Introduction au numéro spécial sur l'explicabilité des modèles de TAL

Guillaume Wisniewski*

* *LLF, CNRS, Université Paris-Cité, F-75013, Paris, France*

RÉSUMÉ. *La capacité des modèles neuronaux à construire, sans supervision explicite, des représentations de la langue a contribué aux progrès spectaculaires réalisés au cours de la dernière décennie par les systèmes de traitement de la langue et de la parole. Si ces représentations permettent de développer des systèmes pour de nombreuses langues et de nombreux domaines, leur utilisation se fait au détriment de l'interprétabilité des décisions : il n'est généralement pas possible de savoir pourquoi un système prend telle ou telle décision. Arriver à comprendre les informations encodées dans ces représentations et à expliquer les prédictions de ces systèmes est aujourd'hui une problématique importante, aussi bien d'un point de vue scientifique qu'applicatif, qui suscite de très nombreux travaux. Nous présentons ici les enjeux de cette thématique et résumons les cinq articles du numéro spécial de la revue TAL sur l'explicabilité qui donnent un aperçu des enjeux de cette problématique et illustrent les différentes méthodes d'explicabilité explorées par la communauté.*

MOTS-CLÉS : *explicabilité, analyse des représentations neuronales, évaluation diagnostique.*

TITLE. *Introduction to Special Issue on Explicability of NLP Models*

ABSTRACT. *The ability of neural networks to construct representations of language without explicit supervision has contributed to the spectacular progress in language and speech processing systems over the last decade. While these representations make it possible to develop systems for many languages and domains, their use comes at the expense of their interpretability: it is generally not possible to know why a system makes a particular decision. Understanding the information encoded in these representations and explaining the predictions made by these systems is an important issue today, both from a scientific and an application point of view, and is the subject of much work. In this article, we present the issues at stake and summarize the five articles in the TAL special issue on Explicability, which provide an overview of the issues at stake and illustrate the different methods being explored by the community.*

KEYWORDS: *Explainability, Neural Representation Analysis, Diagnostic Evaluation in NLP.*

1. Introduction

La capacité des modèles neuronaux à construire, sans supervision explicite, des représentations de la langue a contribué aux progrès spectaculaires (qu'on pourrait même qualifier de révolutionnaires) réalisés ces dernières années par les systèmes de traitement de la langue et de la parole. Ces représentations permettent, en effet, de développer des systèmes pour de nombreuses tâches, notamment en affinant des modèles préentraînés tel BERT. Elles sont également au cœur des performances surprenantes des gigamodèles comme les fameux modèles GPT au cœur de l'IA générative et qui ont donné naissance à une nouvelle manière de concevoir les systèmes de TAL à l'aide d'amorces *(prompts)*.

Si ces représentations neuronales sont aujourd'hui le fondement de tous les systèmes de TAL, leur utilisation se fait au détriment de l'interprétabilité : à cause du nombre de paramètres mis en jeu et du caractère non supervisé de l'apprentissage des représentations, il n'est généralement pas possible de savoir pourquoi un système prend telle ou telle décision ni même quelles informations il considère. Les raisons derrière les bonnes performances des modèles de l'état de l'art restent, en grande partie, inconnues. Les approches à base de réseaux de neurones sont, en cela, fondamentalement différentes des approches longtemps au cœur du TAL, qui construisaient et manipulaient une représentation explicite de la structure « abstraite » des phrases (par exemple un arbre syntaxique ou une représentation sémantique de type logique). C'est pourquoi, ils sont généralement qualifiés de *boîtes noires* ou *opaques*.

Cette opacité des représentations et des systèmes et la nécessité de comprendre et de justifier les décisions des systèmes reposant sur l'apprentissage statistique et, en particulier, sur les réseaux de neurones, a donné naissance à un domaine de recherche très riche et dynamique qui dépasse le domaine du TAL : l'intelligence artificielle explicable (XAI pour *eXplainable Artificial Intelligence*). Si elle soulève des défis spécifiques liés à la nature même des données manipulées (notamment le fait que les mots sont des symboles discrets), l'explicabilité des systèmes de TAL s'inscrit pleinement dans ce domaine : les méthodes développées et utilisées ainsi que les problèmes rencontrés rejoignent en tout point les questions au cœur de l'XAI.

L'objectif de ce numéro spécial de la revue TAL est de proposer, via les cinq articles retenus par le comité scientifique, un aperçu des recherches sur l'explicabilité. Après avoir rappelé les principaux enjeux de l'explicabilité à la section 2 et décrit rapidement les différentes catégories de méthodes proposées dans la littérature (section 3), nous résumons à la section 4, ces contributions qui illustrent la variété des méthodes et des questions au cœur de ce domaine.

