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.

Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, par rapport à une collection ou un corpus.

Plus récemment, les transformers (Vaswani et al., 2017) ont été découverts, ouvrant un tout nouveau monde sur le traitement du langage naturel (NLP) et la compréhension du langage naturel (NLU). Comme les réseaux de neurones récurrents (RNNs (Yin et al., 2017)), les transformers sont conçus pour traiter des données séquentielles, comme le langage naturel, pour des tâches telles que la traduction (Liu et al., 2020) et le résumé de texte (Syed et al., 2021). Ils permettent de résoudre un problème assez important de TF-IDF : la polysémie (Ma et al., 2020). Cependant, ces modèles, pour être entraînés, nécessitent de grandes quantités de données, et il est difficile et coûteux d'entraîner un tel modèle à partir de zéro (Devlin et al., 2019). Heureusement, nous avons aujourd'hui accès à un catalogue de modèles pré-entraînés comme BERT (Devlin et al., 2019) ou RoBERTa (Liu et al., 2019), en plusieurs langues.

Ces modèles pré-entraînés ont été entraînés sur des données de langage naturel de la langue cible. Par conséquent, ils peuvent avoir des difficultés à représenter précisément certains mots spécifiques à un domaine, comme le domaine de l'éducation et de la formation professionnelle, que nous cherchons à représenter. En effet, ce domaine est plus complexe à représenter que la langue naturelle, notamment parce qu'il utilise les mots de la langue naturelle dans un sens différent. De plus, les modèles de base pré-entraînés de BERT ne sont pas conçus pour optimiser la transformation des phrases, ou de paires de phrases, en vecteurs (Reimers et Gurevych, 2019), ce qui rend leur performance dans ce domaine assez faible.

Sentence-BERT, ou SBERT (Reimers et Gurevych, 2019), est une approche qui répond au problème de performance de BERT, notamment sur la recherche sémantique, mais qui surpasse également les performances de BERT sur la classification textuelle. SBERT consiste en une modification du modèle pré-entraîné BERT, en utilisant des structures de réseaux siamois ou triplés. Les vectorisations ou représentations de phrases sémantiques peuvent ensuite être comparées à l'aide de la similarité cosinus. Cependant, cette approche étant basée sur un modèle BERT généraliste, elle conserve la connaissance généraliste de la langue, et éprouve les mêmes difficultés que BERT a représenter le domaine spécifique de l'éducation et de la formation professionnelle.

Pour résoudre ces problèmes, nous avons développé BERTEPro. BERTEPro est un Transformer, basé sur BERT, que nous avons continué à entraîner sur la tâche de modélisation du langage masqué (MLM (Devlin et al., 2019)), sur le domaine de l'éducation et de la formation professionnelle, puis affiné sur les tâches d'implication de paires de phrases (Conneau et al., 2018) et de similarité sémantique (Cer et al., 2017). De nombreuses raisons justifient l'utilisation du pré-entraînement en tandem avec l'affinage (*fine-tuning* en anglais) dans notre approche BERTEPro. La poursuite de l'entraînement de BERT sur des données spécifiques au domaine lui a permis de modifier le contexte de certains mots utilisés différemment dans le langage naturel et dans le langage spécifique au domaine. Selon SBERT (Reimers et Gurevych, 2019), l'affinage du modèle sur l'implication des paires de phrases, puis sur les tâches de similarité sémantique, améliore considérablement la représentation sémantique des phrases. Ce pré-entraînement et cet affinage combinés, nous permettent d'améliorer considérablement les performances sur le jeu de données de référence STS, mais aussi sur notre jeu de données de similarité de paires de phrases, généré à partir de textes de formation éducative et professionnelle. Le nouveau cadre proposé s'avère prometteur pour traiter le langage naturel courant ainsi que le domaine spécifique de l'éducation et de la formation professionnelle.

Le reste de l'article est organisé comme suit : La section 2 passe en revue les études récentes sur les méthodes de vectorisation de textes. La section 3 présente notre approche BERTEPro. Le protocole expérimental est présenté dans la section 4. Finalement, la section 5 évalue BERTEPro sur des tâches STS, communes et spécifiques à un domaine, et sur des tâches de classification spécifiques sur des corpus de l'éducation et de la formation professionnelle. Nous soulevons plusieurs questions pour les travaux futurs dans la section 6 et concluons par un résumé de nos contributions.

