Étude de la reproductibilité des *word embeddings* : repérage des zones stables et instables dans le lexique

Bénédicte Pierrejean Ludovic Tanguy
CLLE-ERSS : CNRS & Université de Toulouse
{benedicte.pierrejean, ludovic.tanguy}@univ-tlse2.fr

RÉSUMÉ

Les modèles vectoriels de sémantique distributionnelle (ou *word embeddings*), notamment ceux produits par les méthodes neuronales, posent des questions de reproductibilité et donnent des représentations différentes à chaque utilisation, même sans modifier leurs paramètres. Nous présentons ici un ensemble d'expérimentations permettant de mesurer cette instabilité, à la fois globalement et localement. Globalement, nous avons mesuré le taux de variation du voisinage des mots sur trois corpus différents, qui est estimé autour de 17% pour les 25 plus proches voisins d'un mot. Localement, nous avons identifié et caractérisé certaines zones de l'espace sémantique qui montrent une relative stabilité, ainsi que des cas de grande instabilité.

ABSTRACT

Reproducibility of word embeddings : identifying stable and unstable zones in the semantic space

Distributional semantic models trained using neural networks techniques yield different models even when using the same parameters. We describe a series of experiments where we examine the instability of word embeddings both from a global and local perspective for several models trained with the same parameters. We measured the global variation for models trained on three different corpora. This variation is estimated to about 17% for the 25 nearest neighbours of a target word. We also identified and described local zones of stability and instability in the semantic space.

MOTS-CLÉS : plongements lexicaux, évaluation, stabilité, reproductibilité.

KEYWORDS: word embeddings, evaluation, stability, reproducibility.

1 Introduction

La popularité des modèles prédictifs en sémantique distributionnelle est indéniable, et les outils comme *Word2vec* (Mikolov *et al.*, 2013) ont su s'imposer comme de nouveaux standards à la fois dans les applications du TAL et dans les investigations empiriques des différentes facettes de la sémantique lexicale. Ces outils permettent de représenter de façon compacte des unités lexicales (par des vecteurs de quelques centaines de dimensions), en se basant uniquement sur des quantités de texte correspondant aux standards contemporains (centaines de millions ou milliards de mots) et d'obtenir par ce biais des rapprochements convaincants d'unités lexicales sémantiquement liées.

Toutefois, au-delà des différentes méthodes mises en œuvre dans ces techniques, leur utilisation soulève un ensemble de questions à la fois méthodologiques et pratiques. Si les embeddings (ou

plongements lexicaux) ont prouvé leur intérêt, et sont utilisés sans hésitation pour représenter des contenus lexicaux en amont d'une application de TAL (notamment basée sur des méthodes d'apprentissage profond) (Goldberg, 2016), plusieurs questions restent en suspens quant aux conditions d'utilisation de ces techniques.

Depuis longtemps on sait que les méthodes d'analyse distributionnelle sont très sensibles à un ensemble de paramètres. On se base généralement sur des évaluations intrinsèques pour estimer la valeur optimale des paramètres, en comparant par exemple les similarités entre les mots au sein des modèles distributionnels avec celles estimées par des humains (Faruqui *et al.*, 2016), et ce sur un échantillon réduit (généralement quelques centaines de paires de mots) ce qui soulève des problèmes concernant la fiabilité des évaluations effectuées.

Les modèles distributionnels neuronaux de type *Word2vec* soulèvent d'autres questions liées à la *reproductibilité* des modèles construits, puisqu'ils se basent sur des méthodes stochastiques à différentes étapes de leur traitement. Autrement dit, pour un même outil paramétré de la même façon et appliqué à un même corpus, un modèle différent est généré à chaque exécution. Cette instabilité doit bien entendu être prise en compte, surtout lorsque l'on utilise les modèles distributionnels à des fins d'exploration locale d'une partie du lexique ou d'un corpus particulier (Hellrich & Hahn, 2016; Antoniak & Mimno, 2018). La table 1 donne quelques exemples des mots les plus similaires de deux mots-cibles pour trois de ces modèles construits de façon identique (voir en 3.1 pour les détails de ceux-ci). On y voit que si globalement les différents voisins sont sémantiquement pertinents, certains d'entre eux n'apparaissent en bonne place que dans certains modèles (e.g. *sheaf* et *newspaper* pour *paper*, *pink* et *red* pour *white*), et que les variations ne semblent pas du même ordre pour les deux mots-cibles choisis.

Mot-cible	paper (n)			white (adj)		
Rang	Modèle 1	Modèle 2	Modèle 3	Modèle 1	Modèle 2	Modèle 3
1	magazine	book	book	black	black	black
2	book	magazine	sheaf	grey	grey	yellow
3	pamphlet	newspaper	magazine	blue	yellow	grey
4	journal	pamphlet	folder	yellow	red	blue
5	article	folder	parchment	pink	blue	red

TABLE 1. Exemple de la variation des 5 plus proches voisins du nom *paper* et de l'adjectif *white* dans deux modèles distributionnels construits de façon identique

Nous proposons dans cet article une série d'études visant à éclaircir ce point sous différents angles, en traitant les questions suivantes :

— Comment peut-on quantifier la stabilité interne d'un modèle distributionnel (ou plus précisément d'une *configuration* de modèle distributionnel) en comparant plusieurs instances produites dans des conditions identiques ?
— La variation d'une instance d'un modèle à une autre est-elle répartie uniformément dans l'espace vectoriel, ou au contraire y a-t-il des zones de plus grande stabilité ? Si oui, peut-on identifier et caractériser ces zones ?
— Comment la stabilité interne d'une configuration varie-t-elle lorsque l'on change les paramètres, notamment lorsque l'on change le corpus sur lequel le modèle est construit ?

