SeeMantics : un jeu lexical pour l'interprétabilité des vecteurs de mots

Simon Guillot[1, 2] Anna Béranger[1] Nicolas Dugué[1]
(1) LIUM, Le Mans Université, Avenue Laennec, 72085 Le Mans CEDEX 9 France
(2) ERTIM, INaLCO, 2 rue de Lille, 75007 Paris
prénom.nom@univ-lemans.fr

MOTS-CLÉS : représentations vectorielles, interprétabilité, GWAP, traits sémantiques.

KEYWORDS: word embedding, interprtability, GWAP, semantic features.

1 Introduction

La nécessité de pouvoir comprendre et auditer des technologies d'apprentissage automatique de plus en plus performantes mais opaques a récemment fait émerger les notions d'explicabilité et d'interprétabilité. Suivant (Broniatowski, 2021), nous distinguons l'explicabilité, mettant l'emphase sur la compréhension mécanistique du comportement interne d'un modèle pour un expert, et l'interprétabilité, relevant de la capacité d'une personne utilisatrice quelconque à faire sens de la sortie d'un modèle. Ces deux notions peuvent être utilisées pour considérer le traitement de l'information dans le modèle, ou bien adresser la représentation des données dans le modèle (Gilpin *et al.*, 2018). Le présent travail s'intéresse plutôt à cette seconde acception de l'interprétabilité portée sur la représentation du lexique dans les systèmes de TAL.

Les représentations du lexique prennent depuis une dizaine d'années la forme de représentations vectorielles (plongements lexicaux), apprises par des modèles faisant usage de l'hypothèse distributionnelle (Harris, 1954; Firth, 1957) et constituent déjà une méthode d'investigation pour des travaux en linguistique ou en humanités numériques (Hamilton *et al.*, 2016, 2018; Dubossarsky *et al.*, 2015; Rotaru *et al.*, 2018; Kozlowski *et al.*, 2019). Dans ce travail, nous proposons plus spécifiquement d'étendre la notion d'interprétabilité, depuis une définition reposant sur l'interprétabilité des dimensions individuelles d'un espace de représentation, à l'interprétabilité des vecteurs de mots en eux-mêmes. La manipulation et la composition de dimensions pour définir des éléments du lexique admet par ailleurs une similarité importante avec les sémantiques lexicales faisant usage de traits sémantiques ou sèmes (Pottier, 1963; Rastier, 2009) émergeants d'un corpus, en opposition à des travaux utilisant des traits sémantiques supposés universels s'exprimant dans les corpus (Şenel *et al.*, 2018; Chersoni *et al.*, 2021). L'interprétabilité des représentations lexicales apparaît donc

comme une interface possible entre modèles en traitement de la langue naturelle et linguistique en rapprochant les notions de dimension interprétable d'un espace de représentation du lexique et de trait sémantique.

2 Revue de littérature et objectif de recherche

L'article séminal de Murphy *et al.* (2012) ouvre la littérature des représentations interprétables en caractérisant ces dernières comme respectant des contraintes de parcimonie, de positivité et de performance. La contrainte de parcimonie vise à ce que les dimensions de l'espace vectoriel issu de l'apprentissage des plongements ne soient utilisées dans la représentation que d'une petite partie du lexique pour qu'elles puissent correspondre à des thèmes cohérents. La positivité tient à ce qu'il est improbable cognitivement que des faits négatifs soient stockés en rapport à des thématiques. La performance enfin intervient comme critère de contrôle, des représentations interprétables ne sont pas très intéressantes si elles n'ont pas les qualités de représentations des modèles plus opaques. Un certain nombre de modèles ont adopté ces critères comme SPINE (Subramanian *et al.*, 2017) ou SINr (Prouteau *et al.*, 2021).

