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

Une étude statistique des plongements dans les modèles *transformers* pour le français

Loïc Fosse Duc Hau Nguyen Pascale Sébillot Guillaume Gravier
Univ Rennes, Inria, CNRS, IRISA
prenom.nom@irisa.fr

RÉSUMÉ

Nous étudions les propriétés statistiques des plongements dans les modèles *transformers* pour le français. Nous nous appuyons sur une analyse de la variance, des similarités cosinus intra-phrase et du rang effectif des plongements aux différents niveaux d'un *transformer*, pour des modèles pré-entraînés et des modèles adaptés à la classification de textes. Nous montrons que les modèles FlauBERT et CamemBERT pré-entraînés ont des comportements très différents même si les deux ont une tendance à générer des représentations anisotropiques, c'est-à-dire se concentrant dans un cône au sein de l'espace des plongements, comme observé pour l'anglais. L'adaptation à la classification de textes modifie le comportement des modèles, notamment dans les dernières couches, et procure une tendance forte à l'alignement des plongements, réduisant également la dimension effective de l'espace au final. Nous mettons également en évidence un lien entre convergence des plongements au sein d'une phrase et classification de texte, lien dont la nature reste difficile à appréhender.

ABSTRACT

An empirical statistical study of embeddings in French transformers

We investigate statistical properties of contextual word embeddings in transformer models for the French language. We rely on an analysis of variance, intra-phrase cosine similarities and effective rank of embbedings at different levels of pre-trained models and models adapted to document classification. We show that the pre-trained FlauBERT and CamemBERT models have very different behaviors even if both have a tendency to generate anisotropic representations, i.e., concentrating representations in a cone within the embedded space, as observed for English. Adaptation to text classification modifies the behavior of the models, especially in the last layers, and provides a strong tendency to align the embeddings, also reducing the effective dimension of the space in the end. We also evidence a relation between the convergence of the embeddings within a sentence and text classification accuracy, yet not fully understood.

MOTS-CLÉS : plongements de mots, *transfomer*, FlauBERT, CamemBERT, analyse statistique.

KEYWORDS: word embedding, transformer, FlauBERT, CamemBERT, statistical analysis.

1 Introduction

Les modèles de type *transformers*, pré-entraînés comme modèles de langage (Devlin *et al.*, 2019; Radford *et al.*, 2019; Martin *et al.*, 2020; Le *et al.*, 2020), sont au cœur du succès de nombreuses approches récentes en traitement automatique des langues. Le principe général de ces modèles consiste à transformer progressivement la représentation des mots (ou plus exactement, des *tokens*) en entrée

Volume 1 : conférence principale, pages 247–256.

de manière à en donner en sortie une représentation contextualisée sous la forme d'un plongement dans un espace de grande dimension. La transformation des plongements se fait au travers de couches d'auto-attention, qui prennent en entrée une séquence de plongements pour en produire une nouvelle en s'appuyant sur les liens que les plongements en entrée entretiennent entre eux.

Une des raisons du succès des modèles *transformers* réside notamment dans l'existence de modèles pré-entraînés sur une tâche générique de modélisation du langage, dans laquelle des *tokens* masqués en entrée doivent être prédits à partir de leur contexte. Plusieurs variantes de ces modèles ont été proposées, avec deux grandes familles : des modèles causaux, dans lesquels seul le contexte gauche d'un *token* est pris en compte, l'apprentissage se fondant sur la prédiction d'un mot en fonction de son contexte gauche (Radford *et al.*, 2019) ; des modèles non causaux qui considèrent les contextes gauche et droit, pour lesquels l'apprentissage repose sur le masquage de *tokens* (Devlin *et al.*, 2019; Martin *et al.*, 2020; Le *et al.*, 2020). En pratique, de nombreuses tâches en traitement automatique de langues sont aujourd'hui résolues en partant des ces modèles génériques pré-entraînés et en y ajoutant des couches de classification qui prennent en entrée les plongements contextualisés par le *transformer*. La couche de classification ainsi que les paramètres du *transformer* sont adaptés à la tâche visée par quelques époques d'apprentissage sur des données d'adaptation.

