Apprentissage multimodal modulaire pour l'extraction de théorèmes et de preuves dans des documents scientifiques longs

Shrey Mishra*, Antoine Gauquier*, Pierre Senellart*,**

* DI ENS, ENS, CNRS, Université PSL, Inria
** IUF

1 Introduction

Les articles scientifiques dans les domaines mathématiques incluent des théorèmes (ainsi que d'autres environnements similaires) avec leurs preuves. Dans le cadre du projet TheoremKB (Mishra et al., 2024a), qui vise à transformer la littérature scientifique d'une collection de PDF en une base de connaissance de théorèmes, nous résumons nos travaux sur l'extraction automatique des théorèmes et de leurs preuves depuis des PDF (Mishra et al., 2024b).

Pour être précis, nous utilisons le terme *théorème* de la même manière qu'il est utilisé en LaTeX (p. ex., par la commande `\newtheorem`) : un environnement de type théorème est un énoncé structuré, éventuellement numéroté, formaté d'une manière spécifique et utilisé pour représenter un énoncé formel (en général mathématique) : cela peut être un théorème, un lemme, une proposition, etc., mais également une définition, une remarque formelle ou un exemple. Une *preuve* est ce qu'on met habituellement dans un environnement `proof` en LaTeX : la démonstration ou ébauche de démonstration d'un résultat.

Nous approchons ce problème d'identification des théorèmes et des preuves en concevant une approche basée sur l'apprentissage multimodal qui classifie chaque paragraphe d'un article en des classes *Basique*, *Théorème* et *Preuve*, en fonction du texte en anglais, d'informations typographiques et du rendu visuel des articles PDF. De plus, nous prenons en compte la *séquence* des paragraphes, leur coordonnées spatiales et les numéros de pages, afin d'exploiter le fait que l'étiquette d'un paragraphe dépend de celle des paragraphes précédents (ou suivants).

Notre contribution, résumée en Fig. 1, comporte ainsi : (i) Trois modèles unimodaux (vision, texte, information de police) pour le problème d'identification des théorèmes et preuves qui reposent sur des techniques modernes d'apprentissage profond (CNN, transformeurs, LSTM) avec un accent mis sur des modèles efficaces plutôt que de très grands modèles ; le modèle de texte requiert le pré-entraînement d'un modèle de langue spécifique à notre corpus, qui peut avoir d'autres applications. (ii) Un modèle multimodal par fusion tardive qui combine les plongements des trois modalités. (iii) Une approche séquentielle au niveau des paragraphes, à base de transformeur, qui peut être utilisée pour améliorer la performance de n'importe quel modèle unimodal ou multimodal en capturant les dépendances entre paragraphes. (iv) Une évaluation expérimentale sur un jeu de données d'environ 200 000 articles en anglais issus d'arXiv, avec un jeu de données de validation séparé de 3 500 articles (soit $\approx$ 529 000 paragraphes).

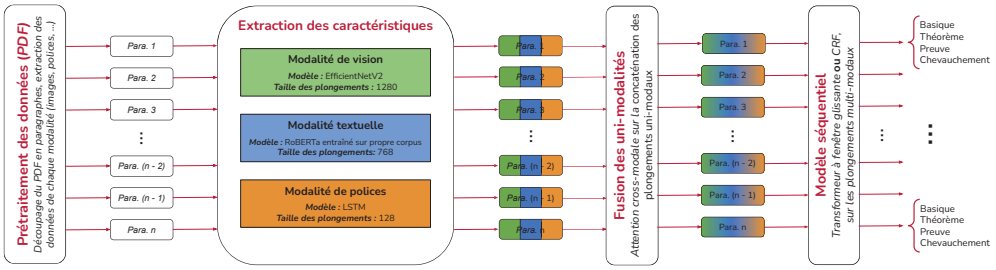

FIG. 1 – *Chaîne de traitement pour l'inférence multimodale*

TAB. 1 – *Comparaison des performances (précision et F_1 moyen sur les trois classes) des modèles unimodaux et multimodaux, avec ou sans approche séquentielle ; 1 lot = 1 000 documents*

Modalité	Modèle	Approche séq.	#Lots	#Paramètres (M)	Précision (%)	F_1 moyen (%)
Basique	Prédit toujours *Basique*	—	—	—	59.41	24.85
Top-k premier mot	Utilise le premier mot	—	—	—	52.84	44.20
Ligne à ligne	Bert (affiné)	—	—	110	57.31	55.71
Polices	LSTM 128 cellules	-	11	2	64.93	45.48
		CRF	11+8	2	71.50	64.51
		Transformeur FG	11+8	2	76.22	71.77
Vision	EfficientNetV2m	-	9	53	69.44	60.33
		CRF	9+8	53	74.63	70.82
		Transformeur FG	9+8	65	79.59	77.66
Texte	RoBERTa entraîné sur propre corpus	-	20	124	76.45	72.33
		CRF	20+8	124	83.10	80.99
		Transformeur FG	20+8	129	87.50	86.67
Multimodal	Attention cross-modale	-	2	185	78.50	75.37
		CRF	2+8	185	84.39	82.91
		Transformeur FG	2+8	198	**87.81**	**87.18**

2 Résultats

Nous comparons dans le Tab. 1 les performances de nos modèles, avec ou sans modèle séquentiel (CRF ou transformeur à fenêtre glissante), à trois compétiteurs : un prédicteur constant, pour référence ; une approche inspirée de (Ginev et Miller, 2020) se focalisant sur le premier mot de chaque paragraphe ; le classifieur textuel ligne à ligne de (Mishra et al., 2021).

Références

Ginev, D. et B. R. Miller (2020). Scientific statement classification over arXiv.org. In *LREC*.

Mishra, S., Y. Brihmouche, T. Delemazure, A. Gauquier, et P. Senellart (2024a). First steps in building a knowledge base of mathematical results. In *SDP*.

Mishra, S., A. Gauquier, et P. Senellart (2024b). Modular multimodal machine learning for extraction of theorems and proofs in long scientific documents. In *JCDL*.

Mishra, S., L. Pluvinage, et P. Senellart (2021). Towards extraction of theorems and proofs in scholarly articles. In *DocEng*.