Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Transport optimal pour le changement sémantique à partir de plongements contextualisés

Syrielle Montariol[1]* Alexandre Allauzen[2]
(1) INRIA Paris, France
(2) PSL, Dauphine-Université et ESPCI, Paris, France
syrielle.montariol@inria.fr, alexandre.allauzen@espci.psl.eu

RÉSUMÉ

Plusieurs méthodes de détection des changements sémantiques utilisant des plongements lexicaux contextualisés sont apparues récemment. Elles permettent une analyse fine du changement d'usage des mots, en agrégeant les plongements contextualisés en clusters qui reflètent les différents usages d'un mot. Nous proposons une nouvelle méthode basée sur le transport optimal. Nous l'évaluons sur plusieurs corpus annotés, montrant un gain de précision par rapport aux autres méthodes utilisant des plongements contextualisés, et l'illustrons sur un corpus d'articles de journaux.

ABSTRACT

Optimal Transport for Semantic Change Detection using Contextualised Embeddings

Several methods for semantic change detection with contextualised embeddings emerged recently. They allow a fine-grained analysis of word usage change by aggregating embeddings into clusters that reflect different usages of a word. We propose a novel method based on optimal transport. We evaluate it on several annotated corpora, showing a gain in accuracy compared to other methods using contextualised embeddings, and illustrate it on a corpus of newspaper articles.

MOTS-CLÉS : Changement sémantique, Transport optimal, Plongements contextualisés.

KEYWORDS: Semantic change, Optimal transport, Contextualised embeddings.

1 Introduction

La *diachronie* désigne l'évolution du langage à travers le temps. L'un des aspects de la diachronie est l'évolution de la signification des mots. Détecter et comprendre ces changements sémantiques est utile, par exemple, en sociolinguistique et en linguistique historique. Ce domaine a rapidement évolué avec l'essor de la sémantique distributionnelle ; les modèles de plongements lexicaux diachroniques ont connu une vague d'intérêt ces dernières années (Tahmasebi *et al.*, 2018). Ils sont utilisés pour des tâches d'analyse de flux de texte, telles que la détection d'événements (Kutuzov *et al.*, 2017) ou la surveillance des changements de discours lors de crises (Stewart *et al.*, 2017).

Suite à l'émergence des plongements lexicaux (e.g. Word2Vec, (Mikolov *et al.*, 2013)), une approche classique pour le changement sémantique implique d'apprendre des plongement pour chaque strate temporelles d'un corpus et de les rendre comparables en alignant les espaces vectoriels (Hamilton *et al.*, 2016). Cette méthode s'est révélée efficace sur un ensemble de dérives sémantiques synthé-

* Travaux effectués au sein du laboratoire LISN-CNRS en partenariat avec l'Université Paris-Saclay.

tiques (Shoemark *et al.*, 2019) et a été largement utilisée dans la littérature (Dubossarsky *et al.*, 2019; Schlechtweg *et al.*, 2020). Une autre approche compare les voisins d'un mot dans son espaces de représentation à différentes périodes (Yin *et al.*, 2018; Gonen *et al.*, 2020). Dans toutes ces méthodes, chaque mot n'a qu'une seule représentation dans une tranche de temps, ce qui limite la sensibilité et l'interprétabilité de ces techniques. Les plongements contextualisés issus de modèles de langues tels que BERT (Devlin *et al.*, 2019) permettent à chaque occurrence de mot d'avoir une représentation vectorielle qui lui est propre. Des travaux récents montrent que de tels plongements peuvent être utilisés pour la détection des changements sémantiques, en agrégeant les informations de l'ensemble des plongements d'un mot selon différentes méthodes (Martinc *et al.*, 2020b; Giulianelli *et al.*, 2020).

Dans cet article, nous résumons ces méthodes et proposons une nouvelle approche basée sur le transport optimal. Bien qu'ancienne (Monge, 1781), la théorie du transport optimal a connu des avancées importantes au XXe siècle, notamment avec Brenier (1987) qui lie le problème avec la théorie des probabilités. Avec l'essor de l'apprentissage automatique, les applications du transport optimal se sont élargies. Dans le cadre du TAL, il est appliqué au transport de mots ou d'ensembles de mots pour diverses tâches : alignement d'espaces de représentation (Alvarez-Melis & Jaakkola, 2018; Alaux *et al.*, 2019), classification de documents (Kusner *et al.*, 2015), mais aussi apprentissage de représentation et *topic modelling* (Xu *et al.*, 2018). Dans le cadre de l'évolution langagière, Huang & Paul (2019) utilisent la distance de Wasserstein pour mesurer le changement sémantique entre plusieurs corpus. La détection de variation sémantique au niveau lexical est un nouveau cadre d'application prometteur. Nous évaluons notre méthode et la comparons avec celles de la littérature sur plusieurs corpus annotés, puis l'appliquons à un corpus d'articles de journaux pour l'illustrer.