2 Vectorisation de textes : État de l'art

Dans cette section, nous passons en revue les principales méthodes utilisées pour la vectorisation de textes.

2.1 TF IDF

TF-IDF (Term Frequency, Inverse Document Frequency) (Jones, 1972) est une méthode statistique largement utilisée consistant à représenter un document en fonction de l'importance des termes qu'il contient par rapport à l'ensemble des documents du corpus. La valeur TF-IDF pour un document d et un terme t est calculée comme la fréquence d'occurrence de t dans d par rapport à l'information relative fournie par t dans l'ensemble du corpus.

Chaque document est ainsi représenté par un vecteur dont la dimension correspond à la taille du vocabulaire et nous pouvons calculer la similarité entre deux documents comme la distance entre ces vecteurs (normalisés). Cette méthode est simple à mettre en œuvre et a donné des résultats satisfaisants dans plusieurs cas d'application. Elle est utilisée par certains moteurs de recherche, notamment avec l'algorithme BM25 (Robertson et al., 2009). Cependant, cette méthode nécessite un vocabulaire fixe et la dimensionalité augmente avec sa taille. Comme les poids TF-IDF ne prennent pas en compte le contexte, l'ambiguïté de certains mots est également le principal problème de cette approche. En effet, ce type d'approche ne permet pas de traiter les problèmes de synonymie et de polysémie.

2.2 Transformers

Les Transformers sont un ensemble de modèles qui ont obtenu de très bons résultats ces dernières années, dépassant les LSTM (Hochreiter et Schmidhuber, 1997) et les GRU (Cho et al., 2014). Parmi les applications les plus importantes, on peut citer la traduction de textes ou les agents conversationnels (avec notamment GPT (Radford et al., 2018)).

Les Transformers sont des modèles de séquence à séquence (*seq2seq*). Un modèle *seq2seq* est un modèle qui prend en entrée une séquence (une suite d'éléments du même type) et renvoie une séquence en sortie. Contrairement aux premiers modèles *seq2seq* utilisant dans une architecture Encoder-Decoder les LSTM ou les GRU, l'architecture du Transfomer a hérité aussi du pattern Encoder-Decoder en n'utilisant pas de réseaux récurrents mais seulement le mécanisme d'attention qui est au centre de cette architecture (Vaswani et al., 2017; Niu et al., 2021). En effet, le problème majeur des réseaux récurrents est la limite de mémoire du contexte sur une certaine fenêtre en raison de la nature itérative du flux d'information, qui reste limité à la taille de l'état interne. Bien que LSTM et GRU l'aient grandement amélioré avec un mécanisme

de régulation et d'oubli, la solution apportée pour les modèles transformers par l'attention consiste à retenir et utiliser tous les états cachés des entrées dans le décodeur. Pour chaque sortie générée, le décodeur sélectionne les états cachés associés à certaines entrées. Cela permet également de donner plus de poids aux parties importantes du contexte de manière indépendante.

2.3 Représentation de BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) est un modèle type Transformer développé par Google pour lequel il existe de nombreuses variantes en fonction du langage ou du type d'application. Nous nous intéressons ici à la représentation des mots appris par les modèles BERT. Les Transformers fournissent un mécanisme puissant d'attention sur le contexte pour construire l'état interne de l'encodeur. De cet état interne, nous pouvons extraire une représentation des mots dans le contexte de leur document. Contrairement à un plongement lexical où la représentation du mot est globale, cette représentation de BERT est locale, dépendant du contexte. Dans le cas de la recherche sémantique, cette représentation donne un sens aux mots dans le contexte de la requête ou d'une phrase en général. On pourra alors représenter la requête par un vecteur, et effectuer une recherche de similarité dans l'espace vectoriel du corpus. Différentes variantes linguistiques de BERT ont été proposées, comme CamemBERT (Martin et al., 2020) et FlauBERT (Le et al., 2020) pour la langue française, GottBERT(Scheible et al., 2020) pour l'allemand, etc.