Nous nous intéressons dans cet article aux propriétés statistiques des plongements générés par les différentes couches de transformation pour deux modèles entraînés pour le français, que ce soit au niveau des modèles pré-entraînés ou dans des modèles adaptés à une tâche de classification de textes.

De nombreux travaux se sont intéressés aux informations linguistiques portées par les plongements contextuels issus des modèles *transformers* pré-entraînés, notamment (Kobayashi *et al.*, 2020; Van Aken *et al.*, 2019; Rogers *et al.*, 2020; Htut *et al.*, 2019; Ethayarajh, 2019; Jawahar *et al.*, 2019). En particulier, Rogers *et al.* (2020) donne un aperçu assez complet des différentes études menées à ce sujet et des principales conclusions qui en découlent. Pour la plupart, ces études s'appuient sur l'utilisation de classifieurs prenant en entrée les plongements contextuels (désignés comme *sondes*), l'hypothèse sous-jacente étant qu'une sonde arrive à une performance satisfaisante si les plongements portent l'information nécessaire. Dans le contexte de cet article, les enseignements les plus pertinents sont que l'ordre des *tokens* en entrée perd de l'importance à partir de la quatrième couche (Lin *et al.*, 2019), que les informations syntaxiques sont principalement représentées dans les couches du milieu, la dernière couche étant typiquement très dépendante de la tâche (Kovaleva *et al.*, 2019).

Peu de ces études portent cependant sur la géométrie et les propriétés statistiques des plongements – *cf*. Sec. 4 dans Rogers *et al.* (2020) – à l'exception notable des travaux de Ethayarajh (2019) et Hernandez & Andreas (2021) qui comparent d'un point de vue géométrique plusieurs modèles. En particulier, nous nous inspirons de la comparaison de ELMO (Peters *et al.*, 2018), BERT (Devlin *et al.*, 2019) et GPT-2 (Radford *et al.*, 2019) dans Ethayarajh (2019), qui met en évidence deux faits importants. D'une part, l'anisotropie des plongements, c'est-à-dire leur concentration dans une direction, augmente au travers des couches d'attention, se concentrant progressivement sur un cône étroit. D'autre part, l'influence du contexte croît au travers des couches, éloignant progressivement les plongements d'un même *token* apparaissant dans des contextes différents. Ces observations sont corroborées par Jawahar *et al.* (2019) dont la figure 1 montre une convergence des représentations des différents types de *spans* et servent de point de départ à notre étude.

Pour notre part, nous étudions ici les propriétés statistiques des plongements après les différentes couches pour mieux comprendre ces représentations et le fonctionnement des modèles pour le français qui, à notre connaissance, n'a pas été étudié à ce jour. Nous nous intéressons plus particulièrement à la convergence des représentations lorsque l'on monte en abstraction aux travers des différentes

couches d'attention, convergence absolue et/ou en direction, élargissant ainsi l'étude de Ethayarajh (2019) à de nouvelles mesures. Nous étudions cette convergence sur les modèles FlauBERT et CamemBERT pré-entraînés et regardons l'impact de l'adaptation à une tâche de classification de textes sur la géométrie des plongements. Bien qu'il n'y ait *a priori* pas de raison fondamentale pour un comportement différent des modèles français par rapport aux modèles anglais, nous verrons que les modèles FlauBERT et CamemBERT pré-entraînés ne montrent étrangement pas les mêmes tendances en dépit de leur forte similitude.

2 Méthodologie

Dans notre étude, nous comparons deux modèles pour le français, FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2020), à partir des modèles de la librairie HuggingFace `flaubert_bert_cased` et `camembert_base` respectivement. Ces deux modèles présentent des architectures proches et génèrent tous deux des plongements de dimension $d = 768$ contextualisés au travers de 12 couches d'attention comportant chacune 12 têtes. FlauBERT s'appuie sur des couches d'attention classiques (Vaswani *et al.*, 2017), avec une couche de plongement initial pour un vocabulaire d'environ 69 000 *tokens*. CamemBERT utilise des couches de transformation de type Roberta (Liu *et al.*, 2019) avec un plongement initial sur un vocabulaire plus restreint d'environ 32 000 *tokens*. Les deux modèles ont peu ou prou le même nombre de paramètres (110M pour CamemBERT, 140M pour FlauBERT) et sont entraînés sur une tâche de modèle de langage masqué. Au-delà de l'architecture, FlauBERT est entraîné sur moins de données que CamemBERT (71 GB *vs*. 138 GB) mais sur des données plus contrôlées, ce qui résulte en pratique en des modèles complémentaires comme le souligne Le *et al.* (2020).