2. Une problématique aux enjeux multiples

La recherche d'explications pour les prédictions des systèmes reposant sur des méthodes d'apprentissage statistique, ainsi que plus généralement des systèmes relevant de l'IA, est une problématique ancienne qui s'est développée depuis que ces méthodes

ont commencé à être utilisées dans des applications (voir Barredo Arrieta *et al.* (2020) pour un aperçu général et un historique de cette problématique). Elle répond, en effet, à de nombreux besoins et questions rendus encore plus prégnants par le développement des systèmes opaques, et notamment ceux fondés sur des réseaux de neurones.

Le premier de ces enjeux est un enjeu applicatif, voire social : les systèmes de TAL sont utilisés quotidiennement par un nombre croissant de personnes et, comme souligné par Goodman et Flaxman (2017), leurs prédictions sont souvent jugées d'une qualité suffisante pour être utilisées sans aucune intervention humaine. Dans la mesure où ces prédictions ont un impact sur la vie de leurs utilisateurs et utilisatrices, il est nécessaire de garantir qu'elles ne leur portent pas préjudice. Expliquer les décisions de ces systèmes est une étape essentielle pour limiter, voire empêcher les erreurs, les discriminations et les injustices causées par ceux-ci et pour développer une *IA de confiance (Trustworthy AI)*.

La nécessité d'expliquer les décisions des systèmes de TAL s'inscrit désormais dans un cadre légal : les discussions autour de la régulation de l'IA (qui concerne directement le TAL), que ce soit en Union européenne avec l'*IA Act*, aux États-Unis avec l'*AI Bill of Rights*, ou en Angleterre avec la *National AI Strategy*, appellent les concepteurs et conceptrices à assurer la transparence et l'explicabilité de leurs systèmes d'IA (Gyevnar *et al.*, 2023). À cet égard, les articles 22(3) et 13-15 du Réglement général sur la protection des données (RGPD) peuvent déjà, selon Goodman et Flaxman (2017), être interprétés comme un « droit à l'explication » pour les personnes ayant fait l'objet d'une « décision automatisée ».

Expliquer les décisions des systèmes de TAL est également d'une importance capitale pour les concepteurs et conceptrices de ces systèmes : les explications produites peuvent en effet fournir des indications sur les limites des systèmes développés et sur les causes des erreurs qu'ils commettent. Elles offrent ainsi des indications précieuses pour la mise au point, l'amélioration et le débogage des systèmes de TAL. Lertvittayakumjorn et Toni (2021) présentent un état de l'art complet de l'utilisation des méthodes d'explication dans ce contexte.

Ces deux enjeux ne doivent pas nous faire oublier la problématique « scientifique » de l'explicabilité : comprendre comment les réseaux de neurones représentent et manipulent le langage, mais également comment ils acquièrent leurs connaissances et leurs compétences, est un véritable défi qui rejoint les objectifs les plus fondamentaux de la science (« comprendre et expliquer le monde » selon la définition donnée par la Wikipédia francophone (Wikipédia, 2024)). En plus d'éclairer notre compréhension des systèmes présents dans notre quotidien, répondre à ces interrogations peut aussi apporter de nouvelles perspectives dans divers domaines scientifiques, tels que les sciences cognitives (Dupoux, 2018), la psychologie (Zhuang *et al.*, 2022), les sciences politiques (Cao et Kosinski, 2024) ou, naturellement, la linguistique (Kirov et Cotterell, 2018 ; Pater, 2019).

3. Des objectifs et des méthodes variés

La question de l'explicabilité a généré un très grand nombre de travaux, notamment en TAL. S'il est illusoire de dresser une liste exhaustive de ceux-ci, les états de l'art sur cette question (par exemple Guidotti *et al.* (2018)) distinguent en général deux types de travaux : ceux proposant des modèles explicables de manière inhérente et ceux cherchant à « ouvrir » la boîte noire (selon l'expression consacrée) en développant des méthodes pour expliquer les comportements ou les décisions de systèmes existants. L'ensemble des articles de ce numéro s'inscrit dans cette deuxième catégorie et offre un aperçu complet des différents types de méthodes qui ont été utilisées pour analyser et comprendre les modèles : l'explicabilité d'un modèle via l'apprentissage d'un modèle de substitution interprétable, l'explicabilité locale à l'échelle d'une décision, l'explicabilité par inspection des paramètres du modèle.

