Proposition de sujet de thèse

Mars 2022

Encadrant de thèse Trois intervenants accompagneront ce projet de recherche.

- Directrice de thèse : **Marie-Jeanne Lesot** – Maître de conférence HDR en informatique au LIP6 / Sorbonne Université
- Co-encadrant : **Jean-Noël Vittaut** – Maître de conférence en informatique au LIP6 / Sorbonne Université
- Co-encadrant au sein de l'entreprise : **Nicolas Chesneau** – Chercheur en *deep learning* et *Tech Lead machine learning* à Ekimetrics

1 Contexte

Par leur développement rapide et leur efficacité en constante amélioration, les algorithmes d'apprentissage automatique (*machine learning*) voient leur importance accrue dans notre société. De l'octroi de crédit immobilier à la tarification d'assurance, l'intelligence artificielle est aujourd'hui au coeur des décisions des experts métier. En particulier, l'apprentissage profond (*deep learning*) y a fortement contribué via l'implémentation de nouveaux modèles de plus en plus complexes. Cette complexité a induit une difficulté croissante à comprendre et interpréter les prédictions effectuées.

Le domaine de l'*eXplainable Artificial Intelligence* (XAI) vise à développer des méthodes dites d' "interprétabilité" ou "explicabilité" afin de pallier ce défaut de transparence. Une partie d'entre elles propose d'appréhender les modèles décrits ci-dessus comme des boîtes noires et d'expliquer leurs prédictions *a posteriori*, sans se référer aux paramètres qui leur sont inhérents : ces approches sont dites *post-hoc*. Ces techniques ont divers champs d'application, permettent notamment d'atteindre un compromis entre performance et explicabilité, et peuvent de même faciliter la détection de divers biais comme celui d'échantillonnage [1].

Une autre catégorie de méthodes d'interprétabilité propose d'expliquer la prédiction d'un algorithme par la génération d'exemples contrefactuels [2]. Un raisonnement contrefactuel consiste à modifier hypothétiquement l'issue d'un évènement en modifiant l'une de ses causes. Générer un exemple contrefactuel revient alors à évaluer dans quelle mesure la modification d'une donnée d'entrée peut impacter la prédiction d'un algorithme. Un exemple contrefactuel doit être "proche" de l'exemple initial afin d'évaluer au mieux la perturbation impactant l'évènement observé .

Par ailleurs, le traitement automatique du langage (TAL) ou *natural language processing* (NLP) - la branche de l'apprentissage automatique traitant spécifiquement les problématiques de modélisation du langage - a été particulièrement influencé par les récentes avancées du *deep learning*. En effet, les données textuelles étant aujourd'hui de plus en plus accessibles notamment grâce aux plateformes de e-commerces (Amazon, Alibaba...) et aux réseaux sociaux (Twitter, Instagram...), leurs cas d'usage associés se sont multipliés dans l'industrie. Ainsi, les algorithmes à l'état de l'art de type *Transformer* permettent d'effectuer avec précision diverses tâches complexes comme l'analyse de sentiment, la traduction, la classification de texte ou bien encore le résumé [3]. Ces modèles peuvent compter jusqu'à plusieurs milliards de paramètres et sont pré-entrainés sur de très larges corpus de textes pouvant contenir divers types de biais (ethnique, genre...). Le risque que ces biais soient répliqués par les modèles existe [4], ce qui renforce la nécessité de tirer profit des méthodes d'interprétabilité afin de les détecter et les corriger.

2 Etat de l'art

Si les algorithmes *post-hoc* de type valeurs de Shapley [5] ou LIME [1] sont certainement les plus connus, les travaux se sont multipliés ces dernières années et d'autres approches ont été proposées. La méthode LORE propose par exemple d'expliquer localement une prédiction à l'aide d'un arbre de décision [6]. Les méthodes *post-hoc* évaluant les algorithmes de classification ou de régression comme des boîtes noires, indépendamment des paramètres qui les constituent, celles-ci peuvent être utilisées dans le cadre du NLP. Leur complexité algorithmique [5] [7] rend cependant leur application aux modèles de *deep learning* parfois difficile. dans le cadre du NLP.