L'adaptation de ces deux modèles à une tâche de classification en polarité se fait sur le *corpus* CLS du *benchmark* FLUE (Le *et al.*, 2020). Les modèles initiaux sont complétés avec une tête de classification sur le *token* CLS de manière à prédire la polarité d'un document : l'ensemble des paramètres est réestimé sur 5 000 exemples [1] avec une ou deux époques. Nous obtenons sur un échantillon de 1 000 phrases de test une précision de 92.8 % pour FlauBERT et 94 % pour CamemBERT après une époque, sans variation notable après la seconde époque.

Notons $\mathbf{e}_i^k(s) = \{e_{ij}^k(s), \quad j \in [1, d]\}$ le plongement contextuel du i-ème *token* d'une phrase s au niveau k du modèle *transformer*. Le niveau $k = 0$ correspond aux plongements non contextuels tandis que $k = 12$ correspond aux plongements finaux.

Pour caractériser la variance des plongements, on s'intéresse avant tout à la variance pour une dimension j donnée par rapport à l'ensemble des *tokens* i d'une phrase, $v_j^k(s) = \text{var}(e_{ij}^k(s))$. L'intérêt de cette mesure est de mettre en évidence une éventuelle convergence des plongements des *tokens* d'une même phrase qui se traduirait alors par une diminution de la variance lorsque le niveau d'abstraction k augmente. Pour caractériser un niveau k donné de manière globale, on s'intéressera à la trace de la matrice de covariance, soit $\sum_j v_j^k(s)$.

La variance regarde les dimensions de manière indépendante. Une autre manière d'étudier une éventuelle convergence des représentations des *tokens* d'une phrase consiste à étudier leur colinéarité.

1. Pour des raisons d'efficacité de calcul, nous n'avons pas pris l'intégralité des données du corpus FLUE/CLS mais 5 000 échantillons (sur 6 000) après mélange aléatoire. Les expériences restent reproductibles l'amorce du tirage aléatoire étant fixé – *cf*. fin de section pour plus de détails.

Pour ce faire, nous nous intéressons à la distribution des similarités cosinus entre les plongements de deux *tokens* d'une même phrase s, soit

$$c_{ij}^k(s) = \frac{\mathbf{e}_i^k(s) \cdot \mathbf{e}_j^k(s)}{\|\mathbf{e}_i^k(s)\| \|\mathbf{e}_j^k(s)\|} \qquad i \in [1, N_s], \quad j > i \ , \tag{1}$$

où N_s désigne le nombre de *tokens* dans la phrase s. Comme pour la variance, cette distribution devrait tendre vers une distribution resserrée autour d'une moyenne proche de 1 en cas de convergence des représentations au sein d'une phrase.

Tant la variance que la distribution des similarités se calculent au niveau d'une phrase. En pratique, nous mesurons ces quantités sur 50 ou 100 phrases issues des données de test de FLUE/CLS et rapportons la distribution de la métrique.

Nous regardons enfin le rang effectif des plongements (Torregrossa *et al.*, 2021) que nous adaptons aux plongements contextuels. Le rang effectif (Roy & Vetterli, 2007) se fonde sur les valeurs singulières de la matrice des plongements pour donner une indication sur le nombre de dimensions utiles en s'appuyant sur le nombre de valeurs singulières qui ont un impact significatif dans la décomposition. Pour une matrice de plongements donnée W, le rang effectif est une valeur continue dans $[1, d]$, une valeur de d signifiant que toutes les dimensions sont utiles. Dans notre cas, la matrice W correspond à un ensemble de plongements contextuels pour les *tokens* des 10 premières phrases des données de test de FLUE/CLS afin de limiter la taille et permettre la décomposition en valeurs singulières.