Cette typologie (à très gros grain) ne doit pas faire perdre de vue que les travaux publiés ont généralement deux objectifs distincts. Un premier type de travaux (Jawahar *et al.*, 2019 ; Li *et al.*, 2023a) vise principalement à expliciter les connaissances linguistiques capturées par les modèles en établissant un lien entre les représentations neuronales et les représentations « classiques » utilisées en TAL (partie du discours, arbres syntaxiques, etc.), et plus généralement dans la modélisation « linguistique » des énoncés, qu'ils soient écrits ou parlés. Deux des articles de ce numéro constituent des exemples représentatifs de ce type de travaux : *Détection de la nasalité en parole à partir de* `wav2vec 2.0` et *Context-Aware Neural Machine Translation Models Analysis and Evaluation Through Attention*.

Le second type de travaux (Li *et al.*, 2023b ; Stahlberg *et al.*, 2018), s'inscrit dans un cadre plus applicatif et se concentre sur l'explication des prédictions aux utilisateurs et utilisatrices finaux plutôt qu'aux concepteurs et conceptrices des systèmes. Ces recherches mettent l'accent sur la transparence et sur l'interprétabilité des modèles développés, visant à rendre leurs décisions compréhensibles et justifiables pour les utilisateurs et utilisatrices non experts. Trois des articles de ce numéro spécial illustrent cette catégorie de travaux : *Sensibilité des explications à l'aléa des grands modèles de langage : le cas de la classification de textes journalistiques*, *Expliquer une boîte noire sans boîte noire* et *La recherche sur les biais dans les modèles de langue est biaisée : état de l'art en abyme*.

4. Contenu du numéro spécial

Ce numéro spécial de la revue TAL contient cinq articles qui illustrent la variété des travaux conduits autour de la problématique de l'explicabilité et de l'analyse des représentations neuronales. Ces articles illustrent parfaitement les différentes méthodes utilisées dans la littérature et la diversité des questions abordées.

Dans le premier article de ce numéro spécial, *Sensibilité des explications à l'aléa des grands modèles de langage : le cas de la classification de textes journalistiques*, Jérémie Bogaert et ses coauteurs abordent une tâche de classification dont l'objectif

est de distinguer les articles de journaux exprimant une opinion de ceux relatant une information. Il s'appuie sur une r méthodde d'analyse, la méthode LRP *(Layer-Wise Relevance Propagation)* de Bach *et al.* (2015), qui construit des *cartes d'importance* identifiant les mots jouant un rôle prépondérant lors de la prise de décision.

Les auteurs observent toutefois, en comparant des modèles obtenus à partir de différentes initialisations aléatoires, que les mots identifiés comme importants par cette méthode varient fortement d'un modèle à l'autre, même si leurs performances sont similaires, soulevant la difficulté d'interpréter les résultats de la méthode LPR. L'article aborde également la question de l'évaluation des méthodes d'explication, notamment en termes de *fidélité* et de *plausibilité*, en comparant les explications obtenues par cette méthode à celles d'une méthode « classique » (un classifieur linéaire considérant des caractéristiques définies par un expert ou une experte).

Dans *Détection de la nasalité en parole à partir de wav2vec 2.0*, Lila Kim et Cédric Gendrot étudient la capacité du modèle préentraîné multilingue de la parole wav2vec2 (Baevski *et al.*, 2020), reposant sur une architecture *Transformer*, à capturer des informations sur la manière dont certains sons sont produits. Leurs expériences reposent sur un second type de méthode d'analyse, les sondes linguistiques *(linguistic probes)* (Köhn, 2015 ; Gupta *et al.*, 2015) ou classifieur de diagnostic *(diagnostic classifier)* (Hupkes et Zuidema, 2018) : les deux auteurs rapportent qu'il est possible d'entraîner un classifieur pour prédire la nasalité à partir des représentations vectorielles construites par wav2vec2, démontrant ainsi que cette information est encodée dans les représentations générées par le réseau de neurones.

En mettant en parallèle les prédictions de la sonde avec des mesures physiologiques, les deux auteurs sont également capables d'expliquer les erreurs de celle-ci et d'affiner notre compréhension des informations capturées par le modèle wav2vec2. En plus d'améliorer nos connaissances sur les représentations neuronales de la parole et sur les informations qu'elles encodent, ce travail suggère également que ces représentations peuvent avantageusement remplacer des mesures aérodynamiques difficiles à réaliser, illustrant ainsi la complémentarité entre études linguistiques et analyse des représentations neuronales.