2.4 Sentence-BERT

Sentence-BERT (Reimers et Gurevych, 2019), ou SBERT, consiste en une modification de l'architecture du modèle BERT pré-entraîné, en utilisant des structures de réseaux siamois et triplet, pour dériver des vectorisations de phrases sémantiquement significatives. Les vectorisations de phrases sémantiques peuvent ensuite être comparés à l'aide de la similarité cosinus. L'objectif de SBERT est de répondre à un des problèmes des modèles BERT, qui est la recherche de la paire de phrase la plus similaire, dans le but d'être principalement utilisé pour de la recherche sémantique. Cependant, il a également montré des gains de performance sur des tâches de classification. Il a d'ailleurs montré que l'utilisation de la stratégie de regroupement par moyenne des vecteurs de sortie de SBERT est meilleure que l'utilisation du CLS pour la classification de textes.

3 Approche proposée

L'approche BERTEPro proposée utilise le réseau pré-entraîné BERT et l'étend avec la combinaison de deux mécanismes. Le premier consiste à poursuivre le pré-entraînement de BERT sur le domaine de l'éducation et de la formation professionnelle. Dans le but d'améliorer la représentation de l'espace textuel, le second mécanisme consiste à affiner ce dernier modèle spécifique au domaine sur les tâches d'implication de paires de phrases et de similarité sémantique, conformément à l'approche SBERT.

3.1 Pré-entraînement

Comme expliqué dans (Gururangan et al., 2020), la poursuite du pré-entraînement des modèles linguistiques (LMs) permet d'améliorer drastiquement leurs performances. En effet, cela permet aux Transformers d'adapter la représentation générale du langage naturel apprise sur des corpus généraux, à un nouveau vocabulaire et à une autre utilisation du même vocabulaire. Le domaine de l'éducation et de la formation professionnelle étant particulier, et utilisant un vocabulaire assez proche du langage naturel, nous avons choisi de poursuivre le pré-entraînement de BERT sur des données de formation professionnelle.

Nous avons entraîné notre modèle sur la tâche de masquage de mots (MLM), la tâche principale sur laquelle le modèle de base a été entraîné. Il s'agit d'une tâche auto-supervisée, qui consiste à masquer un mot aléatoire dans une phrase d'entrée. Le modèle génère la phrase avec le mot manquant en sortie, puis l'erreur est mesurée et rétro-propagée. L'objectif est de donner au modèle une connaissance plus précise du domaine dans son état interne. Les connexions d'attention sont renforcées pour les mots qui sont rares ou utilisés dans un contexte différent auparavant.

3.2 Affinage

Lorsque nous utilisons BERT pour extraire une représentation de phrases, nous utilisons les poids des couches de sortie du modèle comme une représentation vectorielle des mots, ou de la phrase en faisant la moyenne des vecteurs de mots d'une phrase. Cependant, la tâche d'apprentissage du modèle de base (MLM) ne donne pas de contraintes particulières pour cette tâche cible au niveau des couches de sortie. Comme expliqué dans (Reimers et Gurevych, 2019) pour améliorer les performances des Transformers sur la représentation sémantique, il est nécessaire d'affiner l'entraînement sur des jeux de données spécifiques. Dans notre cas, les tâches et les jeux de données pertinents sont :

— NLI (Implication de phrases) : la tâche NLI ou *entailment textuel* consiste à prédire pour une paire de phrases en entrée, leur relation logique en sortie (implicative, neutre ou contradictoire).

— STS (Similarité sémantique de textes) : le jeu de données STSb-fr, traduction française du jeu de données de référence STSb (Wang et al., 2019), permet d'évaluer les performances du modèle pour la tâche *Semantic Textual Similarity* (STS). Un affinage rapide du modèle sur ce jeu de données ajuste les poids des couches de sortie pour obtenir une bonne représentation sémantique des phrases.

L'affinage (*fine-tuning*) du modèle sur ces deux tâches enchaînées permet au modèle d'affiner ses couches internes sur l'aspect sémantique de la représentation des phrases.

4 Analyse expérimentale

Cette section présente une analyse expérimentale approfondie des performances de l'approche proposée BERTEPro.

Nos expérimentations sont consacrées au domaine de l'éducation et de la formation professionnelle en langue française. Par conséquent, le transformer FlauBERT basé sur le français est

adopté ici. Nous avons choisi ce modèle car il possède une version sans casse. Ceci est primordial dans notre cas d'application, car notre jeu de données n'est pas assez propre pour utiliser un modèle sensible à la casse. Ensuite, afin que notre modèle ne soit pas trop complexe à entraîner et à utiliser en production, nous avons choisi le modèle FlauBERT base uncased pour la langue française. FlauBERT base uncased accepte un maximum de 512 tokens en entrée, est composé de 12 couches d'encodage, et peut retourner jusqu'à 512 tokens.

