Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle

Duc Hau Nguyen, Guillaume Gravier, Pascale Sébillot

Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle

Duc Hau Nguyen[1] Guillaume Gravier[1] Pascale Sébillot[1]
(1) Univ Rennes, Inria, CNRS, IRISA, Rennes

RÉSUMÉ

Nous étudions la plausibilité d'un mécanisme d'attention pour une tâche d'inférence de phrases (*entailment*), c'est-à-dire sa capacité à fournir une explication plausible pour un humain de la relation entre deux phrases. En s'appuyant sur le corpus *Explanation-Augmented Standford Natural Language Inference*, il a été montré que les poids d'attention sont peu plausibles en pratique et tendent à ne pas se concentrer sur les *tokens* importants. Nous étudions ici différentes approches pour rendre les poids d'attention plus plausibles, en nous appuyant sur des masques issus d'une analyse morphosyntaxique ou sur une régularisation pour forcer la parcimonie. Nous montrons que ces stratégies permettent d'améliorer sensiblement la plausibilité des poids d'attention et s'avèrent plus performantes que les approches par carte de saillance.

ABSTRACT

Filtering and regularization to improve the plausibility of attention weights in NLI.

We investigate the plausibility of an attention mechanism for a natural language inference task, i.e., its ability to provide a human-plausible explanation of the relationship detected between two sentences. Exploiting the *Explanation-Augmented Standford Natural Language Inference* corpus, it has been shown that the attention weights are hardly plausible in practice and tend not to focus on the relevant tokens. We study here different approaches to increase the plausibility of attention weights, based on masks from morphosyntactic analysis or on regularization to force parsimony. We show that these strategies improve the plausibility of the attention weights and perform better than saliency map approaches.

MOTS-CLÉS : mécanisme d'attention, explicabilité, plausibilité, inférence en langue naturelle.

KEYWORDS: attention mechanism, explanability, plausibility, natural language inference.

1 Introduction

Les mécanismes d'attention sont largement utilisés depuis quelques années en traitement automatique des langues (TAL) du fait des gains de performances qu'ils procurent dans de nombreuses tâches (Luong *et al.*, 2015; Bahdanau *et al.*, 2015; Chen *et al.*, 2017; Vaswani *et al.*, 2017; Devlin *et al.*, 2019; Shen *et al.*, 2018). Au-delà de ces gains, ces mécanismes fournissent des poids d'attention pour chaque mot (ou *token*) d'une séquence en entrée, que l'on espère proportionnels au niveau d'influence du *token* sur la décision finale du modèle. Ces poids, visualisables sur une carte de chaleur, jouent potentiellement un rôle explicatif des décisions prises par le modèle pour un humain.

Qualifier ou quantifier l'explicabilité d'une carte d'attention est cependant difficile, le lien entre la performance du modèle sur la tâche à réaliser et l'explicabilité n'étant pas établi. Certaines études (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019) ont d'ailleurs montré que des modèles adverses pouvaient réaliser les mêmes prédictions tout en produisant des cartes de chaleur différentes. Dans cette quête de critères d'évaluation des cartes d'attention à des fins d'explication, la distinction introduite dans Jacovi & Goldberg (2020) entre fidélité et plausibilité est intéressante : la fidélité indique à quel point les poids d'attention reflètent le processus de raisonnement du modèle, là où la plausibilité s'intéresse à l'utilité des poids pour qu'un humain juge et interprète la décision. Si la plupart des travaux se focalisent sur le premier aspect (Xu *et al.*, 2015; Choi *et al.*, 2016; Ghaeini *et al.*, 2018; Lin *et al.*, 2017), très peu se sont jusqu'à présent intéressés à la plausibilité, en particulier du fait du manque de vérité-terrain. Mullenbach *et al.* (2018) est une des rares études de ce type où des annotateurs évaluent l'informativité de cartes de chaleur dans une tâche de classification de textes.

