Alignement non supervisé d'embeddings de mots dans le domaine biomédical

Félix Gaschi, Parisa Rastin, Yannick Toussaint

Alignement non supervisé d'embeddings de mots dans le domaine biomédical

Félix Gaschi*,** Parisa Rastin** Yannick Toussaint**

*SAS Posos, 53 rue de la Boétie, 75008 Paris
prenom@posos.fr,
www.posos.co
**LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy
nom.prenom@loria.fr
www.loria.fr

Résumé. Notre objectif est de créer un alignement non supervisé et multilingue d'embeddings de mots (ou plongements lexicaux) basés sur des corpora de domaines différents. Plus précisément, nous cherchons à aligner un embedding cible anglais du domaine biomédical avec un embedding source du domaine général d'une autre langue, puisque les textes à traiter sont dans diverses langues (français, espagnol...) et que le vocabulaire du domaine biomédical est essentiellement disponible en anglais. Notre méthode pour aligner deux embeddings de domaines et langages différents repose sur un autre embedding pivot de même domaine que la source et de même langage que la cible. Notre méthode aligne d'abord les embeddings de même domaine pour créer un dictionaire qui sert ensuite à aligner les embeddings de domaines et languages distincts. Elle est évaluée sur une tâche de traduction du domaine biomédical dans plusieurs langues. Bien que notre algorithme ne dépasse pas les méthodes d'alignement entre embeddings de même domaine, elle dépasse ces mêmes méthodes appliquées à des embeddings de domaines différents. Ce travail préliminaire montre qu'aligner des embeddings de domaines différents est possible de manière non supervisé.

Mots-clés : embeddings de mots, traitement automatique du langage, multilingue, apprentissage non supervisé

1 Introduction

Les embeddings de mots (ou plongements lexicaux) fournissent des représentations utiles pour de nombreuses tâches de machine learning (Mikolov et al., 2013b; Pennington et al., 2014). Les embeddings ont été généralisés dans un context multilingue en un concept d'embedding multilingue non supervisé (EMN) (Mikolov et al., 2013a; Artetxe et al., 2017; Zhang et al., 2017a; Lample et al., 2018b; Artetxe et al., 2018). Les données spécifiques à un domaine précis, comme les publications scientifiques, peuvent être rares pour des langues autres que l'anglais. En particulier dans le domaine biomédical, les quantités de ressources disponibles en anglais (plus de 21 millions d'articles sur PubMed) dépassent largement celles disponibles

dans d'autres langues. C'est pourquoi nous avons pour but de créer des représentations multilingues qui alignent un embedding cible spécifique au domaine biomédical en anglais avec un embedding source issu du domaine général dans une autre langue. Notre objectif n'est pas tant de surpasser l'état de l'art en termes d'alignement non supervisé d'embedding, mais plutôt d'explorer la possibilité d'aligner de manières non supervisée des embeddings de domaines distincts.

Les EMN visent à apprendre des représentations multilingues uniquement à partir de données monolingues. Les méthodes existantes pour construire de tels embeddings multilingues reposent souvent sur une transformation linéaire orthogonale et montrent de bons résultats tant que les données sont issues du même domaine (Mikolov et al., 2013a; Artetxe et al., 2017; Zhang et al., 2017a; Lample et al., 2018b; Artetxe et al., 2018). Cependant, elles ont tendance à produire de moins bons résultats avec des données issues de domaines différents (Søgaard et al., 2018). Ces méthodes s'appuient sur l'hypothèse que les espaces métriques des embeddings de mots sont approximativement isométriques (Mikolov et al., 2013a; Zhang et al., 2017a; Lample et al., 2018a).

Bien que Søgaard et al. (2018) aient montré que cette similarité isométrique approximative n'est pas conservée pour des embeddings entraînés sur des domaines différents, l'hypothèse initiale peut tout de même rester valable pour des sous-ensembles bien choisis. Nous apportons en effet une amélioration sur les alignements d'embeddings de domaines différents en s'appuyant sur un embedding pivot de la même langue que l'embedding cible et du même domaine que l'embedding source. Nous mettons en pratique notre méthode pour aligner un embedding source entraîné sur Wikipedia en français sur un embedding cible entraîné sur PubMed en anglais. Notre pivot est donc un embedding entraîné sur le domaine général en anglais dans nos expériences.