Pour poursuivre le pré-entraînement du modèle de base, nous avons collecté des données et des méta-données sur environ $500,000$ formations éducatives ou professionnelles françaises à partir de différentes sources disponibles en ligne. La diversité des sources était primordiale pour réduire le biais induit par la syntaxe spécifique de chaque source. Pour automatiser la collecte de données, nous avons utilisé Selenium, avec ChromeDriver pour le navigateur web Chrome. Nous avons ensuite utilisé BeautifulSoup pour extraire les données des différents sites. Nous avons choisi d'extraire le titre et la description de ces formations, ou plus largement, des objets d'apprentissage (*Learning Objects*). Enfin, nous avons découpé notre corpus en séquences de 100 mots, avec un chevauchement de 16 mots entre les séquences S et $S+1$. Cette méthode nous permet de toujours garder un contexte autour des mots dont nous souhaitons apprendre ou améliorer la représentation. L'apprentissage a été effectué pendant 3 époques avec 80000 itérations par époque et une taille de lot de 16. 10% des données sont utilisées pour la validation du modèle. L'apprentissage a duré environ 12 heures sur un GPU Nvidia P100.

Ensuite, nous avons affiné notre modèle pré-entraîné sur les tâches NLI et STS comme dit précédemment, qui sont des jeux de données de référence libres de droits. Pour l'affinage sur ces deux tâches, nous avons choisi les mêmes paramètres d'affinage que dans SBERT (Reimers et Gurevych, 2019). L'affinage sur la tâche NLI a été effectué pendant une époque avec une taille de lot de 16. La durée de cet affinage a été d'environ 3.5 heures sur un GPU Nvidia P100. L'affinage sur la tâche STS a été effectué pendant 4 époques, également avec une taille de lot de 16. La durée de cet affinage n'a été que de 5 minutes sur un GPU Nvidia P100.

5 Évaluation des résultats

Cette section présente le protocole expérimental utilisé ainsi que les résultats obtenus pour tester l'efficacité de notre modèle de vectorisation de textes BERTEPro dans le domaine de l'éducation et de la formation professionnelle. Cette évaluation concerne deux tâches différentes : (1) la similarité sémantique et (2) la classification de données textuelles.

5.1 Évaluation pour une tâche de similarité sémantique

Pour la première tâche, nous avons d'abord évalué notre modèle sur le jeu de données STSb-fr, qui est l'un des jeux de données de référence les plus utilisés pour la tâche de similarité sémantique des textes en langue française. Nous avons également généré un jeu de données de référence dans le domaine de l'éducation et de la formation professionnelle en France. Ce jeu de données, appelé STS-Trainings dans la suite, est un jeu de données composé de paires de phrases issues de formations, et de leur score de similarité. Pour générer ce jeu de données de référence, nous divisons notre corpus de 500000 formations françaises en phrases. Nous sélectionnons aléatoirement la première phrase dans le corpus. L'approche BM25 (Robertson

et al., 2009), un dérivé de TF-IDF, est ensuite utilisée pour calculer la distance entre la première phrase sélectionnée et le reste des phrases du corpus. Pour choisir la deuxième phrase de la paire, nous sélectionnons dans 10% des cas la phrase la plus proche, dans 10% des cas une phrase aléatoire, et dans 80% des cas, une phrase aléatoire parmi les 100 phrases les plus proches. Nous avons ensuite entraîné un encodeur croisé basé sur SFlauBERT sur le jeu de données STSb-fr, afin de créer un modèle capable d'évaluer la similarité entre les phrases. Cet encodeur croisé est ensuite utilisé sur nos paires de phrases extraites de notre corpus d'entraînement, afin d'obtenir un score de similarité sémantique en cosinus entre les phrases de chaque paire. Comme dans (Reimers et Gurevych, 2019; Reimers et al., 2016), les performances de notre modèle BERTEPro sont évaluées en utilisant la corrélation des rangs de Spearman sur les scores de similarité cosinus.