La Section 2 présente un bref aperçu des travaux similaires portant sur l'évaluation et sur la reproductibilité des *embeddings*. Nous présentons en Section 3 notre dispositif expérimental et les choix faits

pour mesurer la variation entre deux modèles. Les deux dernières sections (4 et 5) sont consacrées à l'exploration globale puis locale de cette variation, notamment en identifiant les zones du lexique qui sont le plus (et le moins) sujettes à variation.

2 Travaux similaires

2.1 Evaluation des embeddings

L'étude de l'impact des différents paramètres sur la qualité relative des modèles distributionnels a donné lieu à de nombreuses publications. Il a ainsi été montré que le changement de paramètres peut entraîner des modifications importantes. Ces paramètres peuvent correspondre aux familles de méthodes de construction de modèles distributionnels (Bernier-Colborne & Drouin, 2016; Chiu *et al.*, 2016), aux hyperparamètres de ces méthodes, aux corpus d'apprentissage (nature, taille, prétraitements) (Asr *et al.*, 2016; Sahlgren & Lenci, 2016) ou encore aux types de contextes utilisés pour représenter les unités lexicales (Levy & Goldberg, 2014; Melamud *et al.*, 2016; Li *et al.*, 2017). Ces travaux abordent généralement la comparaison par le biais de bancs de test d'évaluation intrinsèque d'un modèle en comparant les similarités qui y sont calculées avec des mesures relevant de jugements humains. Bien que pratiques et peu coûteuses, ces méthodes d'évaluation présentent plusieurs inconvénients. Les jeux d'évaluation utilisés posent notamment des problèmes de subjectivité (scores plus élevés pour des mots qui sont associés comparé à des mots similaires) ou encore de surapprentissage (même jeux d'évaluation utilisés jusqu'à obtenir les résultats attendus) (Faruqui *et al.*, 2016). De plus, les jeux d'évaluation ne permettent d'évaluer qu'une partie d'un modèle puisqu'ils sont généralement constitués de quelques centaines de paires de mots.

Plusieurs alternatives ont été envisagées pour pallier à ces limites. La première est de faire appel à des évaluations extrinsèques en aval. Ces méthodes consistent à intégrer les embeddings dans un système de TAL dont la performance globale sera évaluée (classification de texte, analyse d'opinions, reconnaissance d'entités nommées etc.). Ces méthodes, beaucoup plus coûteuses que les méthodes d'évaluation intrinsèque, évaluent les embeddings en tant que composant d'un système et ne donnent pas d'information sur la qualité des embeddings eux-mêmes (Schnabel *et al.*, 2015).

Plusieurs travaux récents se penchent également sur la structure globale des espaces vectoriels produits par ces méthodes, comme Trost & Klakow (2017) afin de mettre au jour les différents biais pouvant exister dans les représentations des différents mots du lexique. Du point de vue de la linguistique de corpus, une autre façon intéressante de comparer deux modèles est celle présentée entre autres par Hamilton *et al.* (2016) qui cherchent les variations pour notamment observer les évolutions des représentations dans un corpus diachronique. Enfin, plus récemment et plus proche du travail présenté ici, certaines études se sont concentrées sur l'étude plus précise de la variation entre des modèles distributionnels : Antoniak & Mimno (2018) ont étudié à la fois la stabilité d'un même modèle à travers des changement de taille, de structure et de nature des corpus utilisés mais aussi à travers ses variations aléatoires inhérentes.

2.2 Reproductibilité des embeddings

L'instabilité interne des embeddings est un phénomène globalement ignoré dans les diverses études qui les exploitent ou les explorent. Elle est dûe à plusieurs phases de l'algorithme qui font appel à des processus aléatoires. Les réseaux neuronaux, notamment, sur lesquels se basent les méthodes comme

Word2vec, ont une phase de détermination aléatoire des poids initiaux des connections entre les neurones, qui sont ensuite ajustés en fonction des données fournies durant l'apprentissage. Toujours dans le cas de *Word2vec*, la méthode d'échantillonnage négatif (*negative sampling*) implique de plus la génération aléatoire de paires mot/contexte. Notons que cette présence de l'aléatoire dans la construction de modèles distributionnels n'est pas l'apanage des seuls réseaux de neurones, puisque par exemple la réduction de dimensions couramment utilisée dans les modèles fréquentiels est souvent faite avec des principes similaires, notamment les variantes randomisées de la décomposition en valeurs singulières (Sahlgren, 2005).

Cette instabilité a été soulevée par Hellrich & Hahn (2016) qui ont mesuré plus précisément les variations des premiers voisins des mots à travers des ensembles de 3 modèles *Word2vec* entraînés avec des hyperparamètres identiques. Ils ont notamment étudié le rôle de la fréquence des mots et surtout du nombre d'itérations dans la stabilité de ces voisinages. Plus récemment, Antoniak & Mimno (2018) ont montré la grande variation des voisinages sémantiques obtenus par plusieurs techniques de construction des embeddings et leur sensibilité à des variations minimes des données (comme l'ordre des documents dans le corpus), ainsi qu'au rôle des processus aléatoires impliqués. Ces deux études mettent en garde les utilisateurs de ces outils sur la nécessité de répéter plusieurs entraînements et d'étudier leur divergence avant de conclure, ce qui a bien entendu un coût calculatoire élevé. Notre approche diffère de ces deux études par le fait que nous cherchons plus précisément à étudier comment cette variation se distribue à travers le lexique, et plus précisément à caractériser en quoi certains mots ou ensemble de mots sont plus ou moins affectés par les processus aléatoires.