2 Travaux connexes

Embedding cross-lingues et non supervisés. Peu après avoir introduit les modèles Skip-gram et de Continuous Bag-of-Words (CBOW) (Mikolov et al., 2013c) pour apprendre des embeddings de mots, Mikolov et al. (2013a) proposent d'aligner des embeddings de différentes langues dans un espace partagé avec l'aide d'un dictionaire bilingue. Avec l'apparition de l'auto-apprentissage itératif (Artetxe et al., 2017), qui alterne entre l'apprentissage d'un alignement et celui d'un dictionaire, les méthodes d'alignement d'embeddings progressent et requièrent alors de moins en moins de paires de mot dans le dictionnaire initial d'entraînement. Finalement, des méthodes entièrement non supervisées reposent sur l'apprentissage adversaire (Zhang et al., 2017a; Lample et al., 2018b) et des heuristiques d'initialisation (Artetxe et al., 2018). Ces dernières s'appuient en grande partie sur une transformation linéaire orthogonale appliquée à des embeddings normalisés (Xing et al., 2015; Smith et al., 2017) assurant ainsi une invariance des distances entre les mots au sein d'une même langue.

Limites de la contrainte orthogonale. Søgaard et al. (2018) démontrent que les EMNs qui s'appuient sur une transformation orthogonale ont besoin de trois conditions pour être efficaces : (1) les langues à aligner doivent être morphologiquement similaires, (2) les corpora d'entraînement monolingues doivent être issus du même domaine, et (3) le même modèle doit être utilisé (un embedding CBOW en espagnol ne pourra pas être aligné avec un embedding

Skip-gram en anglais). En effet, utiliser une transformation orthogonale implique que les embeddings soient à-peu-près isométriques (Mikolov et al., 2013a; Zhang et al., 2017a; Lample et al., 2018b). Søgaard et al. (2018) montrent, grâce à une comparaison de valeurs propres, que les graphes de voisinage des mots ne sont pas isomorphiques. Zhang et al. (2017b) montrent que la distance de Wasserstein entre des embeddings est corrélée à la similarité typologique entre les langues. Patra et al. (2019) utilisent une autre métrique basée sur des homologies persistentes pour évaluer la similarité entre des embeddings de mots de langues différentes. Ces observations mènent finalement à la création de plusieurs méthodes utilisant une contrainte d'orthogonalité faible (Zhang et al., 2017a; Patra et al., 2019). Pour tenir compte de variations locales de la densité des embeddings, un critère local de mise à l'échelle (Cross-domain Similarity Local Scaling ou CSLS) (Lample et al., 2018a) est souvent utilisé dans les modèles basés sur des transformations orthogonales (Lample et al., 2018a; Artetxe et al., 2018; Joulin et al., 2018).

Dépasser les limites des embeddings cross-lingues. Søgaard et al. (2018) montrent qu'un des modèles basés sur des tranformations orthogonales (Lample et al., 2018b) obtient une précision proche de zéro lorsqu'on aligne des embeddings de domaines différents et pour des langues éloignées. De plus, ils montrent qu'il est possible d'augmenter la précision en utilisant les mots écrits de manières identiques dans deux langues comme signal de supervision faible. Cependant, pour autant que nous le sachions, avec l'exception de cette supervision faible, il n'y a pas eu de travaux proposant une méthode pour aligner des embeddings de domaines différents de manière non supervisée. Des méthodes semi-supervisées avec une contrainte orthogonale faible ont été proposées (Patra et al., 2019). Shakurova et al. (2019) ont appliqué avec succès des méthodes d'alignement d'embeddings sur des domaines spécifiques, mais toujours pas entre domaines distincts. Si l'alignement entre des embeddings de domaines différents a, semble-t-il, suscité peu d'intérêt, le cas des langues distantes ou faibles en ressources textuelles a été plus largement étudié. À titre d'exemple, Nakashole (2018) développe un modèle basé sur les voisinages pour améliorer l'alignement entre des embeddings de langues distantes et Nakashole et Flauger (2017) utilisent un langage riche en donnée monolingues disponibles comme pivot pour aligner des langues pour lesquelles moins de ressources sont disponibles, ce qui a en partie inspiré notre idée d'utiliser un troisième embedding intermédiaire pour l'alignement entre domaines différents.