L'ensemble des expériences est entièrement reproductible à partir des codes mis à disposition [2]. En particulier, pour la sélection des données d'apprentissage dans les modèles adaptés, l'amorce du tirage aléatoire est fixée de manière à générer systématiquement la même sélection. Pour les 50 ou 100 phrases utilisées pour mesurer la variance et la distribution des similarités, nous avons simplement pris les 50 ou 100 premières phrases dont la distribution des étiquettes est représentative de celle du corpus. Des expériences préliminaires ont montrés que le choix des phrases n'a qu'une influence marginale sur les résultats.

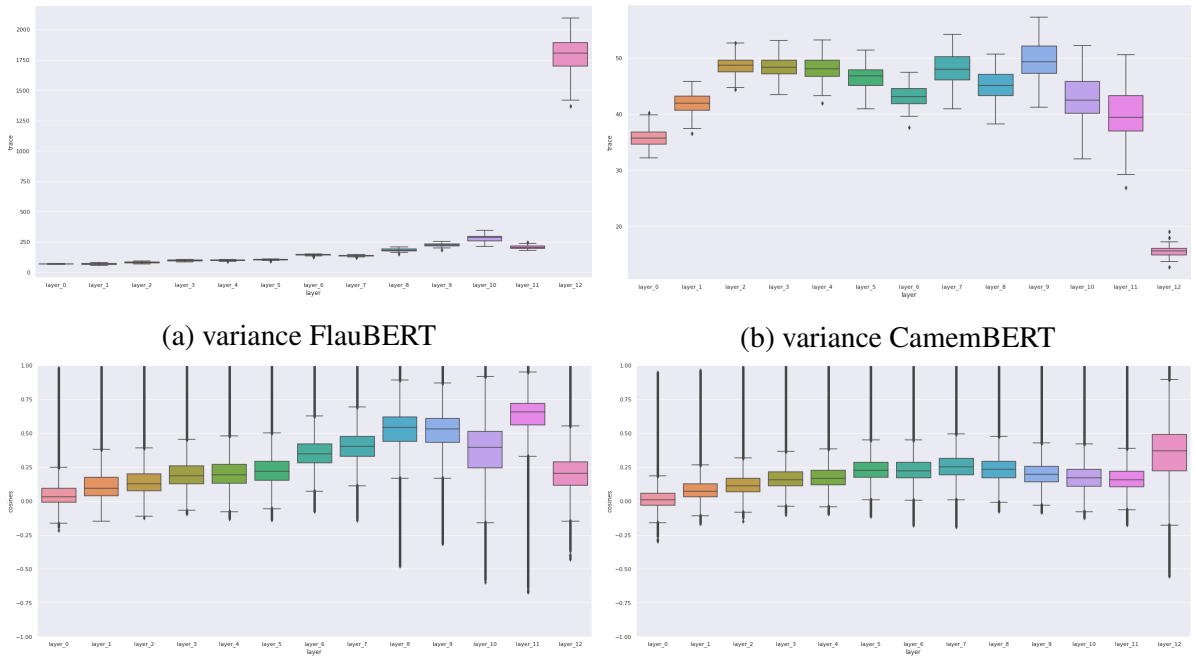
3 Résultats

Dans un premier temps, nous comparons le comportement des deux modèles pré-entraînés. Dans un second temps, nous nous intéressons à l'effet de l'adaptation à la classification de textes.

3.1 FlauBERT *vs.* CamemBERT

Nous analysons tout d'abord la variance des plongements issus des deux modèles. Les figures 1a et 1b montrent l'évolution de la trace de la matrice de covariance après les différentes couches, la distribution de la variance étant calculée sur 100 phrases. Les résultats montrent que le comportement des deux modèles est entièrement opposé : pour FlauBERT, on observe une légère augmentation progressive de la variance jusqu'à la couche 12 avec une explosion après la dernière couche ; pour CamemBERT, on observe une augmentation lente de la variance sur les 3 premières couches puis une

2. https://github.com/lolofo/pir_irisa_insa



(a) variance FlauBERT

(b) variance CamemBERT

(c) similarités FlauBERT

(d) similarités CamemBERT

FIGURE 1 – Distribution de la trace de la matrice de covariance $\sum_j v_j^k(s)$ (en haut) et des similarités intra-phrases $c_{i,j}^k(s)$ (en bas) à travers les couches pour FlauBERT et CamemBERT. En raison d'une différence notable dans les valeurs de variance entre les deux modèles, les échelles des courbes correspondantes ne sont pas les mêmes.

diminution très nette sur la dernière. Dans le premier cas, la variance n'indique aucune convergence des plongements tandis que, dans le second cas, on a une convergence marquée en sortie. Nous notons au passage que pour les deux modèles, la dernière couche d'attention joue un rôle majeur dans la redistribution des plongements, probablement dû à sa proximité avec les têtes de classification utilisées à l'apprentissage. Sur le modèle FlauBERT, une analyse dimension par dimension a montré que l'explosion de variance était dûe à une dizaine de dimensions, indépendamment de la phrase considérée.