L'évaluation de l'interprétabilité des représentations issues de ces méthodes s'est jusqu'à présent concentrée sur les dimensions des espaces de représentations. En particulier, la tâche de *word intrusion* (Chang *et al.*, 2009) a émergé comme évaluation standard (Subramanian *et al.*, 2017; Prouteau *et al.*, 2022). Cette évaluation porte néanmoins une conception assez faible de l'interprétabilité si l'on s'en réfère aux travaux normatifs sur le sujet (Broniatowski, 2021). D'une part, elle n'adresse pas la sortie du système (les vecteurs de mots en eux-mêmes), d'autre part, elle ne vérifie la capacité d'un locuteur non expert en intelligence artificielle à faire sens de ces sorties. Nous proposons de redéfinir l'interprétabilité des représentations du lexique en nous concentrant sur la capacité des locuteurs à faire sens des vecteurs représentant les mots. Les vecteurs de mots interprétables étant définis dans des espaces de très grande dimension, il est impossible de les présenter tels quels à des locuteurs. Pour considérer l'interprétabilité de ces vecteurs de mots, il est nécessaire de trouver une modalité de présentations de l'information contenue dans ces vecteurs de très grande dimension qui reste utilisable par un locuteur (Guillot *et al.*, 2023). Les présents travaux visent à **caractériser les conditions de présentations optimales quant à leur interprétabilité de vecteurs en utilisant un jeu lexical en ligne pour obtenir des annotations**.

3 Méthodologie

Les *games with a purpose* ou GWAPs (von Ahn, 2006) constituent un moyen d'employer de la "puissance de calcul humaine collective" en présentant des tâches de manière ludique via l'outil informatique. Ces jeux sont particulièrement adaptés à la constitution de res-

sources et la production d'annotations nécessaires en traitement automatique de la langue (Lafourcade & Joubert, 2008; Poesio *et al.*, 2013) en utilisant les statistiques de jugements de nombreux locuteurs non experts, plus disponibles et plus robustes que le recours à des experts linguistiques. Nous proposons ici d'employer cette notion de GWAP en utilisant des métriques obtenues sur un jeu lexical dérivé d'exemples contemporains très populaires tels que Semantle [1], Cémantix [2] et Pimantle [3].

Le jeu lexical SeeMantics a été développé au LIUM comme prototype de démonstration de modèles de plongements lexicaux interprétables. L'objectif du jeu est de trouver le mot du jour (ou de la session en cours). Le mot proposé par le joueur ou la joueuse est situé sur une échelle de distance par rapport au mot cible de la session (captures d'écran Fig 1). Une partie se joue en essayant de trouver le mot du jour en s'aidant de ses précédentes tentatives par associations sémantiques. Les distances sont calculées par distance cosine entre les vecteurs de mots d'un modèle de représentation sous-jacent. La spécificité de SeeMantics est d'utiliser un modèle interprétable, SINr, et d'employer les particularités de ce modèle pour donner des indices supplémentaires aux joueurs : les dimensions les plus importantes dans la représentation du mot cible. Nous appelons **stéréotypes** l'ensemble de mots activant le plus fortement une dimension donnée, et nous servons de ceux-ci pour illustrer les dimensions.

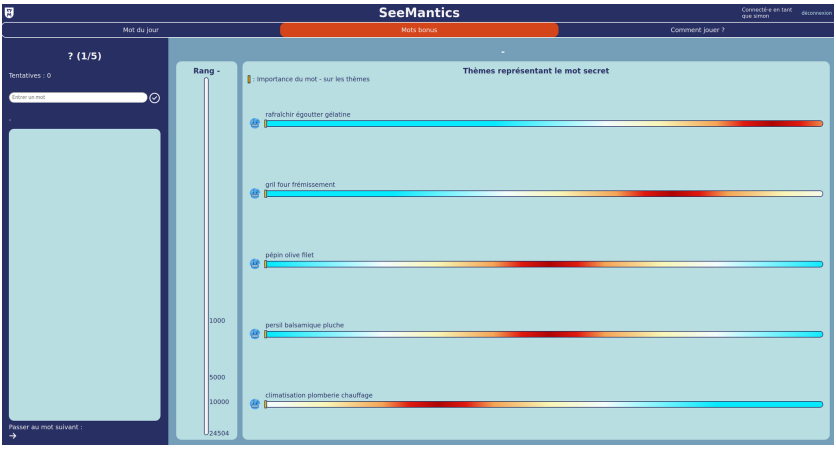

FIGURE 1 – Une partie de SeeMantics où deux dimensions sont proposées, caractérisées par sept stéréotypes chacune, présentant donc en tout quatorze indices verbaux. L'interprétation de cette modalité est difficile selon H0, mais admise par H2, et même facilitée selon H1.