La plupart des approches pour aligner des langues distantes s'appuient sur une contrainte orthogonale faible voire s'en débarrasse complètement, ce qui est justifié par l'apparente absence d'isométrie entre les embeddings de langues distantes. Nous pensons que cette absence profonde d'isométrie ne s'applique pas dans les cas où les embeddings proviennent simplement de domaines différents. Par la suite, nous détaillons brièvement le lien entre transformation orthogonale et isométrie et nous faisons l'hypothèse que, bien que cette condition d'isométrie ne soit pas valable sur l'intégralité des embeddings de domaines différents elle pourrait l'être pour des sous-ensembles bien choisis de ces embeddings.

3 Considérations sur l'isométrie entre des sous-ensembles d'embeddings

Les méthodes supervisées pour l'apprentissage d'embeddings multilingues construisent généralement une transformation linéaire entre les représentations des entrées d'un dictionaire bilingue (Mikolov et al., 2013a). En suivant le formalisme de Lample et al. (2018a), nous avons :

$$W^* = \arg \min_{W \in \mathcal{O}_d} ||AW - B|| \quad (1)$$

Avec $A \in \mathbb{R}^{N \times d}$ et $B \in \mathbb{R}^{N \times d}$ les représentations de dimension d des entrées du dictionnaire bilingue dans les embeddings source et cible. W^* est la transformation linéaire apprise, dans $\mathcal{O}_d$, l'ensemble des matrices orthogonales. La transformation est choisie orthogonale pour conserver les distances au sein des embeddings monolingues. Comme une transformation orthogonale préserve les distances, cela signifie également que la fonction $f : \mathbb{R}^d \to \mathbb{R}^d$ telle que $f(A_i) = B_i$ doit être une isométrie pour que la méthode fonctionne, c'est-à-dire que pour toute paire de lignes issus de A, $A_i, A_j \in \mathbb{R}^d$, on doit avoir $l(A_i, A_j) \approx l(B_i, B_j)$ avec B_i, B_j les lignes de B correspondantes et l une distance.

Pour les méthodes entièrement non supervisées, le dictionnaire doit être appris en même temps que la transformation orthogonale. Pour deux ensembles de vecteurs donnés, représentés par les matrices X pour la source et Z pour la cible, on doit trouver l'application $f : X \to Z$ qui est un dictionnaire bilingue et doit être une quasi-isométrie. Cela signifie que X et Z définissent des espaces métriques qui sont eux-même quasi-isométriques.

Alors qu'une telle fonction peut exister dans le cas mono-domaine, ça n'est pas le cas dans un cas multi-domaine (Søgaard et al., 2018) ; les mots spécifiques à un domaine dans un embedding peuvent ne pas trouver de traduction dans l'embedding d'un autre domaine.

Nous faisons l'hypothèse qu'il est possible d'améliorer les méthodes non supervisées pour la création d'embeddings multilingues et multi-domaines en essayant d'aligner des sous-ensembles bien choisis de chaque vocabulaire. Comme schématisé en figure 1, cet "alignement partiel" des embeddings pourrait être utile pour certaines tâches spécifiques à un domaine, comme le domaine biomédical.

Pour valider cette hypothèse, nous réalisons deux expériences. En suivant les travaux de Patra et al. (2019), nous utilisons la distance de bottleneck pour mesurer à quel point des sous-ensembles d'embeddings de différents domaines sont proche de l'isométrie. Les détails du calcul de la distance de bottleneck sont détaillés dans la section suivante. Ensuite, nous proposons une méthode simple d'alignement non supervisée basé sur une transformation orthogonale qui s'applique à des embeddings de domaines différents. Nous la détaillons dans la section 5.

4 Une métrique pour la quasi-isométrie

Le calcul de la distance de bottleneck nous permet d'évaluer à quel point deux embeddings s'éloignent de l'isométrie. L'utilisation de la distance de bottleneck pour un tel usage est proposée par Patra et al. (2019).