Basées notamment sur les algorithmes d'interprétabilité *post-hoc*, il existe plusieurs méthodes de génération d'exemples contrefactuels fonctionnant sur des données tabulaires [8]. D'autres méthodes de génération de contrefactuels ont été développées afin d'imposer des contraintes d'actionnabilité et/ou de faisabilité [9][10], ceci rendant les explications plus pertinentes et réalistes. A notre connaissance, celles-ci n'ont pas été testées sur des données textuelles ; la génération de contrefactuels restant assez largement circonscrite aux données tabulaires en raison de la grande dimension des espaces latents en NLP [11].

Les modèles de *deep learning* les plus récemment appliqués au NLP présentent de manière inhérente à leur structure des coefficients d'attention. Ceux-ci constituent une piste afin d'interpréter les prédictions effectuées par ces modèles. Divers travaux ont été réalisés afin de remettre en question le fait que les coefficients d'attention puissent être source d'interprétabilité en NLP [12][13], mais ils ne se focalisent que sur des modèles de type LSTM (*long short term memory*). Une analyse plus récente va cependant dans le sens du caractère assimilable des coefficients d'auto-attention aux valeurs de Shapley dans le cadre des BERT [14]. Si la construction de méthodes de génération d'exemples contrefactuels basées sur les coefficients d'attention semble prometteuse grâce à sa complexité algorithmique plus avantageuse, la littérature scientifique n'explore pas cette piste pour le moment.

Les contraintes d'actionnabilité et de faisabilité s'appliquant au cadre classique des données tabulaires ont leur équivalent en NLP : la plausibilité. Deux approches ont été récemment proposées afin de générer des exemples contrefactuels textuels plausibles grammaticalement [15], en intégrant de même des contraintes de diversité (représentativité de l'espace latent initial), d'efficacité, et d'orientation vers un but (déviation de l'état initial vers un aspect particulier, comme un sentiment) [16]. Ces techniques tirent profit du modèle de fondation (*foundation model*) GPT-2 afin de générer du texte sous contrainte. Le nombre particulièrement élevé de paramètres constituant ces types de modèles rend difficile la génération de contrefactuels de longue taille. De plus, l'approche *Polyjuice* [15] nécessite l'intervention humaine via du *prompt engineering* afin de contrôler les contrefactuels générés, ce qui présente plusieurs défauts, comme le coût et l'impossibilité de réaliser un passage à l'échelle. Enfin, une approche s'inscrivant dans un cadre d'apprentissage par renforcement se propose de générer des contrefactuels en entraînant un générateur d'exemples antonymiques [17]. Cette approche a essentiellement pour but de rendre un classifieur plus robuste en augmentant son jeu de données d'entrainement et en évitant les associations fallacieuses (*spurious association*). Si les gains sont significatifs, la méthode ne garantit pas de contrôle sur la contrainte d'orientation vers un but autre que celui imposé par le générateur de texte ayant le rôle d' " agent" dans le paradigme d'apprentissage par renforcement (en l'occurrence un antonyme associé à un sentiment opposé). Ainsi, cette approche mériterait d'être adaptée à un cadre plus flexible dans lequel des exemples autres qu'antinomique pourraient être générés, en jouant par exemple sur les thématiques couvertes par le texte.

3 Les questions posées

Dans ce contexte, l'objectif de cette thèse est d'évaluer la possibilité de générer des contrefactuels dans le cadre du NLP sous diverses formes de contraintes comme celles de plausibilité, de justesse grammaticale ou d'orientation vers un but. Les générateurs de contrefactuels seront évalués comme source d'interprétabilité et comme méthode de renforcement de la robustesse des modèles de langage manipulés.

Ainsi, ce travail permettra de répondre aux questions suivantes :

- Dans quelle mesure les méthodes *post-hoc* agnostiques existantes sont-elles adaptées aux modèles de *deep learning* appliqués au NLP ?
- Comment interpréter les modèles de *deep learning* appliqués au NLP grâce aux paramètres propres à leur structure ? Peut-on en tirer une méthode de génération de contrefactuels ?
- De quelle manière peut-on intégrer les contraintes de plausibilité, d'efficacité et d'orientation vers un but en NLP à la génération de contrefactuels ?