La réplicabilité (répétition à l'identique) et la reproductibilité (répétition en faisant varier un des paramètres, généralement les données) des expérimentations sont des questions de fond dans toutes les sciences expérimentales (on parle de 'reproducibility crisis') et plus récemment en informatique au premier rang de laquelle on trouve les travaux basés sur de l'apprentissage automatique (Hutson, 2018). En témoigne l'organisation récente de conférences et d'ateliers portant spécifiquement sur la question de la reproductibilité de travaux précédents. [1]

Une des façons de contrôler les processus aléatoires consiste à fixer la graine (*seed*) du générateur de nombres aléatoires, (cf. la règle numéro 6 des bonnes pratiques proposées par Sandve *et al.* (2013)), mais on conviendra que cette solution n'est pas satisfaisante intellectuellement, même si elle a l'intérêt de permettre la réplicabilité d'une expérience.

La reproductibilité est une question philosophique qui dépasse notre propos ici et qui nous amènerait à distinguer plus précisément ce qui relève, dans toute manipulation de données, d'un choix informé dans un processus déterministe (le seuil de fréquence, tel ou tel lemmatiseur, supprime-t-on ou non telle catégorie de mots-outils, etc.), de la sélection d'une valeur pour un hyperparamètre aux conséquences mal maîtrisées (nombre de dimensions, d'itérations, taux d'échantillonnage, etc.) et enfin de ce qui n'a simplement pas de sens ni de conséquence prévisible (la graine du générateur de nombres aléatoires). Plus modestement, prendre la mesure de ce phénomène simple et inévitable qu'est la part d'aléatoire dans le processus de construction des embeddings est pour nous une façon d'aborder la prise en main de ces techniques. Notre postulat est qu'un résultat affecté largement par une instabilité dûe aux seuls facteurs véritablement aléatoires est de moindre valeur qu'un phénomène apparaissant de façon répétée à travers des expériences similaires. Partant de là, notre but est de mieux cerner ce qui, pour le cas des embeddings, nous permet de mieux comprendre ce qui est à la portée de ces techniques, et à terme de délimiter les choix sur lesquels on peut véritablement agir en

1. Citons l'atelier *Reproducibility in Machine Learning* à ICML'17 ou encore les tâches participatives proposées par CENTRE (*CLEF, NTCIR, TREC Reproducibility*) à CLEF 2018.

connaissance de cause. Un autre point qui nous distingue des études sur l'instabilité présentées plus haut est que nous ne cherchons pas à contourner celle-ci, mais plutôt à la considérer comme un indice pour une meilleure compréhension de ces méthodes.

3 Dispositif

Nous résumons ici le dispositif expérimental mis en place, à savoir les modèles que nous avons construits et comment nous les comparons.

3.1 Modèles

Nous avons sélectionné une configuration unique en utilisant *Word2vec* avec les paramètres par défaut (code source original, méthode *skip-gram* avec *negative sampling* (taux de 5), fenêtre de taille 5, vecteurs de dimension 100, sous-échantillonnage des mots de fréquence supérieure à 10^{-3}, 5 itérations). Nous l'avons appliquée à un corpus générique de taille moyenne, le BNC (100 millions de mots, lemmatisés et catégorisés par Talismane (Urieli & Tanguy, 2013)), avec un seuil de fréquence minimale de 100. Cette même configuration a été utilisée 5 fois afin d'en tester la stabilité en comparant deux à deux les modèles produits. Ces modèles ne diffèrent donc que par l'effet des processus aléatoires inhérents à l'algorithme de *Word2vec* (voir § 2.2).

Nous avons répété l'opération en utilisant le même outil avec les mêmes paramètres sur deux autres corpus de taille similaire mais de nature très différente. Le premier corpus est le *ACL Anthology Reference corpus* (Bird *et al.*, 2008), constitué d'articles scientifiques dans le domaine du TAL, et le second est un corpus également constitué d'articles scientifiques, mais en biologie, constitué à partir du corpus *All of PLOS*.[2] Ces deux corpus ont eux aussi une taille de 100 millions de mots, et ont été traités de la même façon que le BNC.

3.2 Mesure de la variation

Pour estimer la variation entre deux modèles nous reprenons la méthode déjà utilisée par Sahlgren (2006) et reprise plus récemment par Antoniak & Mimno (2018) qui consiste à mesurer le recouvrement des N mots les plus similaires (plus proches voisins) d'un même mot-cible. Cette comparaison se fait indépendamment de l'ordre des voisins, et utilise généralement une valeur de N limitée. La formule ci-dessous indique comment est calculé le taux de variation (compris entre 0 et 1) pour un mot m entre deux modèles M_1 et M_2, $vois^N_M(m)$ représentant l'ensemble des N mots les plus similaires à m dans le modèle M.

$$var^N_{M_1,M_2}(m) = 1 - \frac{|vois^N_{M_1}(m) \cap vois^N_{M_2}(m)|}{N}$$

La mesure que nous avois choisie pour calculer les plus proches voisins est la similarité cosinus, désormais reconnue comme étant celle qui donne les meilleurs résultats dans l'exploitation des espaces distributionnels.