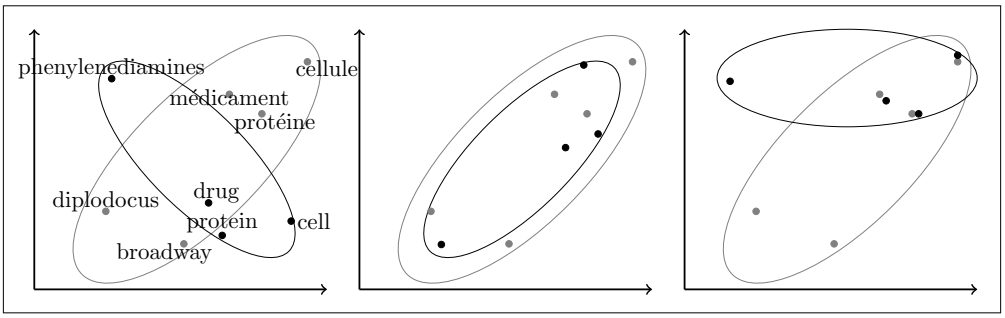


FIG. 1: Exemple jouet. Quand on aligne un domaine (gris) avec un autre (noir), les embeddings initialement non-alignés (gauche) s'alignent mal lorsqu'on essaye d'aligner tous les mots (centre). Notre objectif est de les aligner partiellement (droite).

Les embeddings sont normalisés comme le font Artetxe et al. (2018) : l1-normalisation, centrés à la moyenne, puis l1-normalisation à nouveau. La l1-normalisation permet notamment d'avoir une équivalence entre la distance cosinus et la distance l2.

Une mesure de la distance entre deux espaces métriques $(\mathcal{X}, d)$ et $(\mathcal{Y}, d)$ peut-être donnée par la distance de Hausdorff qui est la distance maximale entre les paires de plus proches voisins :

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = \max\left(\sup_{x\in\mathcal{X}} \inf_{y\in\mathcal{Y}} d(x,y), \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} d(x,y)\right) \tag{2}$$

Cette distance de Hausdorff mesure à quel point deux espaces métriques "coïncident". Pour mesurer à quel point ils sont isométriques, il faut connaître la distance de Hausdorff minimale entre toutes les transformations isométriques possibles de $\mathcal{X}$ et $\mathcal{Y}$. C'est la distance de Gromov-Hausdorff :

$$\mathcal{GH}(\mathcal{XY}) = \min_{f,g} \mathcal{H}(f(\mathcal{X}), g(\mathcal{Y})) \tag{3}$$

Avec f et g des isométries. Calculer cette distance n'est pas réalisable dans notre cas, mais on peut l'approximer en utilisant la distance bottleneck (Chazal et al., 2009). Pour la calculer, on construit d'abord les diagrammes de persistance de premier ordre sur les complexes de Vietoris-Rips des deux espaces métriques. Concrètement, pour une distance t qui varie de 0 à $+\infty$, on on construit des simplexes pour chaque ensemble de points qui se trouvent à une distance inférieure à t. En faisant croître t, on commence avec autant de composante connexes que de points dans l'espace métrique, et on diminue leur nombre en les fusionnant petit à petit. Le diagramme de persistance est constitué des points $(t_{\text{birth}}, t_{\text{death}})$ pour chaque composante connexe qui est apparue à $t = t_{\text{birth}}$ et a été fusionnée avec une autre à $t = t_{\text{death}}$.

Soit nos deux diagrammes de persistance f et g, la distance de bottleneck est le minimum pour toute bijection γ entre f et g de la distance de Manhattan maximale entre les éléments de f et leur image par γ :

$$\mathcal{B}(f,g) = \inf_{\gamma}(\sup_{u\in f} ||u - \gamma(u)||_\infty) \tag{4}$$

5 Méthode d'alignement avec pivot

Søgaard et al. (2018) utilisent les mots identiques comme signal de supervision faible. De manière similaire, nous voulons construire un dictionnaire de manière non supervisée qui pourrait être utilisé pour apprendre une transformation satisfaisant l'equation 1.

Les embeddings source et cible sont de domaines et langages différents. Dans nos expériences, il s'agira par exemple d'un embedding du domaine général entraîné sur Wikipedia en français et d'un embedding du domaine biomédical entraîné sur PubMed en anglais. En plus de ces deux embeddings, on dispose d'un embedding supplémentaire intermédiaire, de même domaine que la source et de même langage que la cible (Wikipedia en anglais dans notre exemple).