Pour ce faire, les approches proposées seront testées sur divers jeux de données comme l'IMDB Database. Les modèles de langage à l'état de l'art de type BERT (*Bidirectionnal Encoder Representation from Transformers*) et autres dérivés d'architecture de *Transformers* seront mobilisés pour traiter ces questions. En particulier, les coefficients d'attention inhérents aux architectures de type *Transformers* seront l'objet d'investigations poussées. Enfin, l'utilisation d'algorithmes d'apprentissage par renforcement (*reinforcement learning*) sera envisagée lors du processus de création de texte nécessaire à la génération d'exemples contrefactuels. Des générateurs de textes autres que antonymiques seront testés afin d'améliorer la qualité des contrefactuels générés. Les méthodes de contrefactuels seront systématiquement testées et utilisées afin d'effectuer de la *data augmentation* et de la détection d'éventuels biais afin de rendre les modèles plus robustes.

4 Activité scientifique

4.1 Cartographie et évaluation des méthodes de contrefactuels issues de données tabulaires, adaptation au NLP

Les méthodes habituelles de génération de contrefactuels [10] issues du traitement de données tabulaires n'ont que peu été utilisées en NLP. Il s'agira de cartographier et évaluer leur pertinence sur de vrais jeux de données textuels dans le cadre de l'analyse de sentiment et de la classification de texte. Nous verrons dans quelle mesure ces approches sont susceptibles de remplir les conditions de plausibilité, d'efficacité et d'orientation vers un but. Nous adapterons alors ces techniques aux données textuelles et entamerons une réflexion autour de l'implémentation d'une nouvelle méthodologie.

4.2 Développement d'une approche de génération de contrefactuels basée sur l'analyse des paramètres inhérents aux modèles de *deep learning* appliqués au NLP

Les méthodes d'interprétabilité de type *post-hoc* étant à la base de nombreuses méthodes de génération de contrefactuels, nous nous proposons de tirer profit des mécanismes d'auto-attention afin de développer une nouvelle approche. Au préalable, un travail empirique sera réalisé afin d'évaluer quelles métriques synthétisant les informations contenues dans les matrices d'auto-attention permettent d'obtenir des mesures d'interprétabilité similaires à d'autres valeurs de référence comme celles de Shapley. Le grand nombre de matrices d'auto-attention (usuellement 12 par couche d'encodeur dans l'architecture) nous permettra d'effectuer de la modélisation de thématiques (*topic modeling*), ce qui facilitera le contrôle du mécanisme de génération d'exemple contrefactuels. D'autres méthodes de *topic modeling* comme par exemple l'allocation de Dirichlet latente (*Latent Dirichlet Allocation, LDA*) pourront être utilisées afin d'affiner notre proposition. En particulier, le problème sera évalué dans un cadre de classification de texte. Les mêmes contraintes que les précédentes approches seront évaluées. Nous tenterons ensuite de détecter les observations mal classées ou bien encore divers potentiels biais grâce à notre proposition de méthode.

4.3 Développement d'une approche de génération de contrefactuels intégrant les contraintes de plausibilité et d'orientation vers un but

Le développement d'heuristiques permettra de contourner la nécessité de l'intervention humaine lors de la manipulation de modèles de fondation (*foundation model*) tout en tirant profit de leur capacité à générer du texte sous contrainte via le *prompt engineering*. Ces heuristiques se baseront notamment sur l'information fournie par les coefficients d'auto-attention des modèles manipulés et sur des règles de grammaire. Enfin, l'inscription de notre approche dans un cadre d'apprentissage par renforcement dans lequel le générateur de texte serait l' "agent" et le modèle de langage l' "environnement" nous permettra de tester plusieurs générateurs de texte afin d'effectuer de la génération artificielle de données supplémentaires (*data augmentation*). Celle-ci nous permettra de rendre le modèle utilisé plus robuste. Enfin, une attention particulière sera portée au respect des contraintes de plausibilité sémantique et d'orientation vers un but. Leur intégration via une hybridation des approches précédentes et du cadre d'apprentissage par renforcement sera envisagée.

5 Conférences envisagées

Le travail effectué tout au long de la thèse sera soumis à des conférences afin de valider scientifiquement le travail réalisé. Leur choix dépendra du contenu des articles produits. Les conférences envisagées peuvent être regroupées en 3 catégories.

Dans un premier temps, les conférences internationales de *machine learning* classiques seront ciblées : *International Conference on Learning Representations* (ICLR), *Neural Information Processing Systems* (NeurIPS), *International Conference on Machine Learning* (ICML).