Afin d'estimer les modalités de présentation optimales des vecteurs de mots, nous prenons en compte deux facteurs : le nombre de dimensions proposées aux joueurs et le nombre de stéréotypes utilisés pour caractériser une dimension (voir Table 1). L'ensemble des

1. https ://semantle.com/
2. https ://cemantix.certitudes.org/
3. https ://semantle.pimanrul.es/

dimensions proposées est donc caractérisé par plusieurs groupes de stéréotypes qui ensembles constituent les **indices verbaux** fournis aux joueurs.

N° dim \ N° stéréo	2	3	4	5	7
2				♡	
3		♡		♣	♠
5	♡	♣	♠		
7	♣	♠			

TABLE 1 – Récapitulatif des modalités de présentations des indices verbaux avec ♡ représentant les modalités environ 10 indices, ♣ représentant les modalités à environ 15 indices et ♠ représentant les modalités à environ 20 indices

Le nombre d'essais nécessaires avant de trouver le mot cible détermine le **score de réussite** nous informant sur l'optimalité des modalités de présentations. Une information de meilleure qualité et plus facile à manipuler devrait permettre de trouver le mot cible en moins d'essais qu'une information insuffisante ou trop confuse.

4 Résultats anticipés et résultats préliminaires

Des contraintes de mémoires de travail entravent la manipulation de trop nombreux items verbaux (Miller, 1956; Peterson & Peterson, 1959). Les différentes combinaisons de nombre de dimensions présentées, et de nombre de stéréotypes (voir Table 1) utilisés pour caractériser ces dimensions, devraient aboutir à des scores de réussite différents, en fonctions du dépassement ou non de cette limite hypothétique (autour de la dizaine de mots). Nous posons néanmoins deux hypothèses précisant les résultats anticipés plus avant, et contredisant possiblement la précédente hypothèse (**H0**) :

— **H1** : des dimensions mieux définies (par plus de mots) apportent davantage d'information qu'un grand nombre de dimensions à nombre de mots utilisés comme indice égal ;

— **H2** : les joueurs synthétisent les dimensions bien définies sous la forme d'un label et sont donc à même de manipuler plus de dimensions qu'ils ne devraient suivant H0.

Les hypothèses H1 et H2 pointent vers l'apparition, pour les joueurs, d'une structure synthétisant les indices verbaux qui leurs sont présentés en des unités sémantiques stabilisées par la cohérence d'un petit champ lexical (les stéréotypes). La capacité pour un locuteur de combiner ces unités sémantiques pour trouver un mot-cible sert de proxy pour mesurer l'interprétabilité de représentations reposant sur l'association de ces unités comme dimensions des vecteurs de mots. Cette expérience de manipulation et de combinaison de traits sémantiques est par ailleurs une tentative d'ancrage des sémantiques à traits dans un cadre expérimental mobilisant l'hypothèse distributionnelle et de l'annotation humaine.

Références

BRONIATOWSKI D. A. (2021). *Psychological Foundations of Explainability and Interpretability in Artificial Intelligence*. Rapport interne, National Institute of Standards and Technology. DOI : 10.6028/NIST.IR.8367.

CHANG J., GERRISH S., WANG C., BOYD-GRABER J. L. & BLEI D. M. (2009). Reading Tea Leaves : How Humans Interpret Topic Models.