La méthode que nous proposons prend comme entrée ces trois embeddings entraînés avec FastText (Bojanowski et al., 2017) : X l'embedding source, Z l'embedding cible et Y le pivot, de même domaine que X et de même langage que Z. Tous les embeddings sont normalisés de la même manière que Artetxe et al. (2018) : normalisés par la norme, centrés à la moyenne, puis normalisé par la norme à nouveau.

D'abord, nous alignons X et Y (la source et le pivot) de manière non supervisée à l'aide de l'algorithme VecMap (Artetxe et al., 2018). X et Y sont des embeddings de même domaines sur lesquels des méthodes comme VecMap ont fait leurs preuves. VecMap apprends d'abord un dictionnaire initial en représentant chaque mot par sa similarité à ses plus proches voisins et construit les paires d'un dictionnaire bilingue par simple recherche de plus proche voisin dans cette nouvelle représentation. Ensuite, une transformation orthogonale et un dictionnaire sont affinés alternativement dans une boucle d'auto-apprentissage.

À partir des embeddings source et pivot de même domaines maintenant alignés grâce à VecMep, $\tilde{X}$ et $\tilde{Y}$, on peut maintenant inférer un dictionnaire. Mais nous pouvons restreindre ce dictionnaire à l'intersection du vocabulaire entre le pivot $\tilde{Y}$ et la cible Z puisque ce sont deux embeddings de même langage. Plus précisément, pour chaque mot de la source $\tilde{X}$ nous constituons une paire avec son plus proche voisin dans l'embedding pivot aligné $\tilde{Y}$ si le mot correspondant se trouve aussi dans le vocabulaire de la cible. Et pour chaque mot qui est à la fois dans Z et $\tilde{Y}$ nous constituons une paire avec le mot de $\tilde{X}$ le plus proche. On peut ainsi construire les matrices A et B de notre dictionnaire bilingue et l'injecter dans l'équation 1 pour aligner la source et la cible. La solution est donnée par la décomposition en valeur singulière (SVD) de $A^\top B = USV^\top$ en écrivant $W^* = UV^\top$.

6 Expériences et résultats

Dans un premier temps, nous voulons mesurer à quel point différents sous-ensembles de paires d'embeddings sont proche de l'isométrie. Dans un second temps, nous évaluons la méthode proposée sur une tâche de traduction spécifique au domaine biomédical.

Dans ce qui suit, l'embedding source est systématiquement un embedding entraîné sur Wikipédia dans une langue autre que l'anglais (français, allemand, espagnol ou portugais). L'embedding cible est entraîné quant à lui sur PubMed [1] pour l'alignement *multi-domaine*,

1. un ensemble d'environ 21 million d'articles scientifiques du domaine biomédical, écrits en anglais, fournis par la U.S. National Library of Medicine `https://www.nlm.nih.gov/databases/download/pubmed_medline.html`

	source-cible	source-pivot
20 000 mots		
fréquents	0.1532	0.0626
MeSH	0.0638	0.0806
Dictionnaire en sortie de VecMap		
dico.	0.1413	0.0763

TAB. 1: Distance de bottleneck entre différents sous-ensembles d'embeddings en français et en anglais.

et sur Wikipedia en anglais pour l'alignement *mono-domaine*. L'embedding intermédiaire de notre méthode est entraîné sur Wikipedia en anglais.

Mesure de la distance bottleneck. Nous avons fait l'hypothèse que certains sous-ensembles d'embeddings de domaines différents peuvent être malgré tout à-peu-près isométriques. Pour s'assurer de la validité de cette hypothèse, nous utilisons la distance de bottleneck comme définie plus haut (section 4). Plus la distance de bottleneck est proche de zéro, plus la paire d'espaces métriques évaluée est proche d'être isométrique.

Nous reportons nos résultats dans le tableau 1. Les premiers 20 000 mots les plus fréquents (*fréquents* dans le tableau) de deux embeddings de même domaine, la source et la cible (*src-cib*), sont plus proches en terme de distance de bottleneck que les 20 000 mots les plus fréquents de deux embeddings de même domaine, la source et l'intermédiaire (*src-pivot*). Cela expliquerait pourquoi des méthodes comme VecMap basées sur une transformation orthogonale fonctionnent mieux dans le cas mono-domaine que dans le cas multi-domaine, puisqu'il s'agit de construire un dictionnaire sur ces 20 000 mots les plus fréquents.

