Etude approfondie des représentations de données textuelles dans l'apprentissage non supervisé

Mira Ait-Saada, Mohamed Nadif

Etude approfondie des représentations de données textuelles dans l'apprentissage non supervisé

Mira Ait-Saada[*,**], Mohamed Nadif[*]

[*]Centre Borelli UMR9010, Université Paris Cité, 75006 Paris
[**]Caisse des Dépôts et Consignations, Datalab, 75013, Paris

Résumé. Les plongements de textes ont récemment suscité un grand intérêt dans plusieurs tâches telles que la classification de textes/documents et la réponse aux questions. Cependant, bien que de nombreux défis soient rencontrés dans le domaine de l'apprentissage non supervisé, on en sait beaucoup moins sur la pertinence de ces différents plongements lorsqu'on dispose d'un ensemble de documents non labellisés. Dans cet article, nous étudions l'utilisation de telles représentations sur des tâches non supervisées : le *clustering* de documents et la visualisation. Ainsi, pour répondre à l'objectif de *clustering*, nous proposons d'utiliser une *approche tandem* combinant des techniques de réduction de dimension et de *clustering*. Nous montrons d'abord l'avantage de s'appuyer sur le sous-espace obtenu par *Uniform Manifold Approximation and Projection* (UMAP) pour le *clustering* plutôt que d'utiliser la réduction de dimension basée sur l'Analyse en composantes principales (ACP), plus souvent utilisée. Ensuite, à travers des expériences réalisées sur des jeux de données réels, nous montrons l'efficacité de l'approche tandem proposée sur des modèles pré-entraînés par rapport aux stratégies de ré-entraînement proposées dans la littérature.

1 Introduction

Pour des besoins de labellisations de données de plus en plus massives, l'apprentissage non supervisé redevient un des principaux challenges de la science des données. De ce fait le *clustering* et la visualisation, par exemple, peuvent être très utiles pour créer de la valeur à partir des données non labellisées et ce notamment dans le contexte de données textuelles. Actuellement, un large éventail de représentations de données textuelles est proposé aux praticiens parmi lesquelles les sacs de mots sparses ou *Bag-Of-Words* (BOW) ainsi que les sacs de mots denses tels que Word2vec (Mikolov et al., 2013) et GloVe (Pennington et al., 2014) également appelés plongements de mots statiques. Plus récemment, Les représentations de textes/documents fournies par les Modèles de Langue Pré-entraînés basés sur les Transformeurs (MLPT) comme BERT (Devlin et al., 2019) et RoBERTa (Liu et al., 2019) qui produisent de différentes manières des représentations mot par mot pour représenter un document.

Malgré la multiplication des méthodes de plongement, il n'y a pas de réponse claire quant aux performances attendues dans un contexte non supervisé, où aucun label n'est disponible. En particulier, les plongements basés sur les Transformeurs suscitent de plus en plus d'intérêt,

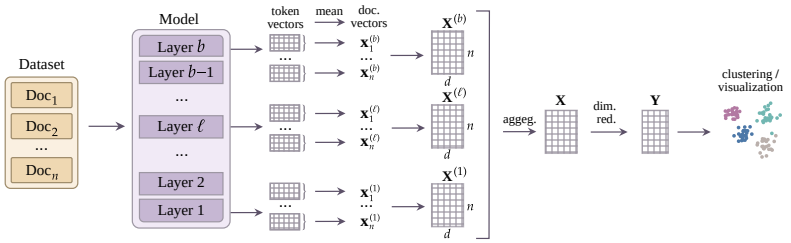

FIG. 1 – Approche Tandem avec un modèle Transformeur pour le *clustering* et la visualisation.

obtenant de très bon résultats dans de nombreuses tâches de traitement automatique des langues (TAL) telles que la réponse aux questions et la Similarité Textuelle Sémantique (STS), mais sont beaucoup moins présents dans le domaine non supervisé.

Par ailleurs, il a été montré dans (Reimers et Gurevych, 2019) que les performances obtenues par BERT sur la version non supervisée de la STS (N-STS) sont moins bonnes que celles obtenues par GloVe, mais l'étude s'est concentrée uniquement sur la dernière couche de BERT et sans aucun post-traitement alors qu'il a été démontré que c'est loin d'être la meilleure stratégie pour tirer pleinement profit des MLPT (Li et al., 2020; Ait-Saada et al., 2021). Dans nos expérimentations, nous étendons l'analyse à l'ensemble des représentations fournies par un modèle Transformeur multi-couches et pas seulement celle fournie par la dernière couche.