2. https://www.plos.org/text-and-data-mining

Cette mesure de la variation a bien entendu un ensemble d'avantages et d'inconvénients. Pour ce qui est des avantages, notons :

— La simplicité et la transparence : le taux de recouvrement est facilement interprétable, puisqu'il donne directement une estimation du nombre de voisins distributionnels que l'on retrouve d'un modèle à l'autre ;

— La facilité de calcul : l'extraction des plus proches voisins d'un mot est une procédure standard et bien optimisée dans les techniques d'exploitation des embeddings ;

— La localité : cette mesure permet d'attribuer un score de variation à chaque mot de l'espace sémantique et permet donc de comparer leur stabilité. Il est toutefois également possible d'utiliser une moyenne sur l'ensemble des mots pour obtenir une mesure globale de la variation entre deux modèles ;

— La pertinence : l'examen des plus proches voisins d'un mot est la façon la plus classique d'observer un espace sémantique, et il s'agit notamment de la procédure qui permet l'utilisation des espaces distributionnels en linguistique de corpus.

Les inconvénients principaux sont :

— L'absence de prise en compte de l'ordre des voisins : le fait de considérer les voisins non ordonnés ne donne qu'une image partielle de la variation locale entre deux modèles ;

— La sensibilité aux phénomènes de *hubness* : il est bien connu que la proximité dans les espaces vectoriels de haute dimension présente des biais que l'on qualifie de malédiction de la dimensionnalité (*dimensionality curse*) et notamment que certains points particuliers (*hubs*) ont tendance à apparaître très fréquemment dans les voisinages proches (Radovanović *et al.*, 2010). Ces phénomènes complexes peuvent naturellement perturber l'estimation de la stabilité ainsi mesurée.

— La nécessité de fixer N : le nombre de voisins considérés pour mesurer le recouvrement est bien entendu un paramètre très important qu'il est nécessaire de fixer a priori.

Nous comptons bien entendu aborder les questions soulevées par ces trois inconvénients, mais nous nous limitons dans cet article au dernier point, comme nous le verrons dans la section suivante.

4 Vue d'ensemble de la variation

Dans cette section nous présentons les mesures globales de la variation entre les différents modèles que nous avons entraînés, en considérant la variation moyenne observée sur l'ensemble des mots, mais également en mesurant la stabilité d'un échantillon de valeurs de similarité. La première question à traiter concerne le choix d'une valeur pour N, le nombre de plus proches voisins considérés.

4.1 Impact du nombre de voisins

Nous avons mesuré la variation moyenne pour chaque mot (sur les 10 paires de modèles, mais en traitant chaque corpus séparément) en utilisant différentes valeurs de N (1, 5, 10, 25, 50 et 100). Comme on peut le voir sur la figure 1 les scores moyens sont autour de 0.2, avec un faible écart-type autour des valeurs moyennes (indiqué par les deux traits pour chaque point de mesure) et décroissent naturellement lorsque N augmente tout en se stabilisant rapidement. La forme de cette évolution et le niveau de variation sont similaires sur les trois corpus.

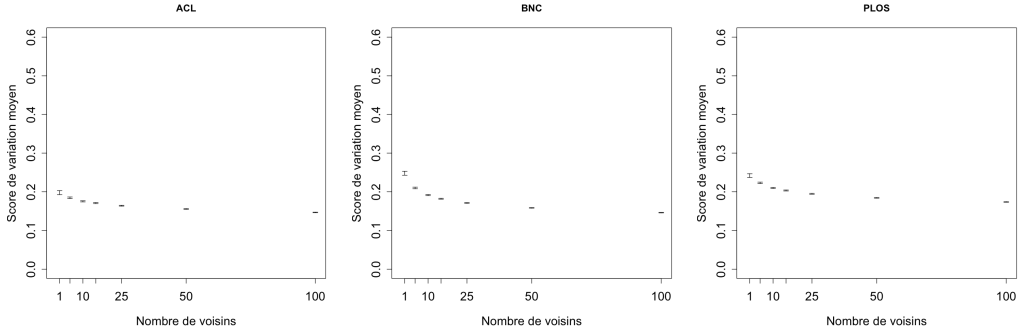

FIGURE 1. Variation moyenne pour différentes valeurs de N

Hellrich & Hahn (2016) ont observé la même stabilisation de leur score suivant le nombre de voisins considérés, mais ont opté pour un choix de N=1. Ce choix ne nous paraissait pas idéal pour l'observation plus détaillée des voisinages et nous avons pour notre part fixé N à 25. Cette valeur a été déterminée de la façon suivante. Pour les modèles entraînés sur le corpus BNC, nous avons calculé les coefficients de corrélation (sur l'ensemble des mots) entre les scores de variation obtenus pour toutes les paires possibles de valeurs de N. En faisant la moyenne de ces coefficients, nous avons observé que les valeurs obtenues pour N=25 maximisaient cette corrélation moyenne, autrement dit qu'elles étaient les plus représentatives des autres valeurs considérées pour N. Toutes les mesures présentées dans la suite de l'article utilisent donc cette valeur de N.

4.2 Variation moyenne

Pour chacun de nos trois corpus, nous avons mesuré la variation (sur les 25 premiers voisins, comme indiqué précédemment) entre chaque paire de modèles (5 modèles, 10 paires) et pour chaque mot du corpus atteignant le seuil minimal de 100 occurrences mais en nous limitant pour la suite aux seules classes ouvertes (adjectifs, adverbes, substantifs, noms propres et verbes). La taille du vocabulaire ainsi sélectionné dans chaque corpus est indiquée en table 2, avec les scores moyens de variation, l'écart-type moyen mesuré sur les 10 paires de modèles et enfin l'écart-type de la variation moyenne mesurée pour chaque mot.