Pour mieux évaluer l'efficacité des deux mécanismes de pré-entraînement et d'affinage de notre modèle BERTEPro, la qualité de notre similarité sémantique obtenue sur les jeux de données STSb-fr et STS-Trainings a été comparée à celle obtenue par SFlauBERT, dans lequel seul le mécanisme d'affinage (sur les tâches NLI et STS) est effectué, et flaubert-base-uncased-xnli ajusté uniquement sur la tâche NLI. Ces deux approches comparées n'ont pas été pré-entraînées sur les données spécifiques du domaine de l'éducation et de la formation professionnelle. L'idée étant d'évaluer l'apport des différentes étapes de notre méthodologie sur à la fois un jeu de données général STSb-fr et spécifique STS-Trainings. D'autre part et par souci d'exhaustivité, BERTEPro est également comparé à la méthode classique de vectorisation TF-IDF.

	STSb-fr	STS-Trainings
TF-IDF	54.60	67.95
flaubert-base-uncased-xnli	80.68	72.23
SFlauBERT	83.06	73.78
BERTEPro	83.05	75.90

TAB. 1 – *Évaluation sur le jeu de données de test STSb-fr et STS-Trainings en termes de corrélation de rang de Spearman.*

Les performances détaillées en termes de corrélation des rangs de Spearman de chaque modèle de vectorisation pour les deux ensembles de données sont indiquées dans le tableau 1. Nous avons constaté que la représentation donnée par BERTEPro est plus efficace pour la représentation de textes dans des domaines spécifiques (*i.e.* STS-Trainings pour le domaine de l'éducation et de la formation professionnelle). BERTEPro présente les meilleures performances en termes de mesure de corrélation de rang de Spearman que les autres méthodes. Comme on peut l'observer dans le tableau 1, la combinaison des mécanismes de pré-entraînement et d'affinage a un effet remarquable sur l'amélioration significative de la vectorisation des phrases et améliore clairement les connaissances pour un ensemble de données du domaine spécifique en comparaison à des approches non pré-entraînées et affinées en partie ou totalement sur les tâches NLI et STS. Même si BERTEPro a été pré-entraîné sur un jeu de données spécifique, une inspection plus approfondie des résultats sur le jeu de données STSb-fr démontre la capacité de BERTEPro à conserver d'excellents résultats (comparables à ceux de SFlauBERT) sur la représentation du langage naturel.

Par souci d'exhaustivité et une évaluation qualitative des similarités sémantiques issues de BERTEPro, nous rapportons les résultats de la comparaison entre BERTEPro et SFlauBERT sur certaines phrases de requêtes du domaine de l'éducation et de la formation professionnelle. Les résultats de ces modèles dans la recherche de similarités entre un exemple de requête et plusieurs phrases sont donnés dans le tableau 2. Nous avons choisi les phrases de la manière suivante. La 1ère est la phrase qui est sémantiquement la plus proche de la requête, la 2ème est une phrase proche ou éloignée et la 3ème est une phrase éloignée ou sans rapport. Pour la première requête *Formation en bureautique* par exemple, et ses 3 phrases liées *Maîtriser microsoft excel*, *Passer votre certification tosa* et *Maitriser l'anglais à l'oral*, nous remarquons que BERTEPro est bien meilleur que SFlauBERT pour trouver les phrases les plus proches sémantiquement à la bureautique parmi les trois phrases fournies. En particulier, BERTEPro est capable de comprendre que TOSA (une certification en bureautique) fait référence à la bureautique ainsi qu'à Microsoft Excel.

Un autre exemple intéressant est donné par la troisième requête *Permis de conduire poids lourds*. En France, le permis de conduire nécessaire pour conduire des poids lourds est le permis C. Le permis B est le permis de conduire pour les voitures, et le permis D est le permis de conduire plus de 8 personnes. Nous remarquons que BERTEPro a mieux assimilé, comparé à SFlauBERT, que la catégorie de permis de conduire la plus proche nécessaire pour conduire un camion est le permis 'C'. Ces résultats montrent que BERTEPro est capable d'être plus précis sur des sujets très spécifiques tels que le permis de conduire nécessaire pour conduire un poids lourd, ou que manger du bœuf ne fait pas vraiment partie du métier de boucher.

Ces résultats confirment la capacité du mécanisme de pré-entraînement sur les domaines spécifiques de BERTEPro à générer une meilleure représentation du texte en améliorant la compréhension des corpus de domaines spécifiques.

