probabiliste d'alignement de phrases : FastAlign (Dyer et al., 2013). Ils obtiennent plus de paires, mais certaines sont de moins bonne qualité. Leur expérience est reproduite Fig. 6, avec l'histogramme de la similarité des paires produites par FastAlign (marron foncé), celui des paires aléatoires (vert), celui des paires que nous avons construites (bleu), mais aussi celui de paires aléatoires extraites de paires de phrases traduites (marron clair). Comme observé par Zhao

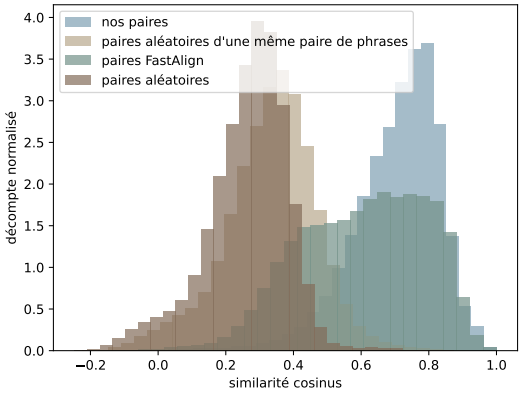


FIG. 6: Similarités en sortie de mBERT pour divers types de paires de mots (en-de).

et al. (2020), les distributions des similarités des paires FastAlign et aléatoires se recouvrent beaucoup, ce qui s'explique bien par le fait que des paires aléatoires extraites dans des paires de phrases (marron clair) ont une similarité comparable à celles de paires tirées dans tout le corpus (marron foncé). Les paires obtenues avec notre méthode (bleu) se distinguent plus clairement.

5 Conclusion

Nous proposons une manière d'extraire des paires de mots traduits en contexte qui permet une analyse plus fine de l'alignement des représentations des différentes langues qu'une représentation des phrases et plus exacte que des paires de mots construites par des modèles probabilistes comme FastAlign. Nous montrons que deux mots de langues différentes, accompagnés chacun de leur contexte, et ayant des significations proches, ont des représentations proches, alors que mBERT n'a jamais été spécifiquement entraîné pour cet alignement. À l'échelle du mot, les représentations des paires sont mieux alignées pour certaines couches que des embeddings non-contextualisés spécifiquement construits pour être alignés.

Des modèles comme mBERT sont donc un bon point de départ pour créer des représentations multilingues cohérentes. Mais il faut tout de même prendre garde à l'information spécifique aux langues qui semble spécialement prendre de l'importance dans les dernières couches du modèle. Notre hypothèse pour expliquer cette diminution de la qualité de l'alignement sur les dernières couches est celle d'une contamination du modèle de langage par l'objectif de prédiction de mot masqué. De futures recherches pourraient essayer d'améliorer l'alignement en modifiant l'objectif de prédiction de mot masqué pour qu'il ne soit pas dépendant de la langue.

Références

Artetxe, M., S. Ruder, et D. Yogatama (2019). On the cross-lingual transferability of monolingual representations. *CoRR abs/1910.11856*.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2016). Enriching word vectors with subword information. *CoRR abs/1607.04606*.

Conneau, A., G. Lample, M. Ranzato, L. Denoyer, et H. Jégou (2017). Word translation without parallel data. *arXiv preprint arXiv :1710.04087*.

Devlin, J., M. Chang, K. Lee, et K. Toutanova (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Dyer, C., V. Chahuneau, et N. A. Smith (2013). A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Joulin, A., P. Bojanowski, T. Mikolov, et E. Grave (2018). Improving supervised bilingual mapping of word embeddings. *CoRR abs/1804.07745*.

Lample, G. et A. Conneau (2019). Cross-lingual language model pretraining. *CoRR abs/1901.07291*.

Liu, Y., J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, et L. Zettlemoyer (2020). Multilingual denoising pre-training for neural machine translation. *CoRR abs/2001.08210*.

Pires, T., E. Schlinger, et D. Garrette (2019). How multilingual is multilingual bert ? *CoRR abs/1906.01502*.

Singh, J., B. McCann, R. Socher, et C. Xiong (2019). Bert is not an interlingua and the bias of tokenization. In *EMNLP*.

Søgaard, A., I. Vulić, S. Ruder, et M. Faruqui (2019). Cross-lingual word embeddings. *Synthesis Lectures on Human Language Technologies 12*, 1–132.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, et I. Polosukhin (2017). Attention is all you need. *CoRR abs/1706.03762*.

Wu, S. et M. Dredze (2019). Beto, bentz, becas : The surprising cross-lingual effectiveness of BERT. *CoRR abs/1904.09077*.

Zhao, W., S. Eger, J. Bjerva, et I. Augenstein (2020). Inducing language-agnostic multilingual representations. *CoRR abs/2008.09112*.

Summary

Deep learning models like BERT, a stack of attention layers with an unsupervised pretraining on large corpora, have become the norm in NLP. mBERT, a multilingual version of BERT, is capable of learning a task in one language and of generalizing it to another. This generalization ability opens the perspective of having efficient models in languages with few annotated data, but remains still largely unexplained. We propose a new method based on in-context translated words rather than translated Sentences in order to analyze the similarity between contextualized representations across languages. We show that the representations learned by mBERT are closer for deep layers, outperforming other representations that were specifically trained to be aligned.