Corpus	Vocabulaire	Var_{25} moyen	Moyenne de l'écart type	Écart-type des moyennes
ACL	22 292	0,16	0,04	0,08
BNC	27 434	0,17	0,04	0,07
PLOS	31 529	0,18	0,05	0,07

TABLE 2. Taux de variation global pour chaque corpus, calculé sur l'ensemble du vocabulaire et sur les 10 paires de modèles comparés

Globalement on observe un taux de variation moyen autour de 0,17. Rappelons que cela signifie que, lorsque l'on extrait les 25 mots les plus proches d'un mot-cible, 4 à 5 d'entre eux ne se retrouvent pas d'un modèle à l'autre. Ce taux semble stable à travers les 10 paires de modèles comparés pour chaque corpus (écart-type de 0,04 en moyenne).

Il existe par contre des différences plus importantes de cette même variation d'un mot à l'autre.

L'écart-type de la variation moyenne est environ de 0,07 (sur plus de 20 000 mots) : certains mots atteignent des scores de variation de 0,8, alors que d'autres ont des scores nuls (indiquant qu'ils ont les mêmes 25 premiers voisins dans tous les modèles, indépendamment de leur ordre).

De plus, il apparaît clairement que la variabilité est bien liée au mot et pas au modèle, puisqu'à travers les 10 paires de modèles comparés les scores de variation sont très stables : le coefficient de corrélation moyen (de Spearman, sur plus de 20 000 mots) entre les 10 paires est de 0,74 pour ACL, 0,72 pour le BNC et 0,77 pour PLOS. Autrement dit, il existe de façon inhérente des mots pour lesquels *Word2vec* va proposer les mêmes voisins dans chaque modèle (pour une configuration et un corpus donnés) et d'autres pour lesquels les voisins seront très sensibles aux aléas de la méthode.

4.3 Variation des scores d'évaluation sur les *benchmarks*

Bien que les méthodes d'évaluation interne classiques soient très critiquées (cf. section 2.1), notamment pour leur taille, l'imprécision des mesures humaines et leur faible taux d'accord inter-annotateur, nous avons voulu étudier l'instabilité de nos modèles face à celles-ci. Nous avons calculé le score de chacun de nos modèles sur les jeux de test WordSim353 (Finkelstein *et al.*, 2002) et Simlex-999 (Hill *et al.*, 2015). La table 3 donne les valeurs minimales et maximales obtenues pour nos trois séries de 5 modèles. La variation de ces scores d'un modèle à un autre sur un même corpus existe bien mais est relativement faible, de l'ordre de 1 à 4% du score (en relatif) et est très inférieure à la variation que l'on peut observer en comparant les modèles obtenus sur des corpus différents (où la variation va de 7 à 36%). Rappelons que de tels scores sont obtenus par la corrélation (des rangs) des scores de similarités avec ceux obtenus en interrogeant des humains sur quelques centaines de paires de mots.

Corpus	WordSim353 (min–max)	Simlex-999 (min–max)
ACL	0.592 – 0.601	0.192 – 0.201
BNC	0.631 – 0.639	0.306 – 0.312
PLOS	0.392 – 0.403	0.273 – 0.279

TABLE 3. Variation des scores sur deux jeux d'évaluation

Nous avons également profité du jeu de test WordSim353 pour regarder dans quelle mesure les scores de similarité (cosinus) entre deux mots varient d'un modèle à l'autre. Le choix de ne regarder que ces 353 paires de mots est dû au temps de calcul nécessaire pour comparer ces scores sur l'ensemble des mots du corpus (près d'un milliard de paires de mots par corpus). Nous avons observé une variation moyenne de ces scores de l'ordre de 4% de leur valeur d'un modèle à un autre, donc légèrement plus importante en valeur relative que les scores globaux du banc de test. Mais là aussi on observe des différences entre les items du jeu de test, sans d'ailleurs que l'ampleur de la variation soit liée à la proximité estimée par les humains entre les deux mots.

Enfin, il se trouve que le vocabulaire utilisé dans ces jeux de test a un taux de variation significativement plus faible que le reste (0.16 vs 0.17, test de Student, $p<0,05$, avec une variation de 1 à 2% en absolu selon les corpus). Il est donc raisonnable d'estimer que les scores de cosinus de ces paires varierait légèrement moins que l'ensemble des autres. Dans tous les cas, il semble clair que ces jeux de test ne permettent pas de prendre en compte à sa juste mesure l'ampleur du phénomène de variabilité interne des modèles.

5 Exploration de la variation

Comme nous l'avons remarqué précédemment, la variation n'est pas homogène à travers le lexique, puisque certains mots ont un score de variation nul et pour certains autres la valeur monte jusqu'à 0.8. Nous avons donc procédé à différentes explorations de ces différences, en cherchant à identifier ce qui permettrait de distinguer les mots stables des mots instables. Dans cette dernière section, nous abordons donc tout d'abord la question en regardant si certaines caractéristiques simples des mots comme leur fréquence ou leur catégorie morphosyntaxique étaient corrélées à leur variation, puis en examinant plus précisément les deux extrémités du spectre de la variation interne.

