Encodeur hybride pour la détection automatique de désinformation

Géraud Faye[1,2] **Wassila Ouerdane**[2] **Sylvain Gatepaille**[1] **Guillaume Gadek**[1]
Souhir Gahbiche[1]

[1] Airbus Defence and Space, Élancourt, France
[2] Université Paris-Saclay, CentraleSupélec, MICS, France

geraud.faye@centralesupelec.fr

Résumé

L'encodage de texte est aujourd'hui basé sur de larges modèles de langue entièrement neuronaux, utilisés comme des *boîtes noires*. Afin de répondre au besoin d'explicabilité, nous proposons **CATS** [1] (Cognitive Attention To Syntax), une couche pouvant être utilisée dans les réseaux de neurones qui permet d'introduire du raisonnement syntaxique pour l'encodage et la classification de textes.

Abstract

Today, text encoding relies mostly on foundation models, used as *black boxes*. To bring more transparency, we propose **CATS** (Cognitive Attention To Syntax), a layer that incorporates syntactic reasoning in neural networks for text classification.

1 CATS, un encodeur neurosymbolique

La détection de désinformation à partir de textes uniquement a fait l'objet du développement de nombreux modèles. Les réseaux convolutionnels [3] et les modèles basés sur l'attention [5] ont montré à travers leurs résultats que les caractéristiques stylistiques du document sont discriminantes pour identifier la désinformation. Ces caractéristiques semblent même être partagées par plusieurs types de désinformation (rumeur, fake news, . . .) [4]. Afin de créer un modèle plus transparent, nous proposons une approche neurosymbolique qui introduit du raisonnement syntaxique pour l'encodage de la phrase, en s'inspirant du principe de compositionnalité.

Dans un premier temps, les phrases sont analysées syntaxiquement, produisant un arbre semblable à celui de la Figure 1.

1. Une version étendue a été publiée dans les actes de EGC 2023 [2]

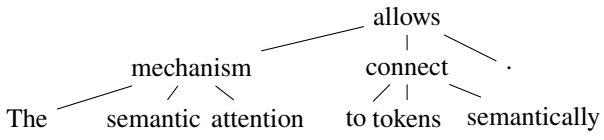

FIGURE 1 – Arbre syntaxique calculé par SpaCy

Cet arbre met en relation les mots entre eux en fonction de leur importance relative. Il est ensuite utilisé pour calculer une matrice d'attention qui relie les mots entre eux du bas vers le haut en fonction de leur distance dans l'arbre syntaxique, donnant la matrice présentée en Figure 2. Cette matrice est ensuite utilisée dans le mécanisme d'attention classique [7]. En retirant les projections vers les espaces *Query*, *Key* et *Value* , nous avons une couche neuronale basée sur du raisonnement symbolique qui permet la propagation des gradients, tout en n'ayant aucun poids entraînable.

2 Évaluation et explicabilité

Cette couche neurosymbolique a été comparée à son équivalent neuronal (noté Standard) dans un modèle utilisant les plongements de fastText [1], suivis de la couche d'attention standard ou de CATS, avant une couche de classification. Ces modèles ont été évalués sur PolitiFact et GossipCop [6] (détection de *fake news* et de rumeurs). Les résultats sont consignés dans la Table 1.

Les modèles basés sur CATS ont des performances légèrement supérieures aux modèles uniquement neuronaux. Mais le grand avantage de CATS est la réduction du besoin en données annotées qui a été mesuré avec des entraîne-

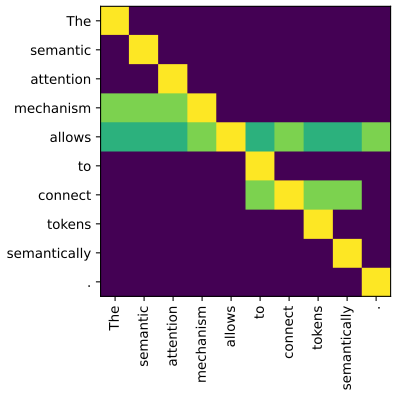

FIGURE 2 – Matrice d'attention correspondant à l'arbre Figure 1. Plus la case est claire, plus le poids correspondant est élevé.