Pour améliorer la qualité des MLPT, plusieurs stratégies de réentraînement sont proposées dans la littérature comme celle proposée par Reimers et Gurevych (2019), qui réentraînent un MLPT siamois sur les tâches NLI et STS, améliorant ainsi les performances obtenues par la dernière couche de BERT et de RoBERTa sur la tâche de N-STS. Cette approche est censée être bien adaptée à des tâches non supervisées, y compris le *clustering*, mais n'a pas été évaluée sur ce dernier. DvBERT (Cheng, 2021) est également réentraîné sur une tâche supervisée basée sur les intéractions entre les mots. D'autre part, plusieurs approches non supervisées sont proposées (Carlsson et al., 2021; Gao et al., 2021; Zhang et al., 2020; Liu et al., 2021; Yan et al., 2021), toutes basées sur des objectifs auto-supervisés et ne nécessitant aucune donnée labellisée. Toutes les approches susmentionnées ont été exclusivement évaluées sur la tâche N-STS et on ne sait pas, à ce jour, si elles sont bien adaptées pour le *clustering*.

Une autre façon d'améliorer les résultats obtenus par ces représentations est de s'appuyer sur des techniques de post-traitement appliquées aux vecteurs en sortie. Ces approches utilisent principalement la réduction de la dimensionnalité (RD) basée sur l'ACP, qui s'est avérée suffisamment efficace pour capturer les informations sémantiques tout en réduisant les dimensions. Dans le cas des plongements statiques, une approche basée sur l'ACP proposée dans (Raunak et al., 2019) est utilisée pour réduire de moitié les dimensions sans altérer les performances. En ce qui concerne les MLPT, la RD basée sur l'ACP avec une étape de *whitening* (cf. Section 2) a été évaluée dans (Su et al., 2021) pour la N-STS et dans (Ait-Saada et al., 2021) pour le *clustering* où elle a montré une amélioration significative des performances.

Dans cet article, nous menons une étude entièrement non supervisée pour déterminer laquelle de ces représentations est la plus appropriée pour effectuer le *clustering* de documents. Chacune d'elles est évaluée après avoir effectué un post-traitement à l'aide de techniques de *réduction de dimension* (RD) telles que l'ACP et UMAP (McInnes et al., 2018). Cette approche *tandem* permet de distiller l'information fournie par des représentations issues de modèles pré-entraînés, obtenant ainsi une amélioration surprenante des résultats ainsi qu'une réduction

drastique des dimensions. Notons qu'à ce jour, ce sujet n'a pas encore été abordé en profondeur dans le contexte des représentations textuelles.
Les principales contributions de cet article sont les suivantes :

— Nous abordons la question du choix de la bonne représentation pour effectuer un *clustering* et une visualisation efficaces. Une étude comparative est réalisée pour évaluer les performances des modèles pré-entraînés et ré-entraînés ainsi que des représentations statiques.

— Nous évaluons différentes techniques de post-traitement basées sur la RD, dans le cadre d'une approche de *clustering* tandem, montrant qu'on peut faire mieux que les approches précédemment proposées, à la fois en termes de *clustering* et de visualisation 2D.

2 Approche Tandem

Étant donné un corpus $\mathcal{D}$ de n documents, plusieurs façons de le représenter sont possibles. Dans le cas d'un modèle Transformeur à b couches, on se retrouve avec b matrices de données $\mathbf{X}_\ell,\ \ell \in 1, \ldots, b$ (cf. Figure 1), à partir desquelles nous dérivons une matrice unique $\mathbf{X}$ en combinant un certain nombre de couches. Dans notre étude, nous évaluons les vecteurs en sortie de la dernière couche (« `last` ») comme dans (Reimers et Gurevych, 2019; Cheng, 2021; Carlsson et al., 2021), la combinaison (moyenne) des deux dernières couches (« `last-2` ») comme dans (Li et al., 2020; Yan et al., 2021) ainsi que la combinaison de toutes les couches (« `all` ») comme suggéré dans (Ait-Saada et al., 2021). Dans cette étude, nous écartons l'utilisation du *token* `[CLS]` qui est encore utilisé dans certaines approches de plongement de phrases (Gao et al., 2021; Liu et al., 2021) que nous n'utilisons pas dans cette étude.

