



HAL
open science

SINr-filtered: Favoriser l'émergence du sens en filtrant les communautés extraites des réseaux de cooccurrences de mots

Anna Béranger, Nicolas Dugué, Simon Guillot, Thibault Prouteau

► To cite this version:

Anna Béranger, Nicolas Dugué, Simon Guillot, Thibault Prouteau. SINr-filtered: Favoriser l'émergence du sens en filtrant les communautés extraites des réseaux de cooccurrences de mots. Extraction et Gestion des Connaissances, Jan 2024, Dijon, France. pp.429-430. hal-04470451

HAL Id: hal-04470451

<https://hal.science/hal-04470451v1>

Submitted on 21 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

SIN_r-filtered : Favoriser l'émergence du sens en filtrant les communautés extraites des réseaux de cooccurrences de mots

Anna Béranger*, Nicolas Dugué*
Simon Guillot*, Thibault Prouteau*

*Le Mans Université, LIUM, avenue Olivier Messiaen, 72085 Le Mans

SIN_r. La représentation vectorielle du lexique est une problématique classique du traitement automatique du langage naturel. Les plongements lexicaux sont des vecteurs dans un espace où deux mots sémantiquement proches ont des représentations peu éloignées. Les méthodes telles que *Word2vec* utilisent des réseaux de neurones sur les cooccurrences des mots pour construire ces représentations. Les *transformers* ont démontré leur pertinence en contextualisant les vecteurs avec des architectures de plus en plus grandes. *SIN_r* (Sparse Interpretable Node Representations) est introduit par Prouteau et al. (2021) pour ne pas se concentrer uniquement sur des objectifs de performance, mais pour entraîner des vecteurs de mots **interprétables** de façon **frugale** (Prouteau et al., 2022). Pour cela, *SIN_r* s'appuie sur un graphe de cooccurrences pondéré : les nœuds représentent les mots et sont connectés entre eux par des arêtes valuées en fonction de leur nombre de cooccurrences dans le corpus. Des communautés sont alors détectées sur le graphe de cooccurrences et utilisées comme dimensions de l'espace latent : le vecteur d'un mot est extrait en utilisant ses liens avec les communautés extraites. En accord avec l'hypothèse distributionnelle, Prouteau et al. (2021) considèrent que les mots qui distribuent leurs liens de manière similaire sur les communautés ont un sens similaire. Or, si la méthode est à l'état de l'art de l'interprétabilité tout en étant considérablement plus frugale que les approches concurrentes comme *SPINE*, et que *SIN_r* obtient des performances en similarité identiques à *SPINE*, celles-ci sont néanmoins légèrement inférieures à *Word2vec*.

De SIN_r à SIN_r-filtered. Nous proposons une approche qui filtre les communautés de *SIN_r* afin d'améliorer significativement les performances de la méthode. L'approche, nommée *SIN_r-filtered*, rattrape les performances de *Word2vec* et réduit l'empreinte mémoire des représentations extraites tout en conservant l'interprétabilité du modèle et sa frugalité. Pour ce faire, nous nous basons sur l'activation des dimensions. Une dimension est dite activée par un mot si, pour cette dimension, la composante du vecteur de ce mot est non-nulle. L'activation des dimensions par les vecteurs de mots suit une loi de puissance (Figure 1). Les **dimensions très activées** sont le résultat de communautés regroupant des mots très fréquents qui apparaissent dans des contextes très divers, elles présentent difficilement une cohérence sémantique. Ainsi, nous filtrons ces dimensions, et nous **améliorons la performance du modèle**. Par ailleurs, beaucoup de **dimensions sont très peu activées**. Nous filtrons ces dimensions : cela **divise l'empreinte mémoire du modèle par 5** en préservant les performances.

SIN_r -filtered : filtrer les communautés pour mieux découvrir le sens

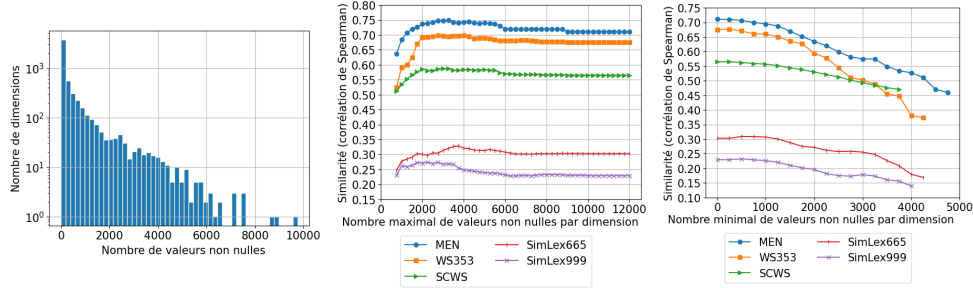


FIG. 1 – Les trois figures résultent du corpus $UkWac$. Distribution du nombre d’activations à gauche. Au centre et à droite, score de similarité. Les dimensions les **plus** (resp. **moins**) activées sont retirées d’après le seuil en abscisse au centre (resp. à droite).

Résultats. Les résultats sur la tâche de similarité Table 1 pour les corpus d’apprentissage BNC et $UkWac$ montrent que SIN_r -filtered devance systématiquement SIN_r , et surpasse $W2V$ sur le corpus de similarité $WS353$. Ces résultats sont obtenus avec les seuils de filtrage optimaux déterminés Figure 1. Le seuil de 4000 activations maximum par dimensions (Figure 1, centre) permet de gagner 5 points sur MEN. Le seuil de 500 activations au minimum (Figure 1, droite) ne dégrade pas les performances, et les vecteurs obtenus à l’issue de son application sont de taille 5 fois inférieurs à ceux de SIN_r , soit environ 1000 dimensions.

	MEN		WS353		SCWS	
	BNC	$UkWac$	BNC	$UkWac$	BNC	$UkWac$
$W2V$.73	.75	.64	.66	.61	.64
SIN_r	.67	.70	.63	.68	.56	.56
SIN_r -filtered	.72	.75	.65	.70	.58	.59

TAB. 1 – Corrélations de Spearman moyennes sur 10 exécutions de la tâche de similarité.

Reproductibilité. La méthodologie est décrite exhaustivement dans (Béranger et al., 2023). De plus, SIN_r et SIN_r -filtered sont disponibles via un **package Python** (<https://github.com/SINr-Embeddings/sinr>) qui permet d’apprendre des plongements, de les évaluer et d’en interpréter les dimensions à partir de grands corpus textuels sur un simple laptop sans GPUs.

Références

- Béranger, A., N. Dugué, S. Guillot, et T. Prouteau (2023). Filtering communities in word co-occurrence networks to foster the emergence of meaning. In *Complex Networks*.
- Prouteau, T., V. Connes, N. Dugué, A. Perez, et al. (2021). Sinr : fast computing of sparse interpretable node representations is not a sin ! In *IDA*, pp. 325–337.
- Prouteau, T., N. Dugué, N. Camelin, et S. Meignier (2022). Are embedding spaces interpretable? results of an intrusion detection evaluation on a large french corpus. In *LREC*.