Une Librairie pour Évaluer l'Interprétabilité des Modèles de Langue

Khalil Maachou[1], Jesús Lovón-Melgarejo[1], Jose G. Moreno[1] and Lynda Tamine[1]

[1]*Université Paul Sabatier, IRIT, UMR 5505 CNRS, Toulouse, France*

Abstract

Les modèles les plus récents basés sur les Transformers, dans les domaines du traitement automatique des langues (TAL) et de la recherche d'information (RI), sont réputés pour leur opacité dans le processus décisionnel. Pour pallier cette limitation, des techniques d'interprétabilité ont émergé, visant à rendre les modèles plus transparents. Malgré la disponibilité de nombreuses ressources pour ces techniques, leur intégration reste complexe. Ce travail propose une librairie intégrée pour évaluer l'interprétabilité des modèles, facilitant ainsi des évaluations rapides et robustes. Cet article est une version résumée de K. Maachou et al. (2024), "eval-rationales: An End-to-End Toolkit to Explain and Evaluate Transformers-Based Models", accepté comme article de démonstration à ECIR 2024.

Keywords

interpretable, évaluation, modèles de langue

1. Introduction

L'évolution rapide des modèles neuronaux dans le traitement automatique des langues (TAL) et la recherche d'information (RI) a conduit à des performances historiques dans plusieurs tâches, mais leur opacité est devenue un défi majeur [1]. Par conséquent, l'intérêt pour l'intelligence artificielle explicable (XAI) a augmenté, en particulier sur l'amélioration des approches pour rendre ces modèles plus interprétables [2, 3].

Dans le contexte des modèles basés sur les Transformers, différents efforts ont été déployés pour expliquer les prédictions à différents niveaux de granularité : locales et globales [4, 5]. Les explications *locales* visent à clarifier les raisons derrière une prédiction individuelle, tandis que les explications *globales* s'intéressent aux prédictions pour un ensemble d'entrées.

Nous nous concentrons sur l'utilisation de *rationales locales* pour expliquer les prédictions, ces *rationales* étant des sous-ensembles d'éléments d'entrée qui expliquent une prédiction [6]. Les techniques XAI basées sur les *rationales* visent à élucider le "pourquoi" derrière les résultats d'un modèle. Cependant, l'évaluation de la qualité des explications reste un défi, d'où l'émergence de *benchmarks* de référence comme ERASER [7].

Des outils et des bibliothèques ont été développés pour faciliter l'utilisation de ces techniques XAI, mais leur intégration reste difficile. Dans ce contexte, nous présentons "eval-rationales", une librairie qui intègre des techniques XAI locales avec le *benchmark* ERASER, ainsi que des

Conférence en Recherche d'Information et Applications (CORIA) 2024, 3 – 4 avril, La Rochelle, France

khalil.maachou@univ-tlse3.fr (K. Maachou); jesus.lovon@irit.fr (J. Lovón-Melgarejo); jose.moreno@irit.fr (J. G. Moreno); tamine@irit.fr (L. Tamine)

0009-0000-9410-5935 (K. Maachou); 0000-0001-6243-0864 (J. Lovón-Melgarejo); 0000-0002-8852-5797 (J. G. Moreno); 0000-0002-3615-8032 (L. Tamine)

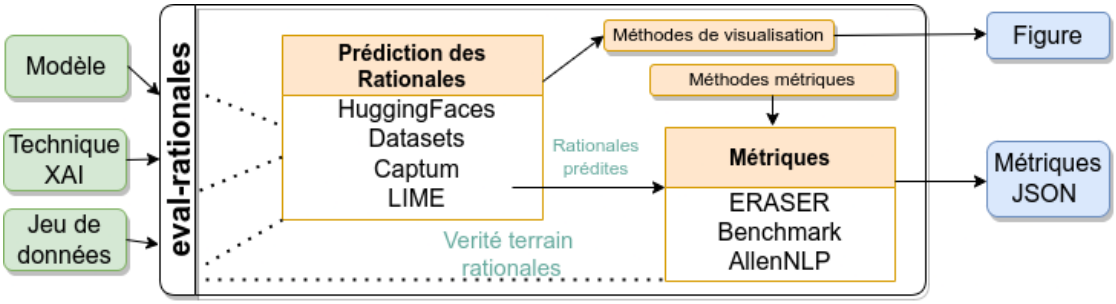

Figure 1: La librairie "eval-rationales" se compose de deux modules : les prédictions des *rationales* et les métriques. Le résultat est un fichier json contenant les métriques calculées.

modèles et des ensembles de données du hub HuggingFace. Notre librairie cherche à simplifier l'évaluation et l'exploration des modèles, offrant des métriques détaillées pour évaluer la qualité des prédictions.

2. Description de la librairie

L'outil présenté offre un ensemble de métriques basées sur les *rationales* pour évaluer la qualité des explications générées par les modèles basés sur les Transformers sur des ensembles de données d'évaluation. Il est composé de deux modules interconnectés : i) le module de prédiction des *rationales*, chargé de prédire les *rationales* à partir des entrées du modèle, et ii) le module de métriques, qui calcule les métriques basées sur les *rationales* prédites (Figure 1).

Le module de prédiction des *rationales* de notre outil prend en entrée trois éléments : un modèle, un ensemble de données d'évaluation et une technique XAI. Notre contribution est adaptée pour évaluer les modèles orientés vers des tâches de classification binaire de séquences. Pour favoriser l'intégration avec les derniers développements en TAL et en RI, nous avons adapté notre outil pour prendre en charge les modèles basés sur la bibliothèque HuggingFace. Aussi, nous considérons les techniques XAI basées sur l'attention [8, 9], LIME [10] et des approches basées sur les gradients [11], principalement basées sur la bibliothèque Captum. De plus, nous avons inclus une méthode aléatoire, servant de référence [7].

Le module de métriques utilise des métriques basées sur les *rationales* suivant le *benchmark* ERASER. Nous calculons deux métriques principales : la complétude et la suffisance. La complétude évalue si les *rationales* prédites incluent toutes les caractéristiques nécessaires pour faire une prédiction. La suffisance évalue si les *rationales* extraites contiennent suffisamment de signal pour prendre une décision éclairée. De plus, lorsque l'ensemble des données d'évaluation inclut la vérité terrain des *rationales*, nous calculons des métriques complémentaires comme la précision, le rappel et le score F1.

3. Conclusion

Cet article présente une librairie pour évaluer les modèles interprétables, intégrant les modèles HuggingFace et le benchmark ERASER. Cette nouvelle ressource atténue les défis de l'intégration de sources différentes, et simplifie le processus d'évaluation des modèles.

References

[1] G. Ras, N. Xie, M. Van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, Journal of Artificial Intelligence Research 73 (2022) 329–396.

[2] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable ai (xai): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 1–33.

[3] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable ai for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 447–459.

[4] L. Lyu, A. Anand, Listwise explanations for ranking models using multiple explainers, in: European Conference on Information Retrieval, Springer, 2023, pp. 653–668.

[5] B. Bhattarai, O.-C. Granmo, L. Jiao, An interpretable knowledge representation framework for natural language processing with cross-domain application, in: European Conference on Information Retrieval, Springer, 2023, pp. 167–181.

[6] S. Wiegreffe, A. Marasovic, Teach me to explain: A review of datasets for explainable natural language processing, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021, p. .

[7] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4443–4458. doi:10.18653/v1/2020.acl-main.408.

[8] D. Bahdanau, K. H. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: 3rd International Conference on Learning Representations, ICLR 2015, 2015, p. .

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[10] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.

[11] S. Karlekar, T. Niu, M. Bansal, Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 701–707.