L'approche tandem consiste à combiner la RD et le *clustering* comme le montre la Figure 1. Dans ce cas, la RD est considérée comme une étape de post-traitement et vise à compresser $\mathbf{X}$ et à améliorer la qualité du *clustering*. Étant donné une matrice $\mathbf{X}_{(n\times d)}$, nous appelons sa version réduite $\mathbf{Y}_{(n\times d')}$, $d' \ll d$. Pour respecter le contexte non supervisé de l'étude et faire une comparaison équitable entre les techniques de RD, nous prenons $d' = 10$. En outre, les techniques de RD sont également couramment utilisées pour la visualisation avec $d' = 2$ où $\mathbf{Y}$ peut être visualisé sur un plan 2D.

Dans les travaux précédents, le post-traitement des plongements a généralement été effectué en se basant sur l'ACP (Raunak et al., 2019; Ait-Saada et al., 2021; Su et al., 2021). On définit la représentation réduite de $\mathbf{X}$ classiquement dérivée par l'ACP comme les projections $\mathbf{Y} = \mathbf{XQ}$, où $\mathbf{Q}$ est composé des d' premiers vecteurs propres de $\mathbf{X}^T\mathbf{X}$. Dans cet article, nous évaluons l'impact de l'opération de *whitening* qui a prouvé son efficacité sur les plongements Transformeur (Su et al., 2021; Ait-Saada et al., 2021) mais n'a pas été comparée à d'autres approches de RD. L'opération de *whitening* consiste à utiliser $\mathbf{Y} = \mathbf{XQ}/\sqrt{\Delta}$ au lieu de $\mathbf{XQ}$, Δ contenant les d' premières valeurs propres de $\mathbf{X}^T\mathbf{X}$. Nous notons l'ACP classique par « PCA » et la version avec *whitening* par « PCA_w ».

Une autre façon de réduire la dimension utilisée dans cet article est UMAP, qui est une technique de RD non-linéaire visant à construire un graphe qui se rapproche de la structure des données dans l'espace d'origine, suivi d'une projection dans un espace de dimension réduite. Dans le cas du post-processing des vecteurs avec $d' > 2$, avons préféré utiliser UMAP au lieu de t-SNE en raison des restrictions de calcul de t-SNE qui la rendent inadaptée lorsque $d' > 3$, ce qui en fait une méthode principalement utilisée pour la visualisation. Une autre différence

notable est que UMAP permet un meilleur équilibre entre les structures locales et globales. En particulier, UMAP préserve efficacement la structure globale grâce à son approximation topologique basée sur des hypothèses inspirées de la géométrie Reimannienne.

3 Etude expérimentale

Dans cette étude, 6 modèles sont utilisés : BERT et RoBERTa, SBERT et SRoBERTa (Reimers et Gurevych, 2019), ainsi que SBERT-CT et SRoBERTa-CT (Carlsson et al., 2021), réentraînés sur une fonction objectif non supervisée appelée *Contrastive Tension*. On utilise les versions larges avec $b = 24$ et $d = 1024$. Ces modèles sont comparés à plusieurs références : BOW pondéré avec TFIDF ainsi que Word2vec et GloVe avec $d = 300$. Les expériences de *clustering* sont réalisées en utilisant l'algorithme k-means, à l'exception du BOW pour lequel Spherical k-means (Dhillon et Modha, 2001) est préféré, conduisant à des résultats significativement meilleurs. Nous effectuons un *clustering* sur 30 initialisations et gardons celle qui fournit la valeur la plus élevée d'inertie intra-classes. Afin d'évaluer la qualité des résultats de *clustering*, nous nous appuyons sur l'information mutuelle normalisée (NMI) (Strehl et Ghosh, 2002) qui est une mesure externe souvent utilisée pour évaluer des résultats de *clustering*.

Jeux de données Les données utilisées sont décrites dans la Table 1. Nous utilisons classic3 et classic4 de l'Université de Cornell, BBC news de Greene et Cunningham (2006) et des extraits aléatoires de DBPedia (Lehmann et al., 2015) et AG-news (Zhang et al., 2015) de taille 12,000 et 8,000 respectivement.

	classic3	classic4	BBC	DBPedia	AG-news
Taille	3,891	7,095	2,225	12,000	8,000
Classes / Équilibre	3 / 0.71	4 / 0.32	5 / 0.76	14 / 0.92	4 / 0.97

TAB. 1 – Description des *datasets*. Équilibre = ratio entre la plus petite et la plus grande classe.