CHERSONI E., SANTUS E., HUANG C.-R. & LENCI A. (2021). Decoding Word Embeddings with Brain-Based Semantic Features. *Computational Linguistics*, **47**(3), 663–698. DOI : 10.1162/coli_a_00412.

DUBOSSARSKY H., TSVETKOV Y., DYER C. & GROSSMAN E. (2015). A bottom up approach to category mapping and meaning change. In *NetWordS*, p. 66–70.

FIRTH J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*. Publisher : Basil Blackwell.

GILPIN L. H., BAU D., YUAN B. Z., BAJWA A., SPECTER M. A. & KAGAL L. (2018). Explaining explanations : An approach to evaluating interpretability of machine learning. *CoRR*, **abs/1806.00069**.

GUILLOT S., PROUTEAU T. & DUGUE N. (2023). Sparser is better : one step closer to word embedding interpretability.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Cultural Shift or Linguistic Drift ? Comparing Two Computational Measures of Semantic Change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, **2016**, 2116–2121.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2018). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv :1605.09096 [cs]*. arXiv : 1605.09096.

HARRIS Z. S. (1954). Distributional Structure. *WORD*, **10**(2-3), 146–162. DOI : 10.1080/00437956.1954.11659520.

KOZLOWSKI A. C., TADDY M. & EVANS J. A. (2019). The Geometry of Culture : Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, **84**(5), 905–949. Publisher : SAGE Publications Inc, DOI : 10.1177/0003122419877135.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes.

MILLER G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review*, **63**(2), 81–97.

MURPHY B., TALUKDAR P. & MITCHELL T. (2012). Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding.

PETERSON L. & PETERSON M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, **58**(3), 193. DOI : 10.1037/h0049234.

POESIO M., CHAMBERLAIN J., KRUSCHWITZ U., ROBALDO L. & DUCCESCHI L. (2013). Phrase detectives : Utilizing collective intelligence for internet-scale language

resource creation. *ACM Transactions on Interactive Intelligent Systems*, **3**(1), 1–44. DOI : 10.1145/2448116.2448119.

POTTIER B. (1963). *Recherches sur l'analyse sémantique en linguistique et en traduction mécanique*. Publications linguistiques de la Faculté des lettres et sciences humaines de Nancy.

PROUTEAU T., CONNES V., DUGUÉ N., PEREZ A., LAMIREL J.-C., CAMELIN N. & MEIGNIER S. (2021). SINr : Fast Computing of Sparse Interpretable Node Representations is not a Sin ! In *Advances in Intelligent Data Analysis XIX, 19th International Symposium on Intelligent Data Analysis, IDA 2021*, Lecture Notes in Computer Science, p. 325–337, Porto, Portugal : Springer, Cham. Issue : 12695, DOI : 10.1007/978-3-030-74251-5_26.

PROUTEAU T., DUGUÉ N., CAMELIN N. & MEIGNIER S. (2022). Are Embedding Spaces Interpretable ? Results of an Intrusion Detection Evaluation on a Large French Corpus. p.6̃.

RASTIER F. (2009). Principes et conditions de la sémantique componentielle. In *Sémantique interprétative*, Formes sémiotiques, p. 17–37. Presses Universitaires de France.

ROTARU A. S., VIGLIOCCO G. & FRANK S. L. (2018). Modeling the Structure and Dynamics of Semantic Processing. *Cognitive Science*, **42**(8), 2890–2917. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12690, DOI : 10.1111/cogs.12690.

SUBRAMANIAN A., PRUTHI D., JHAMTANI H., BERG-KIRKPATRICK T. & HOVY E. (2017). SPINE : SParse Interpretable Neural Embeddings. arXiv :1711.08792 [cs], DOI : 10.48550/arXiv.1711.08792.

VON AHN L. (2006). Games with a purpose. *Computer*, **39**(6), 92–94. DOI : 10.1109/MC.2006.196.

ŞENEL L. K., UTLU , YÜCESOY V., KOÇ A. & ÇUKUR T. (2018). Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(10), 1769–1779. DOI : 10.1109/TASLP.2018.2837384.