2 Méthodologie

Nous utilisons un modèle BERT pré-entraîné pour extraire des plongements contextualisés [1] d'un corpus divisé en strates temporelles, pour une liste de mots-cibles. Ces plongements peuvent être comparés entre deux strates temporelles pour mesurer le degré de changement sémantique du mot. Une première méthode de comparaison, le "moyennage", consiste à moyenner tous les plongements contextualisés d'un mot apparaissant à une période donnée (Martinc *et al.*, 2020b). On obtient une unique représentation vectorielle du mot pour chaque période ; ces vecteurs peuvent être comparés à l'aide de la distance cosinus (DC). [2] Dans la section qui suit, nous présentons une seconde méthode basée sur le *clustering* des plongements, sur laquelle s'appuie notre méthode de transport optimal.

2.1 Méthode de clustering

Nous effectuons un clustering sur tous les plongements contextualisés d'un mot, et considérons chaque cluster comme un usage du mot. Nous en déduisons la distribution des usages à chaque période. Un algorithme communément utilisé pour cette tâche est la propagation par affinité (Martinc *et al.*, 2020a), un clustering itératif qui déduit automatiquement le nombre de clusters pendant l'entraînement (Frey & Dueck, 2007). Néanmoins, les clusters sont généralement nombreux et de tailles très inégales.

1. Nous obtenons des plongements contextualisés en additionnant les quatre dernières couches de sortie des encodeurs de BERT. Le plongement d'un mot est calculé à partir de la moyenne des plongements de ses *wordpieces*.

2. Par abus de langage, nous définissons ici la distance cosinus comme le complément de la similarité cosinus dans l'espace des réels positifs (1 - similarité cosinus).

Pour surmonter cette limitation — diminuer le nombre de clusters a posteriori, afin de se concentrer sur les usages "principaux" des mots tout en limitant la perte d'information — nous utilisons la méthode de *filtrage* de Montariol *et al.* (2021). Un cluster est représenté par la moyenne de tous les plongements qu'il contient. Nous fusionnons chaque cluster avec le cluster le plus proche selon la distance cosinus entre leurs moyennes. Si le plus proche se trouve à une distance supérieure à un seuil [3] et que le cluster comporte moins de 10 éléments, [4] alors il est supprimé. Cette procédure est appliquée de manière récursive jusqu'à ce que la distance minimale entre deux clusters soit supérieure au seuil, ou qu'il ne reste que 2 clusters.

Pour un mot donné, le clustering est effectué sur les plongements issus de toutes les strates temporelles conjointement, afin de déduire une distribution de clusters unique pour toutes les périodes. Les distributions sont normalisées par la fréquence des mots dans leur strate, et comparées en utilisant la divergence de Jensen-Shannon (DJS) (Lin, 2006).

2.2 Méthode du transport optimal

La méthode du moyennage conserve la dimension originale des plongements de BERT (768) et permet une comparaison précise du contexte moyen d'un mot entre strates temporelles ; mais l'information sur la diversité du contexte intra-période est perdue. À l'inverse, la méthode de clustering capture la variabilité du contexte d'un mot en décomposant ses représentations en une distribution de faible dimension ; cependant, l'information sémantique apprise par le modèle et enregistrée dans les plongements est perdue. Pour conserver les deux types d'informations lors de la comparaison de l'usage d'un mot entre deux périodes, nous nous appuyons sur le cadre du transport optimal.

Formulation. L'ensemble des plongements contextualisés d'un mot sont regroupés ; soit avec un clustering unique comme dans la section précédente, soit avec un clustering différent pour les plongements de chaque période. Ensuite, nous calculons la moyenne de tous les plongements d'une période à l'intérieur d'un cluster. Dans une situation avec K clusters et T périodes, on obtient une matrice de plongements (de taille $T \times K \times 768$) et une distribution des clusters ($T \times K$) pour chaque mot. Nous résumons ainsi toutes les informations du nuage de plongements contextualisés à chaque période en K points dans un espace de dimension 768, les *centroïdes*, pondérés par le nombre de plongements dans le cluster associé. Nous souhaitons comparer ces centroïdes entre les périodes.