3.1 Clustering de documents

La Table 2 montre les performances obtenues en utilisant l'approche tandem avec différentes représentations. On peut observer ce qui suit :

— Le modèle BOW montre ses limites face à Word2vec et GloVe quand UMAP est utilisé. Word2vec et GloVe affichent même des résultats compétitifs par rapport aux Transformeurs sauf dans le cas de DBpedia et AG-news pour lesquels les classes sont mal séparées.

— L'approche basée sur UMAP avec le modèle RoBERTa semble être le choix le plus judicieux pour le clustering avec toutes les stratégies (last, last2, all). Il affiche de meilleures performances par rapport à BERT, avec et sans réentraînement.

— L'utilisation de seulement quelques composants de l'ACP ne modifie pas considérablement les résultats et améliore même la qualité du regroupement lorsque le *whitening* est utilisé (PCA_w), bien que cela reste en dessous de l'utilisation d'UMAP.

— Ni les approches supervisées (Reimers et Gurevych, 2019) ni les approches non supervisées (Carlsson et al., 2021) n'apportent d'amélioration au *clustering* et sont même surpassées par les modèles pré-entraînés. Par exemple, si on examine la dernière couche, comme le

Représentation		classic3				classic4				BBC				DBPedia				AG-news			
		orig.	pca	pca_w	umap	orig.	pca	pca_w	umap	orig.	pca	pca_w	umap	orig.	pca	pca_w	umap	orig.	pca	pca_w	umap
Bag-Of-Words		95.2				68.9				81.0				71.4				48.7			
Word2vec		86.7	86.5	91.1	96.2	22.8	22.8	45.9	75.0	79.6	79.1	81.8	88.4	66.8	61.4	60.6	71.7	55.7	54.6	46.9	59.5
GloVe		88.7	88.3	89.6	96.2	54.7	54.2	65.1	73.2	73.8	72.7	79.4	87.8	72.5	63.7	63.0	74.9	52.9	52.4	50.5	55.9
BERT	last	93.3	93.1	94.8	97.1	20.3	20.1	55.1	72.8	76.7	75.6	66.9	77.9	53.3	45.1	43.7	55.8	0.2	0.2	0.2	19.6
	last2	93.3	92.8	95.0	95.7	55.2	55.0	57.3	73.2	77.0	76.2	64.1	77.3	47.4	44.3	44.6	55.6	18.1	17.7	21.0	19.1
	all	90.0	89.9	94.8	96.4	68.0	67.7	71.2	74.4	78.5	76.3	79.7	86.7	67.7	62.0	61.3	71.8	48.5	43.6	47.4	55.6
SBERT	last	87.3	86.9	88.8	93.5	60.7	59.9	62.6	67.7	79.7	72.4	79.7	81.0	37.5	27.2	26.9	39.7	19.8	18.6	24.9	38.8
	last2	88.7	87.7	89.7	94.1	60.4	59.7	62.8	67.6	78.8	75.6	81.4	81.6	43.5	30.1	30.2	40.2	26.9	24.9	27.0	38.7
	all	89.5	89.0	90.8	95.1	46.2	45.8	66.3	61.8	69.4	68.2	67.3	82.7	50.5	49.2	54.9	54.4	35.5	34.1	33.6	41.8
SBERT-CT	last	91.2	90.4	93.2	96.1	66.1	65.0	67.0	68.3	80.7	76.9	80.7	81.4	51.7	37.8	40.0	56.6	41.5	39.8	42.3	53.3
	last2	90.7	90.4	93.0	95.9	66.4	65.9	67.9	69.9	82.1	79.3	84.5	82.3	62.9	41.6	43.3	60.2	43.9	43.4	44.9	53.3
	all	88.7	88.3	90.8	94.9	64.3	63.9	67.5	71.7	74.8	74.2	74.9	83.6	62.9	54.5	57.3	71.4	51.2	49.8	50.6	57.3
RoBERTa	last	91.5	90.8	95.9	98.3	71.9	71.1	72.2	75.1	87.5	86.3	78.0	90.7	68.3	57.3	59.9	68.5	50.0	46.6	51.8	50.9
	last2	89.0	89.1	95.8	98.0	55.7	55.3	71.9	74.9	87.0	85.0	89.1	89.7	71.7	63.0	64.0	68.0	53.7	50.8	56.3	55.2
	all	86.4	86.1	92.8	95.6	51.0	50.5	69.5	73.7	74.1	73.5	83.8	89.2	69.6	60.5	64.3	71.4	51.3	49.8	56.6	53.3
SRoBERTa	last	84.8	83.6	86.7	91.0	63.3	60.7	62.1	65.7	57.1	55.7	63.4	72.1	62.2	40.7	37.6	66.0	33.6	30.6	36.9	51.4
	last2	84.6	84.2	87.9	91.7	63.8	62.1	64.5	65.5	56.3	56.6	63.6	72.2	65.5	49.0	46.6	69.2	34.3	33.0	42.1	52.0
	all	86.1	85.9	90.1	95.1	64.5	64.2	66.9	71.7	70.0	68.9	75.3	83.3	67.0	61.5	60.8	68.9	55.2	53.8	57.9	52.5
SRoBERTa-CT	last	90.4	90.3	94.4	97.1	67.3	66.9	69.1	70.6	83.7	81.9	84.1	83.8	68.7	57.3	55.1	63.8	40.3	37.5	48.6	55.4
	last2	90.9	90.9	93.7	97.3	67.7	67.5	69.4	70.1	84.3	82.4	84.8	85.3	70.0	64.0	62.5	64.6	44.6	41.7	53.5	56.0
	all	89.6	89.0	93.4	96.6	68.6	68.2	69.7	72.2	84.8	81.8	86.3	89.5	71.1	63.1	64.4	69.3	57.5	56.0	60.3	58.7