À l'inverse, quand on évalue cette même distance entre les 20 000 mots les plus fréquents du MeSH [2], une ontologie biomédicale en anglais avec sa traduction en français [3], on obtient le résultat inverse. Les ensembles des représentation vectorielles des mots les plus fréquents dans le vocabulaire biomédical semblent plus proches en terme de distance de bottleneck quand on utilise l'embedding spécifique au domaine en question.

Il semblerait donc que l'hypothèse d'isométrie puisse toujours être valable en choisissant soigneusement les sous-ensembles du vocabulaire que l'on cherche à aligner. Avec notre méthode, qui s'appuie sur un alignement mono-domaine et l'intersection des vocabulaires pour générer un dictionnaire spécifique au domaine recherché, nous fournissons une première heuristique simple pour montrer qu'une sélection du vocabulaire peut améliorer les résultats sur des alignements multi-domaines. Cependant, nous montrons également dans le tableau 1 que la distance de bottleneck évaluée sur le vocabulaire du dictionaire bilingue inféré grâce à VecMap (*dico.*) est plus grande dans le cas multi-domaine. Le sous-ensemble de vocabulaire que nous avons sélectionné n'est donc pas optimal et pourrait donc être amélioré.

2. fournie par le NLM `https://www.nlm.nih.gov/mesh/meshhome.html`
3. fournie par l'INSERM `http://mesh.inserm.fr/FrenchMesh/`

	fr-en	es-en	de-en	pt-en
Wikipedia xx avec PubMed en non supervisé				
VecMap	0.093	0.062	0.068	0.065
MUSE	0.053	0.064	0.055	0.069
WP	0.081	0.081	0.052	0.053
notre méthode	0.382	0.503	0.313	0.460
Wikipedia xx avec PubMed en faiblement supervisé				
VecMap	0.299	0.365	0.254	0.289
Wikipedia xx avec Wikipedia en non supervisé				
VecMap	**0.455**	**0.582**	0.373	**0.555**
MUSE	0.434	0.579	**0.398**	0.532
WP	0.447	0.571	0.363	0.513
soumission UCAM run 3				
UCAM	-	*0.708*	*0.612*	-

TAB. 2: score BLEU-1 sur la tâche de traduction Biomedical WM19.

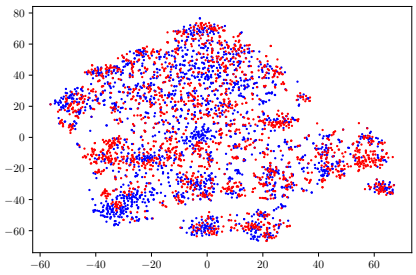
Évaluation de notre méthode. Nous évaluons notre méthode sur l'ensemble de test du dataset Biomedical WMT19[4]. Cette tâche fournie des résumés d'articles PubMed dans divers langages et leur traduction en anglais. Comme notre méthode est une méthode d'alignement de vecteurs de mots et non un modèle de traduction, nous l'utilisons pour traduire les phrases mot-à-mot en la comparant avec d'autres méthodes d'alignement d'embedding, mono-domaines et multi-domaines. On évalue donc les performances à l'aide du score BLEU-1 (Papineni et al., 2002) sur les résumés d'articles entiers.

Nous avons choisi une tâche de traduction plutôt que d'induction de dictionnaire car celle-ci ne tient notamment pas compte des variations morphologiques des mots (Czarnowska et al., 2019) et donne trop d'importance à certain mot comme les noms propres (Kementchedjhieva et al., 2019). De plus, nous pouvons envisager d'utiliser des méthodes d'alignement multi-domaine comme heuristique d'initialisation pour des modèles de traduction non supervisée spécifiques au domaine biomédical, à la manière des travaux de Artetxe et al. (2017); Lample et al. (2018a); Artetxe et al. (2019).