Cette configuration peut être formulée de la manière suivante. On note $\mu^{(1)}, \mu^{(2)} \in \mathbb{R}^{K\times 768}$ les ensembles de K centroïdes dans les deux périodes, et $c^{(1)}, c^{(2)} \in \Delta^{K-1}$ les distributions marginales des clusters telles que $c_i^{(t)} = p(C = i|\mathcal{T} = t, w)$ est la distribution du cluster i des plongements du mot w pour la période t. Δ^{K-1} est le simplexe $K-1$ standard : $c^{(1)}$ et $c^{(2)}$ sont des vecteurs positifs de dimension K et se somment à 1. Ils représentent les poids de chaque centroïde dans les espaces source et cible ($\mu^{(1)}$ et $\mu^{(2)}$). Nous quantifions l'effort de déplacement d'une unité de masse d'un centroïde de $\mu^{(1)}$ à un centroïde de $\mu^{(2)}$ avec une fonction de coût, ici la distance cosinus. Nous résolvons alors le problème en recherchant l'effort minimal requis pour transformer la distribution de masse de $c^{(1)}$ sur $\mu^{(1)}$ en celle de $c^{(2)}$ sur $\mu^{(2)}$.

3. Le seuil est égal à $moy_{dc} - 2 \times ect_{dc}$, où moy_{dc} est la moyenne des distances cosinus entre les clusters et ect_{dc} est l'écart-type. Les valeurs limitées par ce seuil représentent environ 95% d'une distribution normale.

4. Le nombre de 10 est choisi à partir de la procédure d'annotation la tâche SemEval2020–1 (Schlechtweg *et al.*, 2020), où chaque sens doit être annoté au moins 5 fois dans une période afin d'être validé. Nous expérimentons sur des corpus divisés en 2 périodes, d'où le choix de 10 instances.

Distance de Wasserstein (DW). Le transport optimal, également appelé problème de Monge-Kantorovitch, a pour but de résoudre ce problème d'optimisation. Il peut être formulé et résolu avec la programmation linéaire. Ici, nous donnons un bref aperçu du cadre de ce problème ; pour plus de détails, nous renvoyons le lecteur à des articles tels que (Villani, 2004; Solomon, 2018). La DW est positive, symétrique et satisfait l'inégalité triangulaire : autant de propriétés qui en font une distance. Pour notre tâche, elle peut être calculée de la manière suivante :

$$W(c^{(1)}, c^{(2)}) = \min_{\gamma} \sum_{i,j} \gamma_{ij} \cos(\mu_i^{(1)}, \mu_j^{(2)}) \text{ avec } \gamma \mathbf{1} = c^{(1)};\ \gamma^\intercal \mathbf{1} = c^{(2)};\ \gamma \geq 0 \qquad (1)$$

En d'autres termes, nous voulons minimiser le travail total ($\min_\gamma$) pour aller de $c^{(1)}$ à $c^{(2)}$ à l'aide de la distance cosinus ($\cos$), étant donné que la masse transportée est positive ($\gamma \geq 0$). La résolution de cette équation conduit à un plan de transport γ. Elle peut être vue comme une fonction de masse de probabilité sur $K \times K$ dont les marginales sont $c^{(1)}$ et $c^{(2)}$, et qui quantifie la proportion de masse $c_i^{(1)}$ de $\mu_i^{(1)}$ devant être transférée vers $\mu_j^{(2)}$ afin d'obtenir une masse $c_j^{(2)}$ de la manière la plus efficace. La DW représente la somme de tout le travail nécessaire pour résoudre le problème.

Notons que ce problème est complètement différent de la configuration de la section précédente résolue avec la divergence de Jensen-Shannon ; au lieu de comparer deux distributions, nous comparons deux ensembles pondérés de centroïdes. C'est pourquoi nous n'avons pas besoin d'avoir les mêmes clusters pour toutes les strates temporelles ; deux clusterings indépendants, un par période, pourraient permettre un meilleur ajustement pour chaque ensemble de points sans nuire au calcul de la distance.