Les résultats concernant la dispersion des similarités intra-phrases sont donnés dans les figure 1c et 1d pour 50 phrases. Les tendances observées sont similaires à celles que nous avons obtenues pour la variance. Pour le modèle FlauBERT, nous observons un alignement progressif des représentations des *tokens* dans un cône jusqu'à l'avant-dernière couche, la dernière cassant cet alignement pour augmenter l'isotropie. Pour le modèle CamemBERT, nous observons tout d'abord un alignement progressif, qui diminue légèrement entre les couches 8 et 12 avant de croître significativement sur la dernière couche. En d'autres termes, les représentations dans FlauBERT se concentrent dans une direction rapidement avant de s'écarter sur la dernière couche tandis que, dans CamemBERT, la concentration se fait principalement sur la dernière couche. Nous notons dans les deux cas une forte dispersion des valeurs de similarités. Enfin, l'anisotropie de la dernière couche telle que mesurée par la dispersion des similarités qui caractérisent l'étroitesse du cône dans lequel les plongements se concentrent, est plus forte pour Camembert ($\sim$0.4) que pour FaluBERT ($\sim$ 0.25).

Ces deux mesures, la variance et la similarité cosinus, montrent donc que le phénomène de concentration des représentations dans un cône étroit au fur et à mesure que nous montons dans les couches du

modèle (Ethayarajh, 2019) n'est pas vérifiée pour le modèle FlauBERT pré-entraîné. En revanche, elle se vérifie dans une certaine mesure pour le modèle CamemBERT, où les plongements semblent s'aligner dans un cône sur la dernière couche et se concentrer vers une même valeur (diminution de la variance). Ce dernier comportement est cohérent avec les observations faites sur le modèle BERT pour l'anglais, que ce soit dans Ethayarajh (2019) ou dans les expériences que nous avons menées sur ce denier modèle.

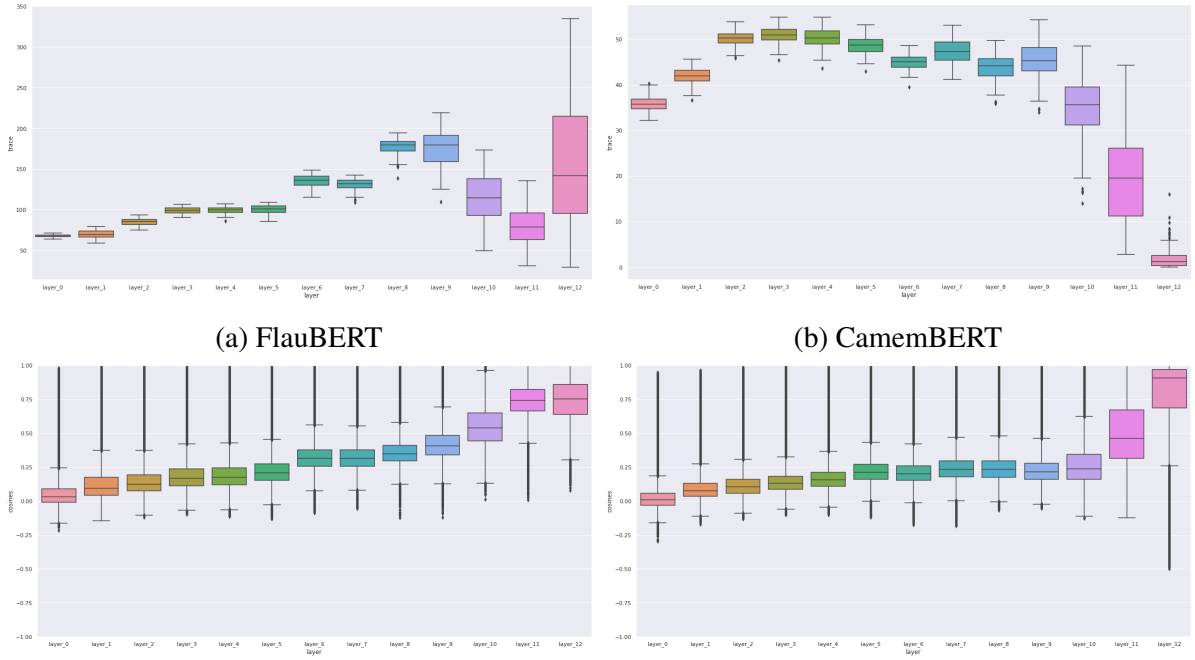
Ces différences de comportement s'observent aussi sur le rang effectif : il est relativement constant ($\sim$550) pour CamemBERT, sauf après la dernière couche où il diminue à 430 ; il oscille pour FlauBERT, augmentant sur les 4 premières couches jusqu'à 567, diminuant ensuite jusqu'à 469 pour réaugmenter légèrement à 522 sur la dernière couche. Avoir plus de dimensions utiles au niveau de la dernière couche pour FlauBERT est cohérent avec les observations faites sur l'anistropie plus forte pour CamemBERT que FlauBERT. Ces valeurs montrent cependant que le nombre de dimensions utiles dans le plongement reste relativement important, malgré l'augmentation de l'anisotropie pour CamemBERT.