TAB. 2 – Scores de *clustering* (NMI en %) obtenus en utilisant l'approche tandem sur différentes représentations textuelles. Trois techniques de RD sont utilisées : PCA, PCA_w et UMAP (toutes avec $d' = 10$) et sont comparées aux représentations "originales" (sans post-traitement).

font Reimers et Gurevych (2019), SRoBERTa fonctionne nettement moins bien que RoBERTa presque dans tous les cas. Cela peut être dû au fait que nous traitons de longs documents alors que SBERT et SRoBERTa sont entraînés sur des phrases courtes.

La Figure 2 montre la distribution de la NMI obtenue avec différentes approches. On observe tout d'abord l'avantage d'utiliser toutes les couches en termes de qualité de clustering et aussi en termes de robustesse. On remarque en effet que le score de NMI est beaucoup moins dépendant de l'initialisation lorsqu'on utilise toutes les couches. La Table 2 montre des résultats similaires entre l'utilisation de la dernière couche et la combinaison des deux dernières, avec un léger avantage pour celle-ci. Cela suggère que chaque couche apporte des informations précieuses pour le *clustering*. De plus, nous pouvons voir sur la Figure 2 à quel point les représentations fournies par UMAP sont robustes, en particulier par rapport à PCA_w qui présente une variance beaucoup plus élevée avec une médiane nettement plus faible. Cela montre la fiabilité et la robustesse de l'utilisation d'UMAP dans le cadre de l'approche tandem.

3.2 Visualisation des données

La Figure 3 montre des projections 2D obtenues en utilisant différentes techniques de RD avec $d' = 2$. Le score NMI donné est calculé à l'aide des étiquettes réelles et le score « Agr » est une mesure non supervisée proposée dans (France et Akkucuk, 2021) qui quantifie la concordance entre l'espace original et l'espace latent. Il est calculé comme suit :

$$Agr_k = \frac{1}{kn} \sum_{i=1}^{n} \left[a_{ik} - \frac{k}{n-1} \right]$$

où k est la taille du voisinage que l'on fait varier entre 1 et 100 tandis que a_{ik} représente les éléments en commun entre les k premières colonnes des matrices de rang $\mathbf{N}_{\mathbf{X}(n\times n)}$ et $\mathbf{N}_{\mathbf{Y}(n\times n)}$

contenant dans chaque ligne i les indices des individus du plus proche au plus éloigné selon la métrique, en utilisant respectivement $\mathbf{X}$ et $\mathbf{Y}$.