3 Évaluation

Extraction des plongements contextualisés. Afin de résumer l'information efficacement, au lieu de garder en mémoire autant de vecteurs que d'occurrences d'un mot, nous utilisons une méthode de regroupement-moyenne (Montariol *et al.*, 2021) en ne stockant qu'un nombre limité de plongements (ici 200) pour chaque strate. À chaque nouvelle occurrence du mot dans la strate, son plongement e_{new} est additionné au vecteur e_m qui est lui est le plus similaire dans la liste de 200 plongements [5]. Le nombre d'éléments ajoutés dans e_m est incrémenté (compteur $c_m \leftarrow c_m + 1$) pour normaliser chaque élément de la liste de plongements à la fin de l'extraction.

Données annotées et modèles. Nous utilisons six jeux de données annotés pour l'évaluation : un jeu en anglais appelé "GEMS", quatre issus d'une tâche d'évaluation SemEval, et "DURel", un jeu en allemand. GEMS (Gulordava & Baroni, 2011) est nommé ainsi après le workshop GEMS où l'article associé a été publié. 5 annotateurs ont évalué le degré de changement sémantique de 100 mots anglais entre les années 1960 et 1990, sans observer les mots en contexte. Afin d'étudier l'évolution sémantique de ces mots, nous générons des plongements contextualisés à partir de textes des décennies 1960 et 1990 du Corpus of Historical American English (COHA) [6] (2,8M et 3,3M de mots respectivement). La tâche SemEval 2020 – 1 : Détection non supervisée de changement lexico-sémantique (Schlechtweg *et al.*, 2020) propose des données annotées en utilisant une nouvelle

5. En utilisant la distance cosinus : $e_m = \arg\min_{e_i \in L} \cos(e_i, e_{\text{new}})$.
6. https://www.english-corpora.org/coha/

Méthode	Mesure	GEMS	SemEval				DURel	Moy
			Anglais	Allemand	Suédois	Latin		
base	DW	0,312	0,386	0,416	0,252	0,283	0,526	0,363
clustering	DW	0,369	**0,456**	0,421	0,264	0,397	0,484	0,399
2× clustering	DW	0,380	0,412	0,457	0,190	0,426	0,530	0,399
clustering + filtrage	DW	0,352	0,437	0,561	0,321	0,488	0,686	0,474
clustering	DJS	**0,394**	0,371	0,498	0,012	0,346	0,512	0,355
clustering + filtrage	DJS	0,403	0,348	**0,583**	0,018	0,408	0,712	0,412
moyennage	DC	0,349	0,315	0,565	0,212	**0,496**	0,656	0,432
SGNS + PO	DC	0,347	0,321	0,712	**0,631**	0,372	**0,814**	**0,533**

TABLE 1 – Corrélation de Spearman entre les classements de chaque système et le classement issu de l'annotation, pour chaque corpus de test. Les valeurs grisées indiquent une corrélation non significative (p-valeur > 0,05). [10]

approche : les annotateurs décident si une paire de phrases de différentes périodes portent la même signification du mot-cible (Schlechtweg & Schulte im Walde, 2020). Les corpus, en quatre langues — anglais (13,4M de mots), allemand (142M), suédois (182M) et latin (11,2M) — sont divisés en deux périodes et les phrases sont mélangées et lemmatisées. Enfin, le jeu de données DURel [7] (Schlechtweg *et al.*, 2018) est composé de 22 mots allemands, classés selon leur degré de changement sémantique entre deux périodes par cinq annotateurs selon la même méthode que pour SemEval. Nous générons des plongements en utilisant le corpus allemand DTA, également lemmatisé (25M de mots pour 1750–1799 et 38M pour 1850–1899). L'adaptation au domaine (tâche *masked language model*) est effectuée sur chaque corpus pour 5 itérations, en utilisant des modèles BERT pré-entraînés adaptés à chaque langue. [8] Les mots-cibles sont classés en fonction de leur degré de changement sémantique à l'aide des méthodes décrites précédemment. Le classement est comparé avec la vérité terrain à l'aide de la corrélation de Spearman.

Résultats. Nous appliquons différentes méthodes reposant sur le transport optimal pour calculer l'évolution de l'usage d'un mot entre deux périodes de temps (Table 1). [11] Nous pouvons soit faire un clustering unique sur les plongements des deux périodes, soit un clustering différent pour chaque période (noté "2× clustering"). Une autre variation consiste à utiliser le compteur c_m du nombre de plongements ayant été additionnés pour former chacun des 200 vecteurs e_m de la liste, pour pondérer la matrice de coût lors du calcul de la DW sans effectuer de clustering (noté "base"). Nous comparons ces méthodes avec le clustering classique, où la distance est calculée avec la DJS, et avec le moyennage, calculé avec la distance cosinus (DC).