5.1 Impact de la fréquence et de la catégorie

Nous avons tout d'abord regardé la variation du score de variation en fonction de la fréquence du mot dans le corpus et sa catégorie morphosyntaxique. Différentes études ont pu montrer que la fréquence d'un mot dans un corpus influence la qualité de sa représentation dans les espaces distributionnels et qu'on obtient de meilleurs résultats suivant les différentes méthodes d'évaluation pour les mots de haute fréquence et par corrolaire sur des corpus de grande taille (Sahlgren & Lenci, 2016).

La figure 2 montre le score de variation moyen obtenu sur nos trois corpus pour différentes classes de fréquence logarithmique. Si une tendance linéaire simple ne semble pas se dégager aussi clairement qu'attendu, on peut résumer le lien en indiquant que ce sont les mots de fréquence intermédiaire (entre 1000 et 10 000 occurrences) qui sont les plus stables. En-deçà ou au-delà de cette zone il y a une légère augmentation de la variation.

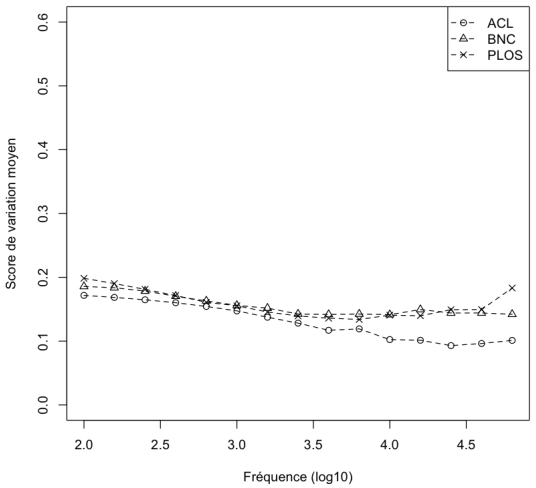

FIGURE 2. Effet de la fréquence sur le score moyen de variation pour ACL, BNC et PLOS

Pour ce qui est de la catégorie morphosyntaxique, on observe une grande homogénéité d'une partie du discours à l'autre. Seuls les noms propres ont une variation légèrement supérieure aux autres catégories, et ce pour les trois corpus.

Il semblerait donc que la variation soit à rechercher en lien avec des caractéristiques plus fines des mots. Pour cela, nous avons eu recours à une observation directe des mots stables et instables.

5.2 Clusters stables

A l'œil nu, en parcourant la liste ordonnée des mots les plus stables, on observe rapidement des régularités sous la forme de classes de mots apparemment similaires et ce, pour chacun des trois corpus (avec cependant des classes différentes d'un corpus à l'autre). C'est donc dans la direction du répérage de ces classes de mots avec une faible variation que nous avons orienté notre investigation.

Si l'on se penche sur le voisinage des mots, on observe plusieurs faits intéressants. Tout d'abord, le score de proximité du premier voisin n'est que partiellement corrélé avec le score de variation (-0,40 en moyenne). Cela signifie que le fait qu'un mot ait un premier voisin très proche explique en partie que son voisinage va rester stable. Il est en effet logique que ce proche voisin résiste aux aléas et qu'il fasse donc partie du voisinage stable de ce mot. En même temps, cela n'est absolument pas systématique, et on trouve de nombreux cas d'instabilité alors que des mots ont des voisins proches, et vice-versa. Ensuite, le score de variation d'un mot stable est corrélé avec la stabilité de ses voisins. Si l'on calcule en effet, lorsque l'on compare deux modèles, la variation moyenne des voisins (en faisant l'union des 25 plus proches voisins d'un mot dans chacun des deux modèles), on obtient une corrélation de 0,5 en moyenne avec le score de variation de ce mot. Cela confirme un mécanisme logique : la stabilité d'un mot entre deux modèles similaires est dûe à la présence de zones stables dans les espaces sémantiques. Il est donc possible d'identifier ces zones de stabilité.

En utilisant la variation mesurée pour chaque paire de modèles, nous avons calculé la variation globale moyenne d'un mot. Nous avons ensuite sélectionné pour chaque corpus les 300 mots montrant la variation moyenne la plus basse et pour chaque modèle entraîné, nous avons calculé leur similarité deux à deux dans chacun des cinq modèles. Nous avons appliqué une classification hiérarchique ascendante, en fixant le nombre de clusters à 10, et avons ainsi obtenu 5 partitions différentes. Pour tester la fiabilité de ces clusters, nous avons calculé l'indice de Rand (Rand, 1971) entre les 5 partitions générées (séparément pour chaque corpus). L'indice de Rand moyen était de 0,93 ($\pm$ 0,01, IC 95%) pour les 3 corpus, confirmant la grande stabilité des clusters extraits. Ceux que nous avons identifés et qualifiés sont indiqués dans la table 4.

On peut voir dans cette table des zones du lexique pour lesquelles on s'attend effectivement à une grande efficacité des méthodes distributionnelles, surtout lorsque les contextes sont restreints, spécifiques et réguliers. Les classes de co-hyponymes apparaissent clairement et répondent bien aux principes de base de la sémantique distributionnelle. Certaines classes enfin sont propres à nos corpus scientifiques comme le lexique transdiciplinaire, dont il a été démontré qu'il était bien capté par ces méthodes (Tutin, 2007). Il s'agit donc au final de zones lexicales classiques au sens de ce que permet la sémantique distributionnelle : on notera que ces clusters sont de taille variée, allant de quelques éléments pour les mesures de performance à plusieurs dizaines pour les ordinaux ou les références internes, ce qui nous permet d'envisager qu'ils ne sont pas dus au biais du nombre de voisins considérés.