Les différences mineures au niveau de l'architecture, du critère et des données d'apprentissage entre les deux modèles ne permettent à notre avis pas d'expliquer une telle différence de comportement, d'autant moins que le modèle anglais BERT, dont l'architecture est très proche de FlauBERT, se comporte de la même manière que CamemBERT. Malheureusement, une étude par ablation pour comprendre l'impact des différents choix dans ces deux modèles demeure hors de portée pour éclairer ces observations : en effet, ré-entraîner un modèle complet de type FlauBERT ou CamemBERT sur la tâche de modèle de langage masqué en faisant évoluer progressivement l'architecture de l'un vers celle de l'autre demanderait trop de temps de calcul. Notons que si, *in fine*, l'anisotropie de la dernière couche est proche pour les deux modèles, les couches de la seconde moitié (7–12) semblent assez différentes de ce point de vue : ce point pourrait en partie expliquer la complémentarité des deux modèles rapportée dans Le *et al.* (2020) et suggère que les informations linguistiques encodées dans ces couches pourraient être significativement différentes d'un modèle à l'autre.

3.2 Impact de l'adaptation à la classification de textes

Nous avons mené la même analyse que précédemment après adaptation à une tâche de classification. Les résultats sont donnés à la figure 2. Contrairement aux modèles pré-entraînés, les deux modèles adaptés se comportent de manière identique et montrent très clairement une convergence des représentations dans un cône au travers des couches, en particulier des couches finales. Nous avons observé que poursuivre l'adaptation renforce cette convergence : nous notons à la fois une concentration dans un cône plus étroit en sortie du modèle, et une concentration qui démarre plus tôt dans les couches. En d'autres termes, une époque supplémentaire modifie le modèle plus en profondeur et renforce son anisotropie. Ces observations se traduisent aussi sur l'évolution du rang effectif qui oscille autour de 550 jusque vers la couche 8 ou 9, avant de chuter rapidement à 207 pour CamemBERT et 341 pour FlauBERT au niveau de la sortie. Enfin, l'analyse de la variance montre que l'évolution de la variance est peu affectée par l'adaptation pour CamemBERT (même tendance) tandis qu'elle l'est grandement pour FlauBERT. Sur ce dernier modèle, nous pouvons observer une augmentation de la variance jusqu'à la couche 10, suivi d'une diminution sur les deux suivantes avant de retrouver l'augmentation finale, cette dernière étant cependant beaucoup moins marquée que pour le modèle pré-entrainé.