Pour notre part, nous nous intéressons à la plausiblité des mécanismes d'attention sur une tâche plus complexe d'inférence en langue naturelle, en nous appuyant sur l'annotation de référence fournie dans le corpus e-SNLI (Camburu *et al.*, 2018). Des travaux récents mettent en évidence que les mécanismes d'attention standards (poids d'attention croisée entre deux encodeurs LSTM) sont peu plausibles, alors qu'une heuristique simple consistant à se focaliser, en accord avec l'annotation humaine du corpus de référence, sur les noms, verbes et adjectifs, obtient une meilleure plausibilité (Nguyen *et al.*, 2021). L'objectif de cet article est donc d'explorer, toujours sur la tâche d'inférence en langue naturelle, des voies permettant de rendre les mécanismes d'attention plus plausibles. Introduire un critère de plausibilité à l'apprentissage, par exemple sous forme de fonction de coût sur les poids d'attention, s'avère difficile en pratique dans la mesure où il existe rarement des données annotées pour cela. Nous cherchons donc à introduire des critères simples pour améliorer la plausibilité. D'une part, nous examinons l'influence d'un filtrage sur les étiquettes morphosyntaxiques des mots des phrases prémisse et hypothèse sur la plausibilité de l'attention, que ce soit pendant ou après l'entraînement du modèle. D'autre part, nous étudions les potentialités d'une régularisation par minimisation d'entropie à accroître cette plausibilité sans avoir à indiquer des catégories morphosyntaxiques de filtrage.

2 Cadre expérimental

Dans cette section, nous présentons la tâche d'inférence en langue naturelle (ILN) – aussi connue sous le nom de reconnaissance d'implications textuelles – et le corpus sur lequel reposent nos expérimentations, avant de présenter l'architecture neuronale, classique pour une tâche d'ILN, qui nous permet de tester nos différentes hypothèses.

2.1 Inférence en langue naturelle : définitions et corpus de référence

L'inférence en langue naturelle consiste, pour une paire de phrases donnée, à reconnaître si une phrase prémisse entraîne, contredit ou est neutre par rapport à une phrase hypothèse. En plus de l'objectif de réaliser l'inférence avec de bons résultats, nous cherchons à obtenir un modèle capable d'expliquer correctement les parties des phrases qui justifient de façon convaincante la décision pour un humain.

Afin d'évaluer si l'explication obtenue est plausible de manière reproductible, nous utilisons le corpus *Explanation-Augmented Standford Natural Language Inference* (e-SNLI) (Camburu *et al.*, 2018). Celui-ci a été annoté par des personnes auxquelles il a été demandé de surligner, dans chaque

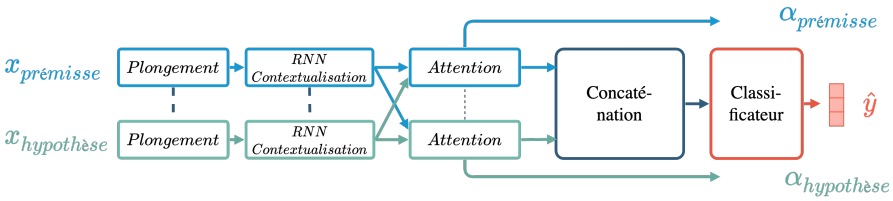


FIGURE 1 – Architecture du modèle utilisé et de la couche d'attention croisée (Nguyen *et al.*, 2021)

paire prémisse-hypothèse, les mots justifiant les étiquettes implication (*entailment*) ou contradiction posées. Ceci a conduit à un total de 18% de mots surlignés. Ce corpus de langue anglaise est, à notre connaissance, le seul de taille conséquente disposant de telles annotations. Ces annotations nous servent de référence comme explication plausible, que nous espérons atteindre par les modèles que nous testons.

L'ensemble d'entraînement est formé d'environ 180 000 paires de phrases, les ensembles de validation et de test en contenant chacun environ 3 000. Chaque phrase prémisse apparaît trois fois avec une hypothèse pour chacune des trois classes. Une analyse des annotations dans e-SNLI montre que les noms, verbes et adjectifs représentent, selon les trois ensembles de données, de 72% à 78% des mots surlignés par les annotateurs, ce qui suggère qu'une carte d'explication plausible doit tendre à contenir la même distribution de ces catégories morphosyntaxiques et se concentrer sur ces trois catégories.