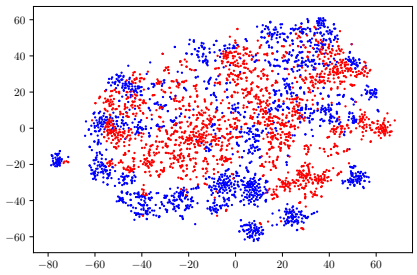
Les résultats sur la tâche de traduction de quatre langues différentes (français, espagnol, allemand, portugais) vers l'anglais sont montrés dans le Tableau 2. Notre méthode est comparée avec les méthodes non supervisées VecMap (Artetxe et al., 2018), MUSE (Lample et al., 2018b) et Wasserstein-Procrustes (WP) (Grave et al., 2018) appliquées à des paires d'embeddings de domaines différents aussi bien qu'à des paires d'embeddings de même domaines. Les résultats de la soumission UCAM (Saunders et al., 2019) au challenge Biomedical WM19 sont aussi donnés à titre de borne supérieure de référence, puisqu'il s'agit d'un modèle de deep learning a priori inégalable avec une simple méthode d'alignement d'embeddings. On présente également les résultats de la méthode faiblement supervisée proposée par Søgaard et al. (2018) basés sur les mots identiques au sein d'une paire de langues.

Dans toutes les langues, notre méthode dépasse les autres méthodes multi-domaines, y compris la méthode faiblement supervisée. En revanche les méthodes mono-domaines appli-

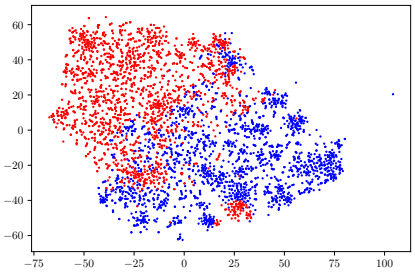
4. http://www.statmt.org/wmt19/biomedical-translation-task.html



(a) VecMap, même domaine.



(b) VecMap, domaines différents.



(c) Notre méthode, domaines différents.

FIG. 2: t-SNE sur des embeddings alignés français (bleu) et anglais (rouge).

quées au domaine général semblent généraliser assez bien au domaine biomédical et obtiennent de meilleures résultats que notre méthode.

Visualisation des alignements. Pour conclure nos expériences, nous avons utilisé une méthode de réduction de dimension, la t-SNE (van der Maaten et Hinton, 2008), pour vérifier qu'un "alignement partiel" comme prévu par notre hypothèse avait bien lieu (cf. figure 1). Nous reportons en figure 2 les résultats d'une telle t-SNE sur les embeddings français (en bleu) et anglais (en rouge) alignés selon trois techniques. L'alignement mono-domaine à l'aide de VecMap des embeddings basés sur Wikipedia en français et en anglais (2a) montre que les nuages de points des deux langues sont globalement superposés et on observe la superposition de certaines grappes. Cette superposition locale est moins, voire pas du tout visible, lorsqu'on utilise la même méthode, VecMap, pour aligner des embeddings de domaines différents (2b). Les nuages de points des deux langues sont toujours globalement superposés, mais peu de grappes semblent se superposer localement. Ce résultat est corrélé au fait que VecMap sur des embeddings de domaines distincts obtient des scores proches de zéro dans notre tâche de traduction. Enfin, nous observons que pour la notre méthode appliquée aux embeddings de domaines différents (2c), les nuages de points ne sont plus globalement alignés. On retrouve ce qui ressemble à l'idée "d'alignement partiel" schématisée dans notre exemple jouet (figure

1). Toutefois, sur la partie où les deux langues se superposent, nous n'observons pas d'alignements significatifs de grappes de points. Cette visualisation tend donc à confirmer l'idée qu'un alignement partiel est possible, mais elle souligne également que notre méthode est perfectible.

7 Conclusion

Nous avons montré que l'hypothèse de quasi-isométrie est encore valable pour des sous-ensembles bien choisis d'embeddings de domaines différents. Nous avons fait la démonstration que les alignements non supervisés basés sur des transformations orthogonales ne sont pas voués à l'échec dans le cas multi-domaine. Cependant, notre méthode reste assez naïve et la distance de bottleneck élevée entre les représentations de notre dictionnaire intermédiaire montrent qu'il y a encore des possibilités d'amélioration.

L'amélioration qu'apportent des méthodes comme la nôtre ou celle faiblement supervisée proposée par Søgaard et al. (2018) par rapport à d'autres sur des embeddings de domaines différents suggèrent que les performances de méthodes d'alignement d'embeddings sont très sensibles à l'initialisation. De futures recherches sont à mener sur les méthodes d'initialisation et leur adaptation à des embeddings de domaines différents.

Références

Artetxe, M., G. Labaka, et E. Agirre (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Vancouver, Canada, pp. 451–462. Association for Computational Linguistics.