Requêtes	Phrases	SFlaubert	BERTEPro
Formation en bureautique	Maîtriser Microsoft Excel	55.70	63.70
	Passer sa Certification tosa	37.70	60.70
	Maîtriser l'anglais à l'oral	27.00	15.80
Devenir ingénieur informatique	Développement informatique	72.40	79.80
	Programmation informatique	55.00	75.90
	Langues étrangères	4.40	3.70
Les bases de la boucherie	Découpage de viande	63.30	64.30
	Manger du boeuf	49.10	35.10
	Aller à l'école	25.30	3.30
Permis de conduire poids lourd	Permis c	45.20	58.90
	Permis d	43.60	51.40
	Permis b	47.50	49.90
Faire la java	Faire la fête	40.03	35.17
	Programmation informatique	27.90	31.31

TAB. 2 – *Score de Similarités des requêtes-phrases de SFlauBERT et BERTEPro.*

Cependant, nous notons que la similarité sémantique de phrases par BERTEPro, hors du domaine spécifique de l'éducation et de la formation professionnelle, peut obtenir des résultats

moins intéressants que SFlauBERT. La cinquième requête nous montre que BERTEPro peut éprouver des difficultés à représenter des mots peu utilisés du langage naturel, qui ont un sens différent dans le domaine spécifique. Ceci s'explique notamment par le fait que 'Java' est présent uniquement comme un langage de programmation dans les données d'entraînement, et non comme un synonyme de fête.

5.2 Évaluation pour une tâche de classification

Dans un deuxième temps, la représentation vectorielle obtenue par BERTEPro a été évaluée sur des tâches de classification supervisée sur plusieurs jeux de données textuels issus du domaine de l'éducation et de la formation professionnelle. Pour ne pas induire de biais sur l'évaluation, nous avons choisi uniquement des jeux de données sur lesquels BERTEPro n'a pas été pré-entraîné.

Le premier jeu de données [1] est une liste de 725 intitulés de métier fournie par Onisep, qui est un opérateur de l'état Français. Ces métiers sont classés par code ROME, une arborescence de familles de métiers construite sur 3 niveaux. Ici, nous chercherons à évaluer BERTEPro sur la classification des libellés de métiers dans le premier niveau hiérarchique des codes ROME associés, composé de 13 classes.

Le deuxième [2] est une liste de formations de l'enseignement supérieur français, proposée aussi par Onisep. Nous chercherons ici à classifier les 4000 intitulés de formations dans leur niveau de sortie attendu (de Bac+1 à Bac+8).

Le troisième jeu de données [3] est issu de Mon Compte Formation (MCF). Celui-ci est composé de 11502 formations et 1104 certifications, chacun associé à un code ROME. Deux jeux de données découlent ainsi de celui-ci. Nous chercherons, pour le premier, à classifier les formations, et les certifications pour le deuxième, dans les 13 classes qui forment le premier niveau hiérarchique des codes ROME.

En ce qui concerne la classification, nous proposons d'utiliser l'algorithme des k-plus proches voisins (avec $k = 1$) comme un apprenant de base. Le k-NN est un algorithme simple et intuitif qui a été largement considéré pour évaluer l'efficacité des techniques de calcul de similarité. Dans notre cas, la distance cosinus sera utilisée comme mesure de distance dans k-NN. En effet, nous souhaitons comparer nos approches sur leur manière de représenter les textes dans l'espace, et la distance cosinus est adaptée pour résoudre ce genre de problèmes.

Nous avons comparé BERTEPro à quatre autres algorithmes de représentation vectorielle de données textuelles que sont (1) TF-IDF, (2) Flaubert-base-uncased qui est la version de base de Flaubert, (3) SFlauBERT, qui correspond à la version affinée de Flaubert-base-uncased sur les tâches de NLI et STS, (4) Flaubert-education, qui est la version de Flaubert dont nous avons continué le pré-entraînement sur les données éducatives.

L'entraînement de l'algorithme k-NN utilisant les représentations vectorielles obtenues par chacun des 4 algorithmes précédents ainsi que BERTEPro a été évalué par une validation croisée à 5-blocs répétée 10 fois. Nous observons dans le tableau 3 les résultats de cette évaluation

1. https://opendata.onisep.fr/data/5fa5949243f97/2-ideo-metiers-onisep.htm
2. https://opendata.onisep.fr/data/605344579a7d7/2-ideo-actions-de-formation-initiale-univers-enseignement-superieur.htm
3. https://opendata.caissedesdepots.fr/explore/dataset/moncompteformation_catalogueformation/

en termes de moyenne et d'écart type de la mesure d'accuracy sur les 50 itérations. Pour examiner si les résultats sont statistiquement significatifs, des tests de student ont été effectués avec un seuil de signification de 5%.