2.2 Modèle d'attention

À des fins de comparaison, nous reprenons le modèle de Nguyen *et al.* (2021) : il s'appuie de manière classique sur une architecture siamoise, illustrée figure 1. Deux encodeurs LSTM (couche de contextualisation) sont liés par une couche d'attention croisée et contribuent à une couche de décision pour réaliser la tâche de classification finale pour une paire de phrases. Nous supposons que les phrases prémisse et hypothèse possèdent les mêmes attributs statistiques et les paramètres des couches de plongement, des LSTM et des couches d'attention sont partagés.

Pour la couche de plongement, nous utilisons un modèle GloVe pré-entraîné de dimension 300. La couche de contextualisation s'appuie quant à elle sur un LSTM bi-directionnel avec un état récurrent de dimension 300 également. Enfin, la couche de classification est de type perceptron avec une activation ReLU et une sortie de dimension 3. Le modèle est entraîné avec l'optimiseur Adam et un taux d'apprentissage de 10^{-3}.

3 Filtrage morphosyntaxique

L'heuristique définie dans (Nguyen *et al.*, 2021), qui consiste à sélectionner les noms, verbes et adjectifs pour l'explication, montre clairement que les étiquettes morphosyntaxiques jouent un rôle prépondérant pour améliorer la plausibilité des poids d'attention. Nous cherchons donc à exploiter

ces étiquettes en introduisant un filtrage des poids d'attention. Les étiquettes morphosyntaxiques sont générées par le modèle pré-entrainé de spaCy (Honnibal & Montani, 2017). Dans la suite, nous appelons *tokens* « peu informatifs », ceux dont l'étiquette n'est ni nom, ni verbe, ni adjectif. Un tel filtrage des *tokens* peu informatifs peut se faire soit *a posteriori*, après entraînement du modèle d'attention et en fonction des étiquettes morphosyntaxiques, soit *a priori* en imposant un masquage des *tokens* peu informatifs au niveau du mécanisme d'attention.

Soit $h_i \in \mathbb{R}^{300}$, $i \in [0, L-1]$, les états cachés pour une phrase de longueur L et $\overline{h}_{L-1}$ l'état caché final de la seconde phrase (plongement de la phrase). Nous mesurons la pertinence de chaque h_i par rapport à son homologue $\overline{h}_{L-1}$ par le score d'alignement $a_i = h_i \ \overline{h}_{L-1}$; les poids d'attention des *tokens* représentés par h_i par rapport à la phrase représentée par $\overline{h}_{L-1}$, notés $\alpha = [\alpha_1, ..., \alpha_i, ...\alpha_{L-1}]$, sont obtenus en normalisant a selon

$$\alpha_i = \frac{\exp(a_i)}{\sum_{k \in [0, L-1]} \exp(a_k)} \quad . \tag{1}$$

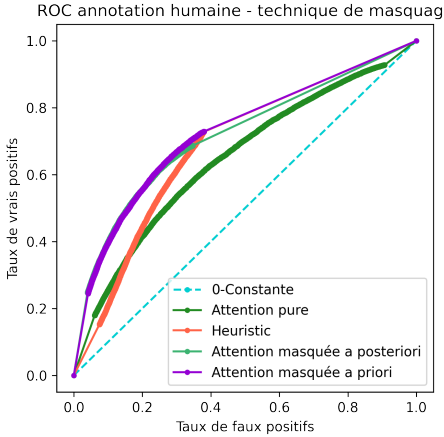
Ce vecteur d'attention α permet de définir une représentation de la première phrase dans le contexte de la seconde comme $c = \sum_i \alpha_i h_i$. Par facilité, nous noterons $\alpha_i = \text{softmax}(a_i)$. Les vecteurs de contexte c sont ensuite concaténés pour être utilisés directement pour la prédiction, $\hat{y} = \text{MLP}[c_p \oplus c_h])$, c_p (resp. c_h) étant le contexte de la prémisse (resp. de l'hypothèse) et $\oplus$ désignant la concaténation.