La réalisation de deux clusterings indépendants n'améliore pas les résultats par rapport à un clustering unique, en moyenne. De grands écarts peuvent être observés entre les corpus, dans les deux sens.

7. https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/durel/

8. Pour l'allemand : bert-base-german-cased (https://deepset.ai/german-bert), pour l'anglais : bert-base-uncased model, pour le latin : bert-base-multilingual-uncased, pour le suédois : bert-base-swedish-uncased (https://github.com/af-ai-center/SweBERT).

10. BERT sur le suédois mène toujours à une corrélation non significative. Nous supposons que ce défaut est dû au modèle utilisé, qui est pré-entraîné sur des textes récents et non lemmatisés ; ils sont donc très éloignés du corpus étudié, composé de textes historiques, lemmatisés et comportant de nombreuses erreurs d'OCR.

11. En utilisant le package POT : https://pythonot.github.io/.

Cependant, comme le nombre de mots-cibles est faible, cela peut s'expliquer par le fait que seuls quelques mots bénéficient ou souffrent du degré de liberté supplémentaire donné par les clusterings indépendants. En effet, sur le plus grand jeu de données de test GEMS (100 mots), les performances entre 1 ou 2 clusterings sont comparables. D'autre part, la comparaison entre les listes complètes de 200 plongements des deux strates sans effectuer de clustering ("base") conduit à une performance moyenne plus faible. L'agrégation apportée par le clustering est donc nécessaire pour limiter la sensibilité au bruit. En outre, le filtrage a un effet positif important sur le clustering, en particulier avec la DW ; en supprimant les clusters minoritaires et extrêmes et en fusionnant les clusters semblables, il vient réduire le bruit et compense le large nombre de clusters inégalement distribués extraits par la propagation par affinité.

Le système SGNS + PO (Schlechtweg *et al.*, 2019) est le seul utilisant des plongements non contextualisés : le modèle Skip-Gram (SGNS) est entraîné sur les deux périodes séparément, et les espaces de représentation sont alignés avec le problème de Procuste orthogonal (PO). La DC est utilisée pour mesurer le changement sémantique. Ce système surpasse largement les autres méthodes en moyenne. Cela peut être lié au fait que les phrases de tous les corpus d'évaluation à l'exception de COHA sont mélangées et lemmatisées. Par conséquent, les modèles BERT ne peuvent exploiter que les phrases au lieu d'une séquence complète de 256 éléments. De plus, SGNS est entraîné intégralement sur les corpus d'évaluation, tandis que les modèles BERT sont pré-entraînés sur du texte brut. Les plongements issus de BERT souffrent dont potentiellement plus de la lemmatisation des corpus.

Parmi les méthodes de plongements contextualisés, le score moyen le plus élevé est obtenu avec clustering + filtrage + DW. Cependant, on trouve de larges disparités selon les jeux de données. Le moyennage surpasse le clustering pour SemEval Latin, tandis que DJS et DW se surpassent alternativement sur les autres corpus. Cette disparité ne semble pas liée à la langue, car des méthodes différentes mènent aux meilleurs scores pour une langue commune (GEMS et SemEval pour l'anglais, DURel et SemEval pour l'allemand). Une cause pourraient être la façon dont chaque méthode distribue les scores de changement sémantique, par rapport à la distribution des scores de la vérité terrain (SemEval latin et allemand et DURel ont des scores de vérité terrain répartis uniformément tandis que SemEval anglais et suédois et GEMS ont une plus forte proportion de scores faibles). En résumé, étant donné que la méthode de transport optimal utilise à la fois les informations du clustering et du moyennage, elle constitue un bon compromis entre ces deux méthodes bien qu'elle ne les surpasse pas systématiquement.

Pour conclure, l'accord inter-annotateurs (moyenne des corrélations de Pearson par paire d'annotateurs) permet de mettre en perspective les performances des méthodes : il est de 0,51 sur le jeu de données GEMS, 0,66 pour DURel et 0,62 pour SemEval (en moyenne sur les 4 jeux de données).