Le troisième article de ce numéro repose sur un autre type d'analyse qui se concentre, cette fois, sur l'attention, une des composantes centrales des *Transformers*. Dans *Context-Aware Neural Machine Translation Models Analysis and Evaluation Through Attention*, Marco Dinarelli et ses coauteurs s'intéressent à la traduction en contexte, une tâche pour laquelle la résolution de l'ambigüité des phénomènes discursifs nécessite de prendre en compte des informations du contexte, parfois au-delà des frontières de phrases.

En introduisant une méthode de normalisation des auto-attentions entre les mots des phrases sources, qui facilite leur visualisation, et en considérant un phénomène linguistique particulier (la coréférence), les auteurs observent à l'aide d'une analyse manuelle qualitative et en définissant plusieurs métriques que l'attention peut être utilisée pour analyser, voire pour expliquer le comportement de différents modèles de

traduction en contexte et même pour évaluer leur capacité à capturer correctement les informations du contexte.

Ces trois articles s'inscrivent dans un même courant de recherche visant à *identifier* les informations qui sont capturées dans les représentations neuronales, sans toutefois pouvoir déterminer si ces informations sont bien *utilisées* par le système pour réaliser ses prédictions. Cette limite des méthodes d'analyse peut se comprendre comme la (fameuse) différence entre corrélation et causalité. Identifiée depuis longtemps (Belinkov, 2022 ; Vanmassenhove *et al.*, 2017), elle a donné lieu à un second paradigme pour l'explicabilité visant à *intervenir* sur les représentations et à découvrir les effets causaux résultants de ces interventions : l'analyse contrefactuelle.

Le quatrième article de ce numéro spécial, *Expliquer une boîte noire sans boîte noire*, s'inscrit dans ce paradigme : Julien Delaunay et ses coauteurs comparent plusieurs méthodes d'explications contrefactuelles (y compris une méthode qu'ils ont développée) pour des tâches de classification de documents. Ces méthodes permettent de déterminer les modifications à apporter à un document pour changer la prédiction d'un classifieur. Les mots ainsi identifiés expliquent (ou du moins justifient) la prédiction initiale. En plus de présenter différentes méthodes pour générer des explications contrefactuelles, les auteurs introduisent un cadre général permettant d'évaluer la qualité de celles-ci et introduisent les métriques afférentes : les expériences décrites dans ce travail soulignent les différents compromis (notamment entre l'interprétabilité d'un modèle et ses performances) qu'il est nécessaire de considérer dans le développement de méthodes d'explicabilité.

Enfin, le dernier article intitulé *La recherche sur les biais dans les modèles de langue est biaisée : état de l'art en abyme*, aborde une question rendue plus prégnante par l'utilisation croissante des systèmes de TAL par le grand public, à savoir les biais, par exemple de genre, qui parsèment les textes générés par ces systèmes et qui nuisent aux minorités et aux groupes historiquement désavantagés. Fanny Ducel et ses coauteures dressent un état de l'art des recherches sur cette question et abordent trois types de travaux complémentaires : les méthodes permettant d'identifier les biais stéréotypés, les méthodes atténuant ces biais et, enfin, les méthodes d'évaluation des biais.

Au-delà de cet état de l'art, l'article propose une analyse des travaux selon divers critères, mettant en évidence certains biais présents dans la recherche sur les biais. Si ceux-ci illustrent bien les préjugés inhérents à la recherche, les auteures formulent plusieurs propositions dans leur conclusion pour améliorer les travaux dans ce domaine, avec pour objectif principal de limiter, notamment auprès du grand public, le biais cognitif selon lequel les machines seraient objectives.

Remerciements

Nous remercions le comité éditorial et scientifique de la revue TAL, ainsi que le comité scientifique invité, en particulier les relecteurs et relectrices, qui ont contribué par leur temps et leurs efforts à la qualité de ce numéro : Loïc Barrault (Meta),

Rachel Bawden (INRIA), Delphine Bernhard (LiLPa, Université de Strasbourg), Nathalie Camelin (LIUM, Le Mans Université), Lina Conti (University of Trento), Maxime Fily (LLF, Université Paris Cité), Aina Garí Soler (Télécom Paris), Nabil Hathout (CLLE, CNRS), Joseph Le Roux (LIPN, Université Sorbonne Paris Nord), Fabrice Maurel (Greyc, Université de Caen Normandie), Timothee Mickus (University of Helsinki), Philippe Muller (MELODI, Université Paul Sabatier), Lucas Ondel Yang (LISN, Université Paris-Saclay), Xavier Tannier (LIMICS, Sorbonne Université) et Nadi Tomeh (LIPN, Université Sorbonne Paris Nord).