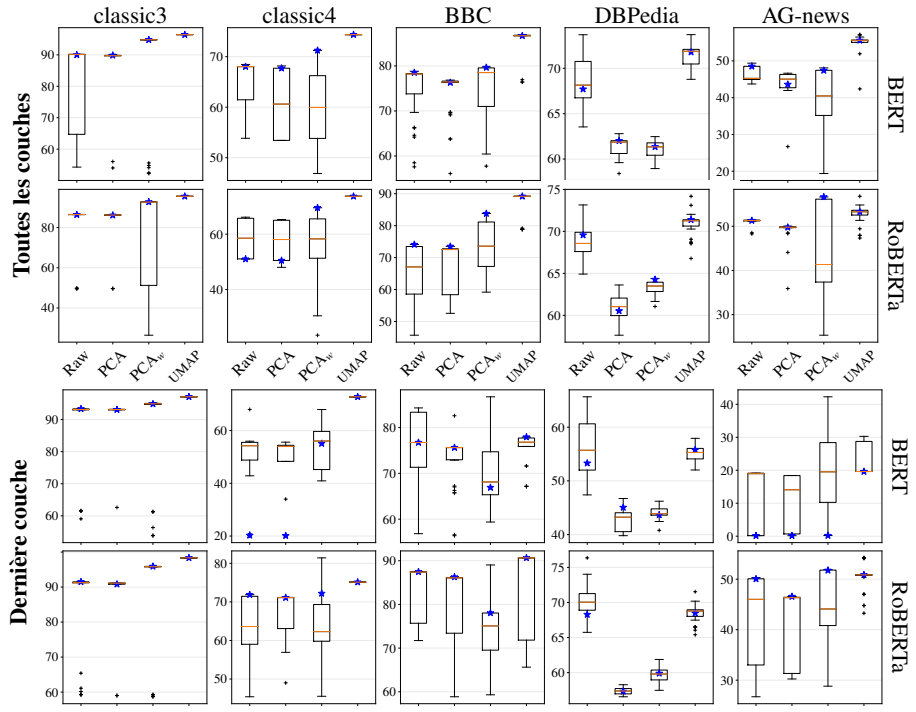

FIG. 2 – Répartition de la NMI sur 30 initialisations. L'étoile correspond au score de la solution sélectionnée avec le critère de k-means (valeurs du Tableau 2) et la ligne orange à la médiane.

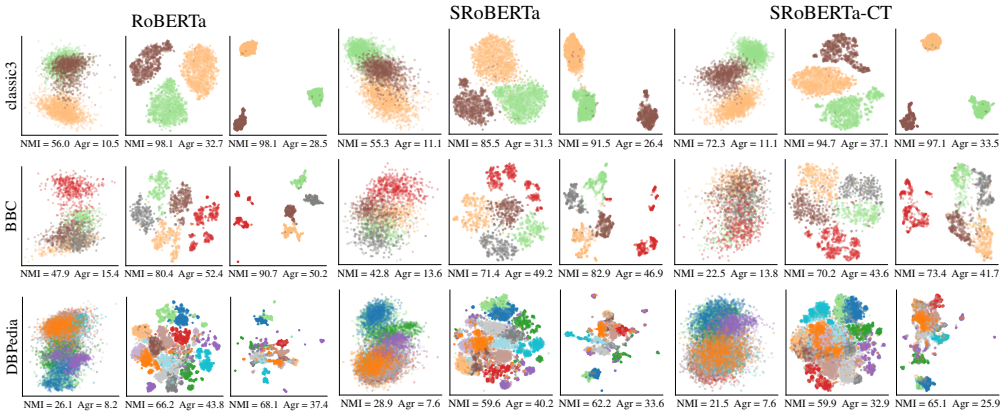

FIG. 3 – Projections 2D obtenues respectivement par PCA$_w$, t-SNE et UMAP. Le score NMI (%) correspond aux performances de clustering appliquées à la version réduite $\mathbf{Y}_{(n\times d')}$ avec $d' = 2$. Les points de données sont colorés en fonction des classes réelles.

Comme il est impossible de régler les paramètres de t-SNE et d'UMAP, nous fixons la perplexité et le nombre de voisins à 15 et conservons la distance euclidienne. Dans l'ensemble, on observe la différence en termes de séparabilité. Ainsi, UMAP est capable de mieux séparer les classes, suivi de t-SNE puis de l'ACP qui, sans surprise, montre une très mauvaise séparabilité. Ceci est corroboré par le score NMI, qui est toujours plus élevé pour UMAP. De plus, le score « Agr » est beaucoup plus faible pour l'ACP. Par contre, il est plus élevé pour t-SNE que pour UMAP. Cela signifie que t-SNE conserve davantage la structure originale des données. Cependant, t-SNE ne permet pas une meilleure séparabilité des classes, ce qui fait d'UMAP un bon compromis entre l'*embedding* des données et la séparabilité. Cela suggère également que la distorsion supplémentaire apportée par UMAP est bénéfique pour le *clustering*.