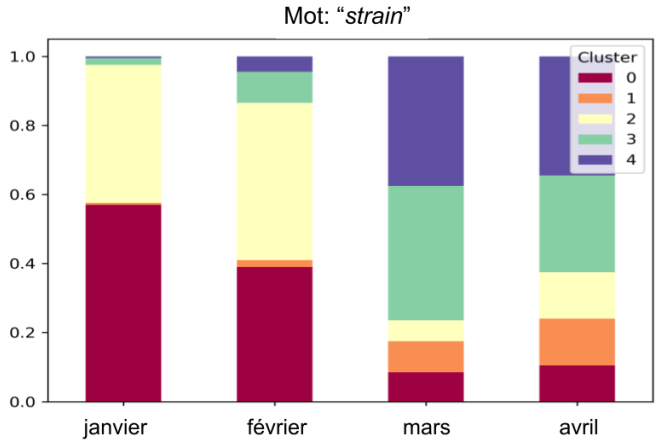
4 Application

Nous montrons un exemple d'exploration diachronique d'un corpus d'articles de journaux en anglais sur le COVID-19. [12] Nous analysons environ 500k articles de janvier à avril 2020, que nous divisons en 4 strates mensuelles de tailles inégales (160M mots en mars, 41M en février, 35M en avril et 10M en janvier). La méthode de transport optimal permet de quantifier l'évolution de l'ensemble des mots du vocabulaire, en extrayant leurs plongements avec la méthode de regroupement-moyenne.

12. https://blog.aylien.com/free-coronavirus-news-dataset/

Nous calculons la DW moyenne entre les mois successifs pour chaque mot du vocabulaire. *Strain* est le 38ème mot avec la DW moyenne la plus élevée, et le 15ème entre février et mars 2020 ; nous nous concentrons dessus pour illustrer les phénomènes d'évolution sémantique pouvant être détectés. *Strain* est un mot polysémique ayant deux sens principaux en anglais apparaissant dans notre corpus : comme la variante d'un virus ou d'une bactérie (terme biologique) et comme "une demande sévère ou excessive sur les ressources ou les capacités de quelqu'un ou de quelque chose" (dictionnaire Oxford). Nous regroupons ses plongements contextuels avec l'algorithme k-means ($k = 5$). Puis, en calculant un score de *tf-idf* sur les unigrammes et les bigrammes des phrases dans lesquelles ce mot apparaît, nous extrayons un ensemble de mots-clés pour chaque cluster – les mots ayant le score de tf-idf le plus élevé – afin d'interpréter les variations de leur distribution (Figure 1).

Les clusters 1, 3 et 4, qui correspondent au deuxième sens du terme (pression sur les systèmes de santé dans le cluster 4, pression financière dans le cluster 3 et pression sur les ressources et les infrastructures dans le cluster 1), voient leur proportion augmenter au fil du temps ; tandis que les clusters 0 et 2, qui correspondent au premier sens du terme (en tant que nouvelle souche de virus), diminuent. Ce comportement souligne l'évolution des préoccupations liées à la pandémie dans les journaux. Ainsi, on observe l'évolution de la répartition des différents sens du mot en terme lexicographique au cours du temps ; mais la méthode permet aussi de révéler les variations d'usage au sein d'une même signification, en fonction des événements de l'actualité.



#	Mots-clés
0	*strain coronavirus, new strain, city wuhan, novel strain, strain virus, chinese city*
1	*strain health, strain resources, stream, network infrastructure, international resources, likely strain*
2	*new strain, acute respiratory, 2019 ncov, respiratory syndrome, severe acute, identified humans*
3	*financial strain, feeling strain, strain coronavirus, economic strain, signs strain, strain said*
4	*ease strain, putting strain, strain health, reduce strain, care system, strain hospitals*

FIGURE 1 – Distributions des clusters par mois et mots-clés principaux pour le mot *strain*.

5 Conclusion

L'évaluation sur données annotées a montré que parmi les méthodes à base de plongements contextualisés, la méthode la plus performante utilise la distance de Wasserstein sur les clusters issus de la propagation par affinité des plongements de BERT. Néanmoins, elle est en moyenne moins performante que la méthode utilisant des plongements non contextualisés (Skip-Gram avec alignement). Malgré ses performances plus faibles, la méthode basée sur le clustering offre une interprétation plus fine que les méthodes basées sur des plongements non contextualisés, car elle tient compte de la diversité des usages et des sens d'un mot ; en particulier, le clustering permet de distinguer les différents usages du mot étudié. C'est pourquoi cette approche peut être utilisée pour détecter l'apparition de nouveaux usages des mots, tracer l'évolution des différents usages, et les interpréter.