5.3 Mots instables

A l'opposé du spectre de la variation, il est bien entendu beaucoup plus difficile d'identifier des régularités par une telle méthode. Puisque le principe même du clustering est inapproprié nous avons eu recours à une observation directe des 300 mots les plus instables et identifié plusieurs cas dont certains sont présentés en table 5.

Type de cluster	Corpus	Exemples
Contextes locaux spécifiques	ACL	mots de langues étrangères utilisés dans les exemples (*para, com, sobre...*), (*der, das, nicht, die...*) ordinaux (*12th, eleventh, 41st...*)
	PLOS	renvois internes (*figures, table, 6b, 1a...*) descriptions de figures (*dot, triangle, filled, orange...*)
	BNC	expressions temporelles (*am, pm, 31st, noon...*) mesures dans les recettes de cuisine (*tsp, tbsp, oz...*)
Classes fermées de co-hyponymes	ACL	mesures de performances (*precision, recall, f-score...*) traitements (*parsing, lemmatization, tokenizing...*)
	PLOS	antibiotiques (*puromycin, blasticidin, cefotaxime...*) voies d'administration (*intraperitoneally, intranasal, intramuscular...*)
	BNC	famille (*wife, grandmother, son, sister...*) pièces et objets de la maison (*kitchen, sitting-room, bathroom, furniture...*)
Phraséologie scientifique	ACL	adverbes de connection (*nevertheless, relatively, secondly, additionally...*) processus scientifique (*discuss, describe, observe...*)
	PLOS	adverbes de connection (*moreover, furthermore, conversely...*) processus scientifique (*hypothetize, reason, elucidate...*)

TABLE 4. Exemples de clusters stables identifiés pour chaque corpus

Parmi les mots les plus instables non reportés dans la table 5, nous avons également repéré des erreurs d'étiquetage, ainsi que des mots de basse fréquence. Cependant, les noms propres en général sont très présents parmi les mots les plus instables dans les trois corpus, qu'il s'agisse de patronymes (ACL et BNC) ou de sigles (PLOS). Pour les autres classes que nous proposons, on observe des génériques du domaine dont la fréquence est élevée et qui ont beaucoup d'hyponymes lointains mais pas de synonymes ni de voisins privilégiés. Nous observons également des adjectifs génériques ainsi que des mots très polysémiques, ces derniers étant plus difficiles à repérer, pour lesquels le mécanisme distributionnel peine effectivement à dégager des régularités dans les contextes d'emplois. Il est intéressant de noter toutefois que cela ne signifie pas que ces mots n'ont pas des voisins pertinents et globalement bien identifiés par les modèles, mais que ceux-ci sont noyés dans un environnement de très haute variation. Par exemple, un adjectif très instable comme *super* dans le BNC a bien des voisins sémantiquement pertinents à travers les différentes instances des modèles (comme *fabulous*,

Corpus	Séries	Exemples
ACL	noms propres	*Steve*, *Joyce*, *Ivan...*
	génériques du domaine	*language*, *sign*
	adjectifs génériques	*free*, *mix*, *special*
	mots polysémiques	*account*, *card*, *zone*, *sign*
PLOS	noms propres	*PCB*, *DMC*, *TLP*, *ACD ...*
	génériques du domaine	*gene*, *cell*, *protein*
	adjectifs génériques	*free*, *current*, *near*, *double*
BNC	noms propres	*Bart*, *Vince*, *Lewis...*
	adjectifs génériques	*whole*, *general*, *super*
	mots polysémiques	*make*, *close*, *cast*

TABLE 5. Exemples de classes de mots instables identifiées pour chaque corpus

stylish, *stunning* et quantité d'autres laudatifs) mais on y trouve en très bonne place de nombreux parasites extrêmement dispersés et variant d'un modèle à l'autre (comme *Granada*, *cracker*, *zeppelin*).

6 Conclusion et perspectives

Cette étude nous a permis de mesurer la stabilité interne des modèles produits par *Word2vec* dans son paramétrage par défaut. Globalement 17% des 25 mots les plus proches sont susceptibles d'être différents, uniquement à cause des facteurs aléatoires intervenant dans la méthode. Ce taux est toutefois inférieur pour certains clusters denses de mots pour lesquels le mécanisme distributionnel est très robuste. Par contre, certaines séries de mots semblent très sensibles à l'instabilité (noms propres, mots génériques du domaine). Cette instabilité inhérente est malheureusement souvent ignorée et est d'ailleurs minime quand on compare deux modèles sur des bancs de test classiques. À ce stade, nous n'avons pas estimé l'impact de cette variabilité sur des tâches externes de TAL qui se basent sur les embeddings (parsing, extraction d'information, classification de documents, etc.) mais il est probable que l'effet soit minime, ou même qu'il se perde dans la variabilité inhérente aux outils qui utilisent ces données et ont de grandes chances d'être soumis aux mêmes variations aléatoires.