	ONISEP		MCF	
	Métiers	Formations	Certifications	Formations
TF-IDF	37.28 ± 0.031 •	66.24 ± 0.016 •	70.11 ± 0.029 •	91.84 ± 0.005 •
Flaubert-base-uncased	21.30 ± 0.034 •	55.11 ± 0.015 •	28.30 ± 0.027 •	47.88 ± 0.009 •
Flaubert-education	50.17 ± 0.033 •	$\mathbf{68.87 \pm 0.015}$ ○	72.73 ± 0.023 •	89.49 ± 0.007 •
SFlauBERT	55.99 ± 0.037 •	66.07 ± 0.016 •	77.57 ± 0.024 •	92.17 ± 0.006 •
BERTEPro	$\mathbf{58.52 \pm 0.034}$	67.74 ± 0.015	$\mathbf{80.96 \pm 0.022}$	$\mathbf{93.56 \pm 0.004}$

TAB. 3 – *Évaluation de l'algorithme 1-NN, sur les jeux de données Onisep et MCF en termes de Moyenne et écart-type de l'accuracy, obtenu sur* 50 *exécutions.* •/○ *indique que* BERTEPro *est significativement meilleur/pire, avec un seuil de signification de 5%.*

Plusieurs constatations peuvent être tirées lors de l'examen des résultats obtenus dans le tableau 3. Nous notons dans un premier temps, que Flaubert-education est meilleur que Flaubert-base-uncased. Ceci montre bien l'importance du pré-entraînement sur les données éducatives. Nous remarquons que BERTEPro affiche des performances statistiquement plus élevées que SFlauBERT et TF-IDF. Rappelons que BERTEPro et SFlauBERT sont respectivement les versions de Flaubert-education et Flaubert-base-uncased, affinées sur les tâches de NLI et STS. Ceci montre encore une fois l'importance du pré-entraînement sur les données éducatives afin de fournir au modèle de vectorisation de textes une connaissance plus précise du domaine étudié. Dans trois expériences sur quatre, il apparaît que l'étape d'affinage par NLI et STS est remarquablement efficace pour améliorer significativement la qualité de la représentation vectorielle des phrases et donc la classification de textes de l'éducation et de la formation professionnelle pour BERTEPro en comparaison avec Flaubert-education.

Nos modèles BERTEPro en version française, et en version anglaise, ont été publiés sur une plateforme de modèles libres de droits, HuggingFace, sur le profil de l'entreprise Inokufu [4]. Ces modèles comptabilisent chaque mois environ 5000 téléchargements.

6 Conclusion

Nous avons montré que FlauBERT, adapté à la tâche de similarité de textes de phrases, ne convient pas aux similarités de phrases spécifiques à un domaine, comme le domaine de l'éducation et de la formation professionnelle. Pour surmonter ce problème, nous avons présenté BERTEPro. L'approche BERTEPro utilise un réseau FlauBERT pré-entraîné et l'étend avec la combinaison de deux mécanismes. Le premier consiste à poursuivre le pré-entraînement de FlauBERT sur le domaine de l'éducation et de la formation professionnelle, puis, dans un deuxième temps, à l'affiner sur deux tâches spécifiques, l'inférence en langage naturel (NLI) et la similitude de textes de phrases (STS). Nous avons évalué BERTEPro sur (1) une tâche commune et spécifique au domaine de la similarité de textes de phrases, où il surpasse notamment

4. https://huggingface.co/inokufu/

les modèles de vectorisation de phrases non spécifiques au domaine de l'éducation et de la formation professionnelle, tout en conservant d'excellents résultats sur le langage naturel et (2) une tâche de classification qui a confirmé l'efficacité de notre stratégie pour mieux catégoriser des données textuelles issues du domaine de l'éducation et de la formation professionnelle.