Artetxe, M., G. Labaka, et E. Agirre (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia, pp. 789–798. Association for Computational Linguistics.

Artetxe, M., G. Labaka, et E. Agirre (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 194–203. Association for Computational Linguistics.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics 5*, 135–146.

Chazal, F., D. Cohen-Steiner, L. J. Guibas, F. Mémoli, et S. Y. Oudot (2009). Gromov-hausdorff stable signatures for shapes using persistence. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, Goslar, DEU, pp. 1393–1403. Eurographics Association.

Czarnowska, P., S. Ruder, E. Grave, R. Cotterell, et A. Copestake (2019). Don't forget the long tail ! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 974–983. Association for Computational Linguistics.

Grave, E., A. Joulin, et Q. Berthet (2018). Unsupervised alignment of embeddings with wasserstein procrustes.

Joulin, A., P. Bojanowski, T. Mikolov, H. Jégou, et E. Grave (2018). Loss in translation : Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 2979–2984. Association for Computational Linguistics.

Kementchedjhieva, Y., M. Hartmann, et A. Søgaard (2019). Lost in evaluation : Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 3336–3341. Association for Computational Linguistics.

Lample, G., A. Conneau, L. Denoyer, et M. Ranzato (2018a). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lample, G., A. Conneau, M. Ranzato, L. Denoyer, et H. Jégou (2018b). Word translation without parallel data. In *International Conference on Learning Representations*.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Q. V. Le, et I. Sutskever (2013b). Exploiting similarities among languages for machine translation. *CoRR abs/1309.4168*.

Mikolov, T., Q. V. Le, et I. Sutskever (2013c). Exploiting similarities among languages for machine translation.

Nakashole, N. (2018). NORMA : Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 512–522. Association for Computational Linguistics.

Nakashole, N. et R. Flauger (2017). Knowledge distillation for bilingual dictionary induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2497–2506. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, et W.-J. Zhu (2002). Bleu : A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, USA, pp. 311–318. Association for Computational Linguistics.

Patra, B., J. R. A. Moniz, S. Garg, M. R. Gormley, et G. Neubig (2019). Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 184–193. Association for Computational Linguistics.

Pennington, J., R. Socher, et C. D. Manning (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Saunders, D., F. Stahlberg, et B. Byrne (2019). UCAM biomedical translation at WMT19 : Transfer learning multi-domain ensembles. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3 : Shared Task Papers, Day 2)*, Florence, Italy, pp. 169–174. Association for Computational Linguistics.

Shakurova, L., B. Nyari, C. Li, et M. Rotaru (2019). Best practices for learning domain-specific cross-lingual embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Florence, Italy, pp. 230–234. Association for Computational Linguistics.

Smith, S. L., D. H. P. Turban, S. Hamblin, et N. Y. Hammerla (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

Søgaard, A., S. Ruder, et I. Vulić (2018). On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia, pp. 778–788. Association for Computational Linguistics.

van der Maaten, L. et G. Hinton (2008). Visualizing data using t-sne. *Journal of Machine Learning Research 9*(86), 2579–2605.

Xing, C., D. Wang, C. Liu, et Y. Lin (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Denver, Colorado, pp. 1006–1011. Association for Computational Linguistics.

Zhang, M., Y. Liu, H. Luan, et M. Sun (2017a). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Vancouver, Canada, pp. 1959–1970. Association for Computational Linguistics.

Zhang, M., Y. Liu, H. Luan, et M. Sun (2017b). Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 1934–1945. Association for Computational Linguistics.

Summary

We aim to create an unsupervised cross-lingual embedding based on corpora from different domains. More precisely we align a biomedical English embedding with a general-domain non-English embedding under the hypothesis that monolingual data in the biomedical domain is mainly available in English. Our method for aligning two embeddings from different domains and languages relies on a proxy embedding of same domain as one embedding and same language as the other. The same-domain embeddings are aligned together in order to generate a dictionary for aligning the cross-domain embeddings. We evaluate our proposed algorithm on a biomedical translation task in several languages. While our method gives results below same-domain alignment approaches, it outperforms other cross-domain alignment techniques. This preliminary work ultimately intends to show that aligning different domains in an unsupervised manner is possible.

Keywords: word embedding, natural language processing, multilingual, unsupervised learning