4 Conclusion

Comme les plongements MLPT ont montré des performances médiocres lorsqu'ils sont utilisés en entrée de tâches d'apprentissage, plusieurs approches ont été proposées afin de les améliorer, le plus souvent celles-ci sont basées sur le ré-entrainement des modèles. Dans cet article, nous évaluons l'impact de telles approches sur deux tâches : le *clustering* de texte et la visualisation. Ainsi, nous montrons que le ré-entraînement, bien que bénéfique pour la tâche de similarité sémantique, n'apporte aucune amélioration significative dans l'une ou l'autre de nos deux tâches. Le post-traitement, cependant, en plus d'être plus simple, montre des améliorations bien plus impressionnantes. Plus spécifiquement, nous soulignons le potentiel d'UMAP qui, même avec un sous-espace de faible dimension, montre une amélioration significative des performances de *clustering* dans le cadre d'une approche tandem.

Remerciements. Ce travail a été financé par la Caisse des Dépôts et Consignations (CDC), l'ANRT et l'Idex-Spectrans d'Université Paris Cité.

Références

Ait-Saada, M., F. Role, et M. Nadif (2021). How to leverage a multi-layered transformer language model for text clustering : an ensemble approach. In *CIKM*, pp. 2837–2841.

Carlsson, F., A. C. Gyllensten, E. Gogoulou, E. Y. Hellqvist, et M. Sahlgren (2021). Semantic re-tuning with contrastive tension. In *ICLR 2021*.

Cheng, X. (2021). *Dual-View Distilled BERT for Sentence Embedding*, pp. 2151–2155. New York, NY, USA : Association for Computing Machinery.

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

Dhillon, I. S. et D. S. Modha (2001). Concept decompositions for large sparse text data using clustering. *Machine learning 42*(1), 143–175.

France, S. L. et U. Akkucuk (2021). A review, framework, and R toolkit for exploring, evaluating, and comparing visualization methods. *The Visual Computer 37*(3), 457–475.

Gao, T., X. Yao, et D. Chen (2021). Simcse : Simple contrastive learning of sentence embeddings. In *EMNLP (1)*, pp. 6894–6910.

Greene, D. et P. Cunningham (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *ICML*, pp. 377–384.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web 6*(2), 167–195.

Li, B., H. Zhou, J. He, M. Wang, Y. Yang, et L. Li (2020). On the sentence embeddings from pre-trained language models. In *EMNLP*, pp. 9119–9130.

Liu, F., I. Vulić, A. Korhonen, et N. Collier (2021). Fast, effective, and self-supervised : Transforming masked language models into universal lexical and sentence encoders. In *EMNLP*, pp. 1442–1459.

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*.

McInnes, L., J. Healy, et J. Melville (2018). UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pp. 3111–3119.

Pennington, J., R. Socher, et C. Manning (2014). GloVe : Global vectors for word representation. In *EMNLP*, Doha, Qatar, pp. 1532–1543.

Raunak, V., V. Gupta, et F. Metze (2019). Effective dimensionality reduction for word embeddings. In *Proceedings of RepL4NLP-2019*, pp. 235–243.

Reimers, N. et I. Gurevych (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*, pp. 3980–3990.

Strehl, A. et J. Ghosh (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research 3*(Dec), 583–617.

Su, J., J. Cao, W. Liu, et Y. Ou (2021). Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv :2103.15316*.

Yan, Y., R. Li, S. Wang, F. Zhang, W. Wu, et W. Xu (2021). ConSERT : A contrastive framework for self-supervised sentence representation transfer. In *ACL-IJCNLP*, pp. 5065–5075.

Zhang, X., J. Zhao, et Y. LeCun (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657.

Zhang, Y., R. He, Z. Liu, K. H. Lim, et L. Bing (2020). An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of EMNLP*, pp. 1601–1610.

Summary

Dense text representations are gaining great interest in several supervised tasks but much less is known about how suitable they are when dealing with an unlabeled dataset. In this paper, we investigate the use of such representations in unsupervised tasks: document clustering and visualization. For that, we propose the use of a *tandem approach* based un UMAP, showing that we can do better than the fine-tuning approaches usually proposed in the literature.