	PolitiFact		GossipCop	
	Fiabilité	F1	Fiabilité	F1
Standard	0.889	0.902	0.727	0. 758
CATS	**0.916**	**0.929**	**0.732**	**0.762**

TABLE 1 – Résultats des différents modèles sur le jeu de données de test.

ments sur un dataset réduit (voir Figure 3). Le modèle neurosymbolique est capable de généraliser avec seulement 50 articles alors que le modèle neuronal n'apprend rien avec si peu de données.

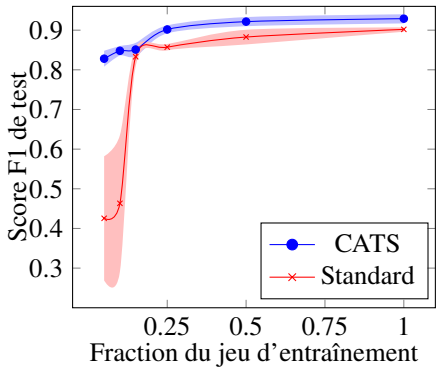

FIGURE 3 – Évolution du score F1 d'un apprentissage sur une portion du jeu de données PolitiFact.

Enfin, car notre matrice d'attention est inversible par construction, on peut calculer la contribution de chaque jeton dans la prédiction. Cela nous a permis de développer un outil identifiant les phrases les plus suspectes dans un texte, afin de faciliter la vérification manuelle des faits ou de mettre en garde les utilisateurs contre les marqueurs probables de désinformation, comme illustré dans la Figure 4.

Le modèle n'identifie pas les faits factuellement faux, mais les expressions et tournures de phrases caractéristiques de la désinformation.

Twelve More Hurricanes Headed Towards US . Twelve More Hurricanes Headed Towards US This is a satirical website . Don ' t take it Seriously . It ' s a joke . Wednesday 06 July 2059 4332 Shares The National Hurricane Center has issued twelve more hurricane warnings for the east coast of the US .

" Regardless of which coast you live on , be prepared to evacuate at least twelve times " the National Weather Services said Thursday , not ruling out the possibility of a thirteenth hurricane by the end of the year . This is a satirical website . Don ' t take it Seriously . It ' s a joke . / ! \ Report Abuse loading Biewty

FIGURE 4 – Les indicateurs explicites de satire sont mis en évidence (rouge) pour le lecteur trop hâtif. Les phrases en vert ne contribuent pas à la classe désinformation.

Il serait intéressant par la suite d'incorporer un mécanisme de coréférence, ce qui permettrait de lier les phrases entre elles à partir de leurs entités communes.

Références

[1] Bojanowski, Piotr, Edouard Grave, Armand Joulin et Tomás Mikolov: *Enriching word vectors with subword information*. CoRR, abs/1607.04606, 2016.

[2] Faye, Géraud, Sylvain Gatepaille, Guillaume Gadek et Souhir Gahbiche: *Encodeur hybride pour la détection automatique de désinformation*. Revue des Nouvelles Technologies de l'Information, Extraction et Gestion des Connaissances, RNTI-E-39 :91–102, 2023.

[3] Gadek, Guillaume et Paul Guélorget: *An interpretable model to measure fakeness and emotion in news*. Procedia Computer Science, 176:78–87, 2020.

[4] Lee, Nayeon, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen tau Yih et Madian Khabsa: *On Unifying Misinformation Detection*, avril 2021.

[5] Pelrine, Kellin, Jacob Danovitch et Reihaneh Rabbany: *The Surprising Performance of Simple Baselines for Misinformation Detection*, avril 2021.

[6] Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee et Huan Liu: *Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media*. CoRR, abs/1809.01286, 2018.

[7] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser et Illia Polosukhin: *Attention is all you need*. CoRR, abs/1706.03762, 2017.