Nous envisageons également de participer aux conférences internationales traitant plus spécifiquement des problématiques relatives à l'*eXplainable Artificial Intelligence* (XAI), comme la conférence *Fairness, Accountability, and Transparency (FAT*)*, *ACM Conference on Human Factors in Computing Systems (CHI)* ou bien encore *Intelligent User Interfaces Conference (IUI)*.

Enfin, nous projetons de présenter nos travaux à des conférences nationales comme la *Conférence sur l'Apprentissage automatique* (CAP), ou *Extraction et Gestion des Connaissances* (EGC).

6 Calendrier prévisionnel

La première année de la thèse sera consacrée à la revue de littérature et la prise en main des méthodes usuelles de génération d'exemples contrefactuels. Il s'agira ensuite d'adapter ces approches au cadre du NLP et de rédiger un premier article scientifique présentant notre première proposition de méthode.

La deuxième année aura pour objectif d'investiguer l'apport des coefficients d'attention propres aux modèles de *deep learning* appliqués au NLP afin de construire une nouvelle méthode de génération de contrefactuels. Il s'agira d'intégrer divers types de contraintes au processus de génération, comme par exemple la plausibilité grammaticale ou bien encore l'orientation vers un but. Un deuxième article scientifique sera proposé.

Enfin, la troisième année se focalisera sur le développement d'une dernière approche s'inscrivant dans le cadre de l'apprentissage par renforcement. Un article scientifique sera de même rédigé. Les six derniers mois seront consacrés à la rédaction de la thèse et à la préparation de la soutenance.

7 Laboratoire, entreprise et organisation

7.1 Laboratoire d'accueil : le LIP6

Le LIP6 est un laboratoire de recherche en informatique placé sous la tutelle du CNRS et de l'université Sorbonne Université. Il regroupe plus de 500 personnes, dont plus de 200 permanents et 200 doctorants. Il se consacre à la modélisation et la résolution de problèmes fondamentaux motivés par les applications, ainsi qu'à la mise en œuvre et la validation des solutions au travers de partenariats académiques et industriels.

L'équipe LFI impliquée dans l'encadrement de ce sujet de thèse développe des recherches dans le cadre de l'intelligence computationnelle, elle étudie et propose de nouvelles approches pour la prise en compte et le traitement de données et de connaissances imparfaites et subjectives. Elle est spécialisée dans la gestion de telles données, en particulier pour les tâches de raisonnement et d'apprentissage automatique, en exploitant les cadres des logiques non classiques et du *soft computing*. Elle met également en œuvre ces approches dans le domaine de l'*eXplainable Artificial Intelligence* (XAI), pour améliorer l'interprétabilité et l'intelligibilité de systèmes d'apprentissage automatique.

7.2 Entreprise : Ekimetrics

Ekimetrics est un cabinet de conseil spécialisé en data science. L'entreprise travaille sur de nombreuses problématiques de *machine learning* classiques (classification, clustering, régression) mais aussi sur des sujets plus complexes comme le NLP, le *reinforcement learning*, ou la vision par ordinateur *computer vision*. Afin d'approfondir son expertise dans ces thématiques, l'entreprise lance l'Eki.Lab, un laboratoire qui a pour but de travailler sur des sujets de recherche ; le projet de thèse s'inscrit dans cette dynamique.

7.3 Organisation du travail

Le travail de recherche s'inscrit dans le cadre d'une thèse industrielle. La répartition du travail entre thèse et activité au sein de l'entreprise se fera dans les proportions 70-30%. Ces proportions seront de même respectées entre le temps passé au laboratoire et en entreprise.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.

[2] Judea Pearl. Causal and counterfactual inference. *The Handbook of Rationality*, pages 1–41, 2018.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[4] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184, 2021.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[8] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

[9] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[10] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[11] Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

[12] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[13] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[14] Kawin Ethayarajh and Dan Jurafsky. Attention flows are shapley value explanations. *arXiv preprint arXiv:2105.14652*, 2021.

[15] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*, 2021.

[16] Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. *arXiv preprint arXiv:2012.04698*, 2020.

[17] Hao Chen, Rui Xia, and Jianfei Yu. Reinforced counterfactual data augmentation for dual sentiment classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 269–278, 2021.