Références

ALAUX J., GRAVE E., CUTURI M. & JOULIN A. (2019). Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*.

ALVAREZ-MELIS D. & JAAKKOLA T. (2018). Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 1881–1890, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1214.

BRENIER Y. (1987). Decomposition polaire et rearrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Ser. I Math.*, **305**, 805–808.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DUBOSSARSKY H., HENGCHEN S., TAHMASEBI N. & SCHLECHTWEG D. (2019). Time-out : Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 457–470, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1044.

FREY B. J. & DUECK D. (2007). Clustering by passing messages between data points. *Science*, **315**(5814), 972–976.

GIULIANELLI M., DEL TREDICI M. & FERNÁNDEZ R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3960–3973, Online : Association for Computational Linguistics.

GONEN H., JAWAHAR G., SEDDAH D. & GOLDBERG Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 538–555, Online : Association for Computational Linguistics.

GULORDAVA K. & BARONI M. (2011). A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, p. 67–71 : Association for Computational Linguistics.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 1489–1501. DOI : 10.18653/v1/P16-1141.

HUANG X. & PAUL M. J. (2019). Neural temporality adaptation for document classification : diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4113–4123, Florence, Italy.

KUSNER M. J., SUN Y., KOLKIN N. I. & WEINBERGER K. Q. (2015). From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 957–966 : JMLR.org.

KUTUZOV A., VELLDAL E. & ØVRELID L. (2017). Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, p. 31–36, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/W17-2705.

LIN J. (2006). Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, **37**(1), 145–151. DOI : 10.1109/18.61115.

MARTINC M., MONTARIOL S., ZOSA E. & PIVOVAROVA L. (2020a). Capturing evolution in word usage : Just add more clusters ? In *Companion Proceedings of the Web Conference 2020*, WWW '20, p. 343–349, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3366424.3382186.

MARTINC M., NOVAK P. K. & POLLAK S. (2020b). Leveraging contextual embeddings for detecting diachronic semantic shift. *LREC 2020*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

MONGE G. (1781). Mémoire sur la théorie des déblais et des remblais. *Imprimerie Royale*.

MONTARIOL S., MARTINC M. & PIVOVAROVA L. (2021). Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

SCHLECHTWEG D., HÄTTY A., DEL TREDICI M. & SCHULTE IM WALDE S. (2019). A wind of change : Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 732–746, Florence, Italy : Association for Computational Linguistics.

SCHLECHTWEG D., IM WALDE S. S. & ECKMANN S. (2018). Diachronic usage relatedness (durel) : A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 169–174.

SCHLECHTWEG D., MCGILLIVRAY B., HENGCHEN S., DUBOSSARSKY H. & TAHMASEBI N. (2020). Semeval-2020 task 1 : Unsupervised lexical semantic change detection. *SemEval@COLING2020*.

SCHLECHTWEG D. & SCHULTE IM WALDE S. (2020). Simulating lexical semantic change from sense-annotated data. *CoRR*, **abs/2001.03216**.

SHOEMARK P., LIZA F. F., NGUYEN D., HALE S. & MCGILLIVRAY B. (2019). Room to Glo : A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, p. 66–76, Hong Kong, China : Association for Computational Linguistics.

SOLOMON J. (2018). Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*.

STEWART I., ARENDT D., BELL E. & VOLKOVA S. (2017). Measuring, predicting and visualizing short-term change in word representation and usage in VKontakte social network. In *Eleventh international AAAI conference on web and social media*.

TAHMASEBI N., BORIN L. & JATOWT A. (2018). Survey of computational approaches to diachronic conceptual change. *CoRR*, **1811.06278**.

VILLANI C. (2004). Transport optimal : coup de neuf pour un très vieux problème. In *Images des Mathématiques* : CNRS.

XU H., WANG W., LIU W. & CARIN L. (2018). Distilled wasserstein learning for word embedding and topic modeling. In S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 31*, p. 1716–1725. Curran Associates, Inc.

YIN Z., SACHIDANANDA V. & PRABHAKAR B. (2018). The global anchor method for quantifying linguistic shifts and domain adaptation. In *Advances in neural information processing systems*, p. 9412–9423.