Il convient de mentionner que notre proposition est générique et peut servir d'approche d'apprentissage de représentation vectorielle de phrases efficace dans de nombreux autres domaines spécifiques, dans lesquels les modèles basés sur BERT manquent de précision. Des expériences supplémentaires sur d'autres modèles de base, tels que les modèles FlauBERT large ou CamemBERT, sont en cours de réalisation. L'approche décrite pourrait également être généralisée à d'autres langues avec BERT ou RoBERTa comme modèles de base, et même à des modèles multilingues, avec la distillation de la connaissance (Reimers et Gurevych, 2020).

Références

Cer, D. M., M. T. Diab, E. Agirre, I. Lopez-Gazpio, et L. Specia (2017). Semeval-2017 task 1 : Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval@ACL 2017, Vancouver, Canada*, pp. 1–14.

Cho, K., B. van Merrienboer, D. Bahdanau, et Y. Bengio (2014). On the properties of neural machine translation : Encoder-decoder approaches. In *SSST@EMNLP, Doha*, pp. 103–111.

Conneau, A., R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, et V. Stoyanov (2018). XNLI : evaluating cross-lingual sentence representations. In *EMNLP, Brussels, Belgium*, pp. 2475–2485.

Devlin, J., M. Chang, K. Lee, et K. Toutanova (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019, Minneapolis, USA*, pp. 4171–4186.

Gururangan, S., A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, et N. A. Smith (2020). Don't stop pretraining : Adapt language models to domains and tasks. In *ACL, Online*, pp. 8342–8360.

Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Le, H., L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, et D. Schwab (2020). Flaubert : Unsupervised language model pre-training for french. In *LREC, Marseille, France*, pp. 2479–2490.

Liu, X., K. Duh, L. Liu, et J. Gao (2020). Very deep transformers for neural machine translation. *arXiv preprint arXiv :2008.07772*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, et V. Stoyanov (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

Ma, R., L. Jin, Q. Liu, L. Chen, et K. Yu (2020). Addressing the polysemy problem in language modeling with attentional multi-sense embeddings. In *ICASSP 2020*, pp. 8129–8133. IEEE.

Martin, L., B. Müller, P. J. O. Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, et B. Sagot (2020). Camembert : a tasty french language model. In *ACL*, pp. 7203–7219.

Niu, Z., G. Zhong, et H. Yu (2021). A review on the attention mechanism of deep learning. *Neurocomputing 452*, 48–62.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever, et al. (2018). Improving language understanding by generative pre-training.

Ranjan, N., K. Mundada, K. Phaltane, et S. Ahmad (2016). A survey on techniques in nlp. *International Journal of Computer Applications 134*(8), 6–9.

Reimers, N., P. Beyer, et I. Gurevych (2016). Task-oriented intrinsic evaluation of semantic textual similarity. In *COLING*, pp. 87–96.

Reimers, N. et I. Gurevych (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP, Hong Kong, China, November 3-7, 2019*, pp. 3980–3990.

Reimers, N. et I. Gurevych (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP, Online*, pp. 4512–4525.

Robertson, S., H. Zaragoza, et al. (2009). The probabilistic relevance framework : Bm25 and beyond. *Foundations and Trends® in Information Retrieval 3*(4), 333–389.

Scheible, R., F. Thomczyk, P. Tippmann, V. Jaravine, et M. Boeker (2020). Gottbert : a pure german language model. *arXiv preprint arXiv :2012.02110*.

Syed, A. A., F. L. Gaol, et T. Matsuo (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access 9*, 13248–13265.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems 30*.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, et S. R. Bowman (2019). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In *ICLR 2019, New Orleans, LA, USA9*.

Yin, W., K. Kann, M. Yu, et H. Schütze (2017). Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv :1702.01923*.

Summary

FlauBERT and CamemBERT have established a new state-of-the-art performance for the understanding. Recently, SBERT has transformed the use of the pre-trained BERT network, to reduce the computational effort of sentence embeddings, while maintaining BERT's accuracy. However, these models were trained on non-specific texts of the French language, which therefore do not allow a fine representation of texts from specific domains, such as education and professional training. In this paper, we present BERTEPro, a language model based on FlauBERT, whose training has been extended on texts of the specific domain of education and professional training, before being fine-tuned on NLI and STS tasks. The performance evaluation of BERTEPro on common and specific STS tasks, as well as on classification tasks on textual data from the domain of education and professional training, confirmed that the proposed methodology enjoys significant advantages over other state-of-the-art methods.