Les différences observés sur la géométrie des plongements, entre les différentes couches d'un même modèle et entre les différents modèles, amènent naturellement à se poser la question quant à un



(a) FlauBERT (b) CamemBERT

(c) FlauBERT (d) CamemBERT

FIGURE 2 – Distribution de la trace de la matrice de covariance $\sum_j v_j^k(s)$ (en haut) et des similarités intra-phrases $c_{i,j}^k(s)$ après adaptation

modèle	couche 5	couche 7	couche 9	couche 11	couche 12
FlauBERT pré-entrainé	69.8	71.4	72.0	72.8	73.0
FlauBERT adapté	77.7	85.6	91.9	92.9	93.1
CamemBERT pré-entrainé	73.1	81.3	89.7	84.4	89.6
CamemBERT adapté	77.9	86.5	93.4	94.6	94.5

TABLE 1 – Précision sur la tâche de classification en polarité fondée sur le plongement du *token* CLS après différentes couches.

possible lien entre cette géométrie et la performance dans la tâche de classification en polarité à l'instar des travaux de Torregrossa *et al.* (2020). La tâche de classification en polarité se basant sur le *token* CLS, nous récupérons ainsi le plongement du *token* CLS en sortie de différentes couches afin de l'utiliser comme entrée pour un classifieur simple, en s'inspirant de l'approche adoptée dans Rogers *et al.* (2020). Le classifieur est composé d'une projection du plongement du *token* CLS vers l'espace de classification binaire, avec un *dropout* de 0,1. Les paramètres du classifieur sont estimés sur les données d'apprentissage du corpus, le plongement restant bien entendu fixe. Nous rapportons dans la table 1 les résultats sur phrases du corpus de test en considérant les plongements de CLS après les couches 5, 7, 9, 11 et 12 respectivement.

Les résultats mettent à nouveau en avant les différences entre les deux modèles pré-entrainés : les performance obtenues avec les plongements du réseaux FlauBERT sont modérées et avec une faible augmentation de la précision lors de la monté dans les couches, là où CamemBERT permet d'obtenir une précision nettement supérieur sur chaque couche avec une augmentation à travers les couches largement visible et permet même d'accéder à une précision proche de 90 % sur la dernière couche. Clairement, la concentration des représentations dans un cône plus étroit pour le modèle CamemBERT

pré-entrainé est bénéfique à la tâche de classification de texte. Sans surprise les performances pour les réseaux adaptés sont proches pour les deux modèles et largement supérieures à celles des réseaux pré-entrainés, avec une augmentation progressive des performances au fur et à mesure que l'anisotropie du modèle augmente. Cependant, l'évolution des similarités cosinus entre deux couches (par exemple, 9 à 11) n'est pas proportionnelle au gain en performance, ce qui tend à indiquer que si le lien qui semble exister entre géométrie des plongements, en particulier anisotropie, et performance en classification de texte est bien réel, il est assez complexe.

Nous avons complété ces expériences en nous penchant sur la répartition des exemples positifs et négatifs au sein de l'espace de plongement final. Pour ce faire, nous considérons le plongement du *token* CLS de la première phrase du corpus de test et mesurons sa similarité cosinus avec le plongement du *token* CLS de chacune des autres phrases du corpus, en distinguant les phrases de la même classe des autres. Cette expérience, qui ne s'appuie que sur la première phrase, se veut purement qualitative. À noter que contrairement aux expériences précédentes, nous nous focalisons ici sur la comparaison de deux phrases et non sur la géométrie des plongements au sein d'une même phrase. Comme attendu, les résultats montrent que pour les modèles pré-entrainés, les deux classes ne sont pas séparables tandis qu'elles le sont facilement pour les modèles adaptés. Pour ces derniers, les directions sont quasi-orthogonales pour FlauBERT, avec une similarité cosinus proche de 0.2 en moyenne pour les exemples de classe différente et de 0.95 pour ceux de la même classe, alors qu'elles sont opposées pour CamemBERT (similarité cosinus proche de -0.8 et 1 resp.). Ces résultats suggèrent donc que les différences géométriques des modèles pré-entrainés induisent des comportements différents lors de l'adaptation qui ne se mesure pas sur les taux de classification pour une tâche simple comme la classification de texte selon la polarité.