Notre point de vue sur la question est plus orienté vers l'utilisation des embeddings comme outil d'exploration sémantique de mots ou de corpus. C'est cette orientation qui nous a guidés vers la mesure de la variation des plus proches voisins, l'observation de ceux-ci étant le mode principal d'investigation de la représentation sémantique d'un mot dans de tels espaces vectoriels. Au vu de l'ampleur du phénomène, nous rejoignons les conclusions et consignes de Antoniak & Mimno (2018) sur les précautions avec lesquelles aborder les résultats. Si la solution la plus directe est de mutliplier les modèles avant de tirer des conclusions, nous espérons toutefois parvenir à une procédure d'identification de ce qui pourrait expliquer la stabilité relative de la représentation d'un mot, et par là même prédire à terme la fiabilité de celle-ci. De premiers résultats encourageants dans ce sens sont présentés dans (Pierrejean & Tanguy, 2018), mais de nombreuses questions restent à ce stade en suspens, notamment concernant le rôle des différents facteurs sur cette variabilité (taille du corpus et hyperparamètres), ainsi que l'utilisation d'une mesure plus fine de la variation qui prendrait l'ordre des voisins en considération.

Au-delà de cette instabilité "interne" des modèles, notre objectif est également de nous pencher sur la variation observable entre des modèles différents et au premier chef lorsque l'on passe d'un corpus à l'autre, à l'instar de ce qu'ont fait Hamilton *et al.* (2016) en étudiant les changements diachroniques. C'est cet objectif initial qui nous avait fait choisir des corpus différents comme ACL et PLOS. Il est bien entendu possible de mesurer par le même recouvrement quels sont les mots (ou classes) dont la représentation change le plus. Toutefois, on ne peut aborder ces questions sans avoir au préalable une estimation de la part du hasard dans chacun des modèles. Plus généralement, c'est une meilleure compréhension des mécanismes et des limites de ces outils qui permettra de les intégrer dans des travaux d'investigation plus fins et plus fiables en linguistique de corpus. Si nous reconnaissons comme bien d'autres leur spectaculaire efficacité et leur facilité d'utilisation, elles ne doivent pas nous en masquer les limites et les défauts.

Remerciements

Les expériences présentées dans cet article ont été réalisées en utilisant la plateforme OSIRIM administrée par l'IRIT et soutenue par le CNRS, la région Midi-Pyrénées, le gouvernement français, et le FEDER (voir `http://osirim.irit.fr/site/fr`).

Références

ANTONIAK M. & MIMNO D. (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, **6**, 107–119.

ASR F. T., WILLITS J. A. & JONES M. N. (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. In *Proceedings of the 37th Meeting of the Cognitive Science Society*.

BERNIER-COLBORNE G. & DROUIN P. (2016). Evaluation of distributional semantic models : a holistic approach. In *Proceedings of the 5th International Workshop on Computational Terminology*, p. 52–61, Osaka, Japan.

BIRD S., DALE R., DORR B., GIBSON B., JOSEPH M., KAN M.-Y., LEE D., POWLEY B., RADEV D. & FAN TAN Y. (2008). The ACL Anthology Reference Corpus : A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of Language Resources and Evaluation Conference (LREC 08)*, Marrakesh, Morrocco.

CHIU B., CRICHTON G., KORHONEN A. & PYYSALO S. (2016). How to Train Good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, p. 166–174, Berlin, Germany.

FARUQUI M., TSVETKOV Y., RASTOGI P. & DYER C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, p. 30–35.

FINKELSTEIN L., GABRILOVICH E., MATIAS Y., RIVLIN E., SOLAN Z., WOLFMAN G. & RUPPIN E. (2002). Placing Search in Context : The Concept Revisited. *ACM Transactions on Information Systems*, **20**(1), 116–131.

GOLDBERG Y. (2016). A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*, **57**, 345–420.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of ACL 2016*, p. 1489–1501.

HELLRICH J. & HAHN U. (2016). Bad Company - Neighborhoods in Neural Embedding Spaces Considered Harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 2785–2796, Osaka, Japan.

HILL F., REICHART R. & KORHONEN A. (2015). SimLex-999 : Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, **41**, 665–695.

HUTSON M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, **359**(6377), 725–726.

LEVY O. & GOLDBERG Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, p. 302–308, Baltimore, Maryland, USA.

LI B., LIU T., ZHAO Z., TANG B., DROZD A., ROGERS A. & DU X. (2017). Investigating Different Syntactic Context Types and Context Representations for Learning Word Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2411–2421.

MELAMUD O., MCCLOSKY D., PATWARDHAN S. & BANSAL M. (2016). The Role of Context Types and Dimensionality in Learning Word Embeddings. In *Proceedings of NAACL-HLT 2016*, San Diego, California.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, **abs/1301.3781**.

PIERREJEAN B. & TANGUY L. (2018). Predicting word embeddings variability. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM 2018)*. To appear.

RADOVANOVIĆ M., NANOPOULOS A. & IVANOVIĆ M. (2010). Hubs in Space : Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, **11**, 2487–2531.

RAND W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

SAHLGREN M. (2005). An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE)*.

SAHLGREN M. (2006). *The Word-Space Model*. PhD thesis, Gothenburg University.

SAHLGREN M. & LENCI A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 975–980, Austin, Texas.

SANDVE G. K., NEKRUTENKO A., TAYLOR J. & HOVIG E. (2013). Ten simple rules for reproducible computational research. *PLoS computational biology*, **9**(10).

SCHNABEL T., LABUTOV I., MIMNO D. & JOACHIMS T. (2015). Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 298–307.

TROST T. A. & KLAKOW D. (2017). Parameter free hierarchical graph-based clustering for analyzing continuous word embeddings. In *Proceedings of TextGraphs-11 : the Workshop on Graph-based Methods for Natural Language Processing*.

TUTIN A. (2007). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2007)*, p. 283–292, Toulouse, France.

URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 188–201, Les Sables d'Olonne, France.