Description du projet doctoral - École doctorale SMI

Proposé par: Olivier Pons olivier.pons@lecnam.net et Éric Gressier Soudan eric.gressier_soudan@cnam.fr

Analyse, structuration et génération de document, le cas des manuels scolaires

Contexte et Introduction

Ce projet doctoral s'inscrit dans la continuité des travaux de l'ANR MAnuels sco-LAires INclusifs (MALIN) dont l'objectif est de rendre utilisables les manuels scolaires numériques par les enfants en situation de handicap (avec un focus sur la dyspraxie et la déficience visuelle).

L'inclusion scolaire des élèves en situation de handicap est un enjeu sociétal majeur en France depuis la loi de 2005. Cependant, l'accessibilité des manuels scolaires numériques reste un défi. Actuellement, les adaptations sont souvent effectuées manuellement, ce qui est long (4 à 6 mois pour un manuel de primaire) et coûteux. De plus, ces adaptations ne sont pas toujours optimales en termes d'accessibilité et d'efficacité pédagogique.

Rendre accessible ces manuels scolaires numériques est une préoccupation essentielle et un challenge technique pour les éditeurs scolaires dont la chaîne de production existante est essentiellement basée sur le modèle des manuels imprimés. Il n'existe pas en France, de normes légales obligatoires pour faciliter la mise en accessibilité des manuels contrairement aux États-Unis. Aux USA, la loi fédérale sur l'éducation des personnes handicapées (Individuals with Disabilities Education Act) impose aux États qui acceptent un financement fédéral le format NIMAS (National Instructional Materials Accessibility Standard). D'autres pays tendent également à imposer ce format XML qui facilite l'accès aux données mais dont la structuration est très insuffisante, notamment pour le traitement des interactions.

Le projet MALIN qui repose sur une collaboration entre quatre laboratoires : LISN (Université Paris Saclay), MICS (Ecole CentraleSupelec), CEDRIC (CNAM), Inserm 1284 (CRI, Université de Paris), et interagit tant avec les éditeurs qu'avec les organismes de transcription, vise à développer des solutions techniques innovantes afin de permettre une automatisation des adaptations.

Une étape fondamentale vers cette tâche consiste a reconnaître la structure de documents lors de leur analyse[2] pour les convertir en un format structuré servant de base à d'autres applications. D'un point de vu purement scientifique, l'analyse, la structuration et la génération de document sont des sujets de recherche importants qui constituent encore souvent un vrai verrou scientifique. Dans ce contexte les manuels scolaires, au vu de leur complexité et de la variété des structures, constituent un excellent challenge.

Objectif et méthodes

L'objectif de la thèse consiste, à partir de manuels au format PDF natif ou d'images scannées, à inférer et comprendre leur structure et leur sémantique et à produire une version dans un(des) format(s) structuré(s) sur lesquels se reposeront les différentes méthodes d'adaptation.

À partir de PDF natifs, plusieurs approches d'extraction automatique de la structure d'un manuel scolaire (cours, exercices, consignes, énoncés, exemples, etc.) et de son contenu multimédia (textes, images, etc.) ont déjà été étudiées. Nos expériences à base de transformers multimodaux ont donné des résultats prometteurs pour les manuels d'