4 Conclusion

L'étude empirique des propriétés statistiques des plongements aux différents niveaux des modèles FlauBERT et CamemBERT permet globalement de vérifier que l'hypothèse d'anisotropie des plongements contextuels observée pour l'anglais s'applique aux modèles pour le français. Elle permet également de mieux comprendre l'impact de l'adaptation des modèles à une tâche de classification, cette dernière renforçant de manière très significative l'anisotropie, la dispersion et le nombre de dimensions utiles du plongement final. On peut se demander quel est l'impact de l'adaptation à d'autres tâches nécessitant de maintenir une distinction plus forte entre les différents *tokens* d'un texte comme les tâches d'étiquetage. De manière plus surprenante, la comparaison des modèles FlauBERT et CamemBERT fait apparaître une différence importante de comportement des modèles pré-entraînés, différence qui reste difficilement explicable par l'architecture et l'entraînement des modèles. Le prolongement de cette étude consistera donc à tenter d'apporter un éclairage à ces différences, par exemple en regardant les corrélations entre les représentations des mêmes *tokens* par les deux modèles ou en étudiant les informations linguistiques portées par les différentes couches d'un modèle et leur éventuel lien avec les caractéristiques géométriques des espaces de plongement. Enfin, les expériences de classification à partir du plongement du *token* CLS suggèrent un lien entre géométrie et performance dans la tâche de classification de texte, une représentation de chaque phrase plus concentrée dans un cône améliorant les performances de classification, ce que nous observons tant pour les modèles adaptés que pour les modèles pré-entrainés. En revanche, la nature du lien entre géométrie et performance reste difficile à appréhender, rejoignant en cela les observations de Torregrossa *et al.* (2020), et mérite d'être approfondie.

Références

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : pre-training of deep bidirectional transformers for language understanding. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4171–4186.

ETHAYARAJH K. (2019). How contextual are contextualized word representations ? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, p. 55–65.

HERNANDEZ E. & ANDREAS J. (2021). The low-dimensional linear geometry of contextualized word representations. In *25th Conference on Computational Natural Language Learning*, p. 82–93.

HTUT P. M., PHANG J., BORDIA S. & BOWMAN S. R. (2019). Do attention heads in BERT track syntactic dependencies ? Unpublished arXiv preprint 1911.12246.

JAWAHAR G., SAGOT B. & SEDDAH D. (2019). What does BERT learn about the structure of language ? In *Annual Meeting of the Association for Computational Linguistics*, p. 3651–3657.

KOBAYASHI G., KURIBAYASHI T., YOKOI S. & INUI K. (2020). Attention is not only a weight : Analyzing transformers with vector norms. In *Conference on Empirical Methods in Natural Language Processing*, p. 7057–7075.

KOVALEVA O., ROMANOV A., ROGERS A. & RUMSHISKY A. (2019). Revealing the dark secrets of BERT. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, p. 4356–4365.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBE B., BESACIER L. & SCHWAB D. (2020). FlauBERT : unsupervised language model pre-training for French. In *Language Resources and Evaluation Conference*, p. 2479–2490.

LIN Y., TAN Y. C. & FRANK R. (2019). Open Sesame : Getting inside BERT's linguistic knowledge. In *ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 241–253.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A robustly optimized bert pretraining approach. Unpublished arXiv preprint 1907.11692.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representation. In *North American Chapter of the ACL - Human Language Technology*, volume 1, p. 2227–2237.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I. *et al.* (2019). *Language models are unsupervised multitask learners*. Rapport interne, OpenAI.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2020). A primer in bertology : What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866.

ROY O. & VETTERLI M. (2007). The effective rank : A measure of effective dimensionality. In *European Signal Processing Conference*, p. 606–610.

TORREGROSSA F., ALLESIARDO R., CLAVEAU V., KOOLI N. & GRAVIER G. (2021). A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, **11**(2), 85–103.

TORREGROSSA F., CLAVEAU V., KOOLI N., GRAVIER G. & ALLESIARDO R. (2020). On the correlation of word embedding evaluation metrics. In *Language Resources and Evaluation Conference*, p. 4789–4797.

VAN AKEN B., WINTER B., LÖSER A. & GERS F. A. (2019). How does BERT answer questions ? a layer-wise analysis of transformer representations. In *28th ACM International Conference on Information and Knowledge Management*, p. 1823–1832.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, **30**.