5. Bibliographie

Bach S., Binder A., Montavon G., Klauschen F., Müller K.-R., Samek W., « On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation », *PLOS ONE*, vol. 10, n° 7, p. 1-46, 07, 2015.

Baevski A., Zhou H., Mohamed A., Auli M., « wav2vec 2.0 : a framework for self-supervised learning of speech representations », *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F., « Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI », *Information Fusion*, vol. 58, p. 82-115, 2020.

Belinkov Y., « Probing Classifiers : Promises, Shortcomings, and Advances », *Computational Linguistics*, vol. 48, n° 1, p. 207-219, 04, 2022.

Cao X., Kosinski M., « Large language models know how the personality of public figures is perceived by the general public », *Scientific Reports*, vol. 14, n° 1, p. 6735, Mar, 2024.

Dupoux E., « Cognitive science in the era of artificial intelligence : A roadmap for reverse-engineering the infant language-learner », *Cognition*, vol. 173, p. 43-59, 2018.

Goodman B., Flaxman S., « European Union Regulations on Algorithmic Decision Making and a "Right to Explanation" », *AI Magazine*, vol. 38, n° 3, p. 50-57, 2017.

Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D., « A Survey of Methods for Explaining Black Box Models », *ACM Comput. Surv.*, aug, 2018.

Gupta A., Boleda G., Baroni M., Padó S., « Distributional vectors encode referential attributes », *in* L. Màrquez, C. Callison-Burch, J. Su (eds), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, p. 12-21, September, 2015.

Gyevnar B., Ferguson N., Schafer B., « Bridging the Transparency Gap : What Can Explainable AI Learn from the AI Act ? », *Proceedings of ECAI*, p. 964-971, 2023.

Hupkes D., Zuidema W., « Visualisation and 'Diagnostic Classifiers' Reveal how Recurrent and Recursive Neural Networks Process Hierarchical Structure (Extended Abstract) », *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, International Joint Conferences on Artificial Intelligence Organization, p. 5617-5621, 7, 2018.

Jawahar G., Sagot B., Seddah D., « What Does BERT Learn about the Structure of Language ? », *in* A. Korhonen, D. Traum, L. Màrquez (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 3651-3657, July, 2019.

Kirov C., Cotterell R., « Recurrent Neural Networks in Linguistic Theory : Revisiting Pinker and Prince (1988) and the Past Tense Debate », *Transactions of the Association for Computational Linguistics*, vol. 6, p. 651-665, 12, 2018.

Köhn A., « What's in an Embedding ? Analyzing Word Embeddings through Multilingual Evaluation », *in* L. Màrquez, C. Callison-Burch, J. Su (eds), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, p. 2067-2073, September, 2015.

Lertvittayakumjorn P., Toni F., « Explanation-Based Human Debugging of NLP Models : A Survey », *Transactions of the Association for Computational Linguistics*, vol. 9, p. 1508-1528, 2021.

Li B., Wisniewski G., Crabbé B., « Assessing the Capacity of Transformer to Abstract Syntactic Representations : A Contrastive Analysis Based on Long-distance Agreement », *Transactions of the Association for Computational Linguistics*, vol. 11, p. 18-33, 2023a.

Li D., Hu B., Chen Q., He S., « Towards Faithful Explanations for Text Classification with Robustness Improvement and Explanation Guided Training », *in* A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, R. Gupta (eds), *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Association for Computational Linguistics, Toronto, Canada, p. 1-14, July, 2023b.

Pater J., « Generative linguistics and neural networks at 60 : Foundation, friction, and fusion. », *Language*, vol. 95, n° 1, p. e41-e74, 2019.

Stahlberg F., Saunders D., Byrne B., « An Operation Sequence Model for Explainable Neural Machine Translation », *in* T. Linzen, G. Chrupała, A. Alishahi (eds), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Brussels, Belgium, p. 175-186, November, 2018.

Vanmassenhove E., Du J., Way A., « Investigating'Aspect'in NMT and SMT : translating the English simple past and present perfect », *Computational Linguistics in the Netherlands Journal (CLIN)*, vol. 7, p. 109-128, 2017.

Wikipédia, « Science — Wikipédia, l'encyclopédie libre », , En ligne (Page disponible le 8 mai 2024), 2024.

Zhuang C., Xiang Z., Bai Y., Jia X., Turk-Browne N., Norman K., DiCarlo J. J., Yamins D., « How Well Do Unsupervised Learning Algorithms Model Human Real-time and Lifelong Learning ? », *in* S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (eds), *Advances in Neural Information Processing Systems*, vol. 35, Curran Associates, Inc., p. 22628-22642, 2022.