Le masquage *a posteriori* est le plus simple. Il consiste à modifier *a posteriori* les poids d'attention α_i pour forcer les poids des *tokens* peu informatifs à être nuls. En pratique, nous définissons un masque associé à chaque *token* d'une phrase, tel que $\text{mask}_i = -\infty$ si le *token* est peu informatif, 0 sinon. Ce masque est ajouté aux poids d'attention α_i de chaque *token*, i.e., $\tilde{\alpha}_i = \alpha_i + \text{mask}_i$, avant normalisation. En d'autres termes, le vecteur d'attention masqué α_i^{post} est défini comme $\text{softmax}(\alpha_i + \text{mask}_i)$, ce qui rend *de facto* les poids d'attention des *tokens* peu informatifs nuls.

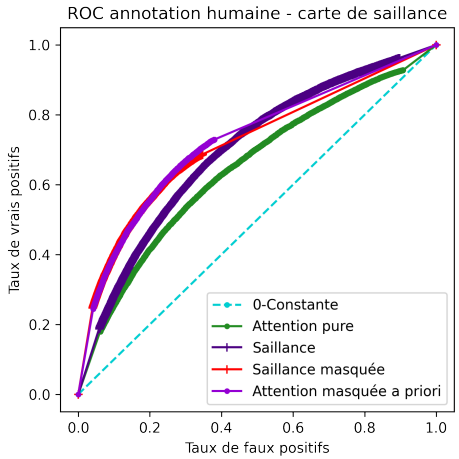
Cependant, le masquage *a posteriori* n'a aucune influence sur la classification qui s'appuie sur le modèle d'attention standard. C'est pourquoi nous proposons une stratégie de masque *a priori* qui permet de tenir compte de la notion de *tokens* peu informatifs directement dans le modèle et donc dans la décision.

L'idée du masque *a priori* est de forcer les poids d'attention des *tokens* peu informatifs à être nuls pendant l'apprentissage, permettant ainsi au modèle d'affiner sa représentation des *tokens* et son modèle d'attention. En utilisant la même définition de masque que précédemment, cela peut être obtenu en ajoutant le masque non plus aux α_i mais directement au score d'alignement a_i, soit $\alpha_i^{\text{prior}} = \text{softmax}(a_i + \text{mask}_i)$. Dans ce cas, le masque constitue une entrée supplémentaire au modèle et permet de concentrer l'attention sur les *tokens* informatifs.

Pour évaluer ces différentes approches, nous comparons les poids d'attention à la référence binaire fournie dans le corpus e-SNLI, ce qui ne peut se faire directement car $\alpha_i \in [0, 1]$. Nous binarisons donc les poids à l'aide d'un seuil de manière à pouvoir compter le nombre de *tokens* surlignés par les annotateurs correctement détectés (t.q. α_i au-dessus du seuil). En faisant varier ce seuil, nous obtenons une courbe de caractéristique de fonctionnement (ROC) : un seuil bas, qui sélectionne beaucoup de *tokens* pour l'explication, génère un fort taux de vrais positifs au prix de nombreux faux positifs tandis qu'un seuil élevé génère peu de faux positifs mais détecte peu de *tokens* pertinents. Les valeurs de seuil sont générées automatiquement par l'algorithme de (Fawcett, 2006) entre $[0, 1]$ afin de produire la courbe de ROC.



(a) ROC des différentes stratégies fondées sur les poids d'attention.



(b) Comparaison des ROC pour les méthodes fondées sur la saillance et celles fondées sur les poids d'attention.

FIGURE 2 – ROC pour l'évaluation de la plausibilité par poids d'attention et saillance

Méthodes	0-Constante	Heuristique	Attention pure	Masquage a posteriori	Masquage a priori
AUC	0.5	0.686	0.648	0.717	0.726

TABLE 1 – AUC des différentes stratégies fondées sur les poids d'attention.

Les résultats sont donnés à la figure 2a pour les trois variantes d'attention ci-dessus (attention classique, filtrage *a priori* ou *a posteriori*) et moyennés sur les paires de phrases de type *entailment* ou *contradiction*. Dans cette figure, la courbe *0-Constante* est une *baseline* correspondant à une carte avec tous les poids d'attention à 0 ; *Attention pure* indique la plausibilité de la carte d'attention du modèle standard par rapport à la vérité-terrain ; *Heuristic* montre celle obtenue grâce à l'heuristique de Nguyen *et al.* (2021) ; *Attention masquée a posteriori* présente la performance du modèle *Attention pure* filtré par les étiquettes morphosyntaxiques ; enfin, *Attention masquée a priori* présente celle du modèle s'entraînant avec le masquage. La moyenne de l'aire sous la courbe de ROC (*Area Under Curve ou AUC*) est une autre façon de quantifier cette performance. Cette valeur correspond à la hauteur de la courbe en moyenne. Le tableau 1 montre la valeur de cette métrique pour les courbes précédentes.

Le filtrage, *a priori* comme *a posteriori*, apporte une nette amélioration de la plausibilité des poids d'attention et surpasse l'heuristique, sans modifier les performances de classification (80 % pour le modèle original, 79.6 % pour le masque *a priori*). En revanche, il n'y a que très peu de différence entre les deux stratégies de filtrage.

Le mécanisme d'attention n'est pas le seul moyen pour obtenir une explication de la décision : en effet, les cartes de saillance sont souvent utilisées en pratique (Bastings & Filippova, 2020; Shahid & Debar, 2021). En calculant la rétro-propagation à partir de la sortie, la dérivée peut se propager jusqu'à l'entrée initiale, permettant ainsi de mettre en évidence la contribution de chacun des éléments

λ	Entropie	Précision	Plausibilité (AUROC)	Plausibilité a posteriori	δ
0.000	1.926513	0.786	0.643	0.710	0.094433
0.001	1.857389	0.789	0.672	0.717	0.045054
0.005	1.064913	0.785	0.721	0.725	0.003136
0.006	0.930867	0.786	0.727	0.722	-0.004391
0.007	0.278311	0.774	0.698	0.693	-0.005489
0.008	0.063424	0.764	0.648	0.584	-0.063735
0.009	0.067032	0.763	0.688	0.607	-0.080979
0.010	0.000446	0.759	0.557	0.551	-0.006205

TABLE 2 – Résultats avec la régularisation par minimum d'entropie pour différentes valeurs de λ (un *run* par valeur de λ)

de l'entrée à la sortie prédite. Nous comparons donc nos approches par attention à une méthode de saillance classique : l'entrée étant une séquence de vecteurs de dimension d, la dérivée à l'entrée est un vecteur de dimension d dont nous prenons la norme infinie pour calculer la saillance de chaque *token*. Les résultats sont données à la figure 2b où *Saillance* concerne l'explication issue de la carte de saillance, *Saillance masquée* considérent en plus un masquage *a posteriori* selon les catégories morpho-syntaxiques. La méthode de saillance offre une carte de chaleur des *tokens* en entrée qui est plus plausible que celle du modèle d'attention, mais cette méthode ne permet pas non plus de se concentrer sur les *tokens* informatifs. En effet, avec le filtrage morphosyntaxique *a posteriori* (c.-à-d. si nous forçons les valeurs de la carte de saillance sur les mots peu informatifs à être à 0), la plausibilité de la carte de saillance peut être améliorée et atteint à peu près le même niveau que la carte d'attention filtrée.

4 Régularisation par minimisation de l'entropie

Une autre approche pour améliorer la plausibilité du modèle d'attention consiste à régulariser les poids d'attention de manière à forcer leur parcimonie sans contraintes sur les caractéristiques morphosyntaxiques. En effet, il a été observé que les poids d'attention du modèle original sont assez diffus pour une phrase, mettant en avant de nombreux *tokens* peu informatifs, ce qui nous a amenés à proposer les approches par filtrage. À travers une régularisation forçant la parcimonie des poids d'attention, nous cherchons à voir si ces poids se concentrent de manière naturelle sur les *tokens* informatifs.

En première approximation, nous choisissons une régularisation par minimisation de l'entropie de la distribution des poids α_i, une entropie faible traduisant une part de déterminisme dans la distribution correspondant aux α_i, et donc une forme de parcimonie (Zhang *et al.*, 2019). Pour ce faire, nous ajoutons à la fonction de coût de classification L_{classif} (entropie croisée catégorielle) sur laquelle se fonde le modèle un coût de régularisation, c.-à-d.,

$$L_{\text{regul}} = L_{\text{classif}} + \lambda H(\alpha) \ , \tag{2}$$

où $H(\alpha) = -\sum_i \alpha_i log(\alpha_i)$.

Les résultats sont donnés dans la table 2 pour différentes valeurs de λ en utilisant la métrique AUC

pour mesurer la plausibilité. La plausibilité *a posteriori* correspond au filtrage *a posteriori* des poids d'attention, la dernière colonne rapportant la différence après et avant filtrage morphosyntaxique. Nous observons qu'avec des valeurs faibles de régularisation ($\lambda \leq 0.006$), la plausibilité de la carte d'attention s'améliore tandis que la précision du classifieur reste stable. Le filtrage par étiquette morphosyntaxique devient progressivement inutile comme on peut le voir sur la dernière colonne. La régularisation par entropie permet donc d'améliorer significativement la plausibilité des poids d'attention sans nécessiter de filtrage, tout en maintenant la qualité de la prédiction.

5 Conclusion et discussion

Les résultats expérimentaux que nous rapportons dans cet article montrent que des stratégies simples, comme le filtrage par les étiquettes morphosyntaxiques ou la régularisation, permettent de modifier les poids d'attention dans un modèle bi-LSTM siamois pour rendre ces poids plus plausibles, c'est-à-dire plus pertinents pour une interprétation par un humain de la décision dans une tâche d'inférence en langue naturelle. Plus précisement, nous montrons que le modèle peut être modifié de manière à rendre des poids d'attention plausibles, sans impact sur la qualité de la prédiction, ces approches surpassant celles fondées sur la saillance. La régularisation par minimum d'entropie, qui tend vers la parcimonie des poids d'attention, s'avère particulièrement efficace pour focaliser l'attention sur les *tokens* importants pour un humain, notamment les noms, verbes et adjectifs, sans nécessiter d'analyse morphosyntaxique.

Ces premiers résultats appellent naturellement des extensions, notamment l'utilisation d'autres mécanismes de régularisation ou différentes manières de prendre en compte les informations morphosyntaxiques en entrée du modèle. On peut également se demander si la régularisation des poids d'attention impacte l'explicabilité fondée sur la saillance. Se pose également la question de l'extension de ces résultats à d'autres modèles, notamment ceux fondés sur l'auto-attention. Enfin, la reproductibilité des résultats de plausibilité obtenus grâce aux modèles testés est aussi à évaluer sur des corpus autres que celui d'inférence en langue naturelle utilisé dans ce travail, afin d'étudier s'ils sont extensibles à des tâches d'explication d'autres natures. Par ailleurs, nous nous sommes volontairement limités à des approches non supervisées, n'utilisant pas l'annotation des *tokens* importants dans la phase d'apprentissage. Si ces approches comportent des avantages évidents, au premier plan desquels le fait de ne pas requérir de données annotées, la question de la supervision de l'attention reste cependant un sujet à explorer plus avant.

Plus généralement, cette notion de plausibilité de l'explication (par opposition à la fidélité de l'explication – *cf*. Jacovi & Goldberg (2020)) est au cœur de l'explicabilité en apprentissage automatique pour le TAL et les résultats de nos travaux contribuent au débat sur la pertinence de l'attention pour l'explicabilité – *cf*. la polémique entre Jain & Wallace (2019) et Wiegreffe & Pinter (2019).

Références

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations*.

BASTINGS J. & FILIPPOVA K. (2020). The Elephant in the Interpretability Room : Why Use Attention as Explanation when we Have Saliency Methods ? In *ACM BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, p. 149–155.

CAMBURU O.-M., ROCKTÄSCHEL T., LUKASIEWICZ T. & BLUNSOM P. (2018). e-SNLI : Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems 31*, p. 9539–9549.

CHEN Q., ZHU X., LING Z.-H., WEI S., JIANG H. & INKPEN D. (2017). Recurrent Neural Network-based Sentence Encoder with Gated Attention for Natural Language Inference. In *2nd Workshop on Evaluating Vector Space Representations for NLP*, p. 36–40.

CHOI E., BAHADORI M. T., SUN J., KULAS J., SCHUETZ A. & STEWART W. (2016). RETAIN : an Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems*, volume 29, p. 3512–3520.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4171–4186.

FAWCETT T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**(8), 861–874.

GHAEINI R., FERN X. & TADEPALLI P. (2018). Interpreting Recurrent and Attention-based Neural Models : a Case Study on Natural Language Inference. In *2018 Conf. on Empirical Methods in Natural Language Processing*, p. 4952–4957.

HONNIBAL M. & MONTANI I. (2017). spaCy 2 : Natural Language Understanding with Bloom Embeddings, Convolutional Neural networks and Incremental Parsing.

JACOVI A. & GOLDBERG Y. (2020). Towards Faithfully Interpretable NLP Systems : How Should we Define and Evaluate Faithfulness ? In *58th Annual Meeting of the Association for Computational Linguistics*, p. 4198–4205.

JAIN S. & WALLACE B. C. (2019). Attention is not Explanation. In *North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 3543–3556.

LIN Z., FENG M., SANTOS C. N. D., YU M., XIANG B., ZHOU B. & BENGIO Y. (2017). A Structured Self-attentive Sentence Embedding. In *5th International Conference on Learning Representations*.

LUONG T., PHAM H. & MANNING C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *2015 Conf. on Empirical Methods in Natural Language Processing*, p. 1412–1421.

MULLENBACH J., WIEGREFFE S., DUKE J., SUN J. & EISENSTEIN J. (2018). Explainable Prediction of Medical Codes from Clinical Text. In *North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1101–1111.

NGUYEN D. H., GRAVIER G. & SÉBILLOT P. (2021). A Study of the Plausibility of Attention between RNN Encoders in Natural Language Inference. In *IEEE International Conference on Machine Learning and Applications*, p. 1623–1629.

SERRANO S. & SMITH N. A. (2019). Is Attention Interpretable ? In *57th Annual Meeting of the Association for Computational Linguistics*, p. 2931–2951.

SHAHID M. R. & DEBAR H. (2021). CVSS-BERT : Explainable Natural Language Processing to Determine the Severity of a Computer Security Vulnerability from its Description. In *IEEE International Conference on Machine Learning and Applications*, p. 1600–1607.

SHEN T., ZHOU T., LONG G., JIANG J., PAN S. & ZHANG C. (2018). DiSAN : Directional Self-Attention Network for RNN/CNN-Free Language Understanding. In *32rd AAAI Conference on Artificial Intelligence*, p. 5446–5455.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you Need. In *Advances in Neural Information Processing Systems*, volume 30.

WIEGREFFE S. & PINTER Y. (2019). Attention is not not Explanation. In *2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, p. 11–20.

XU K., BA J., KIROS R., CHO K., COURVILLE A., SALAKHUDINOV R., ZEMEL R. & BENGIO Y. (2015). Show, Attend and Tell : Neural Image Caption Generation with Visual Attention. In F. BACH & D. BLEI, Éds., *32nd International Conference on Machine Learning*, p. 2048–2057.

ZHANG J., ZHAO Y., LI H. & ZONG C. (2019). Attention with Sparsity Regularization for Neural Machine Translation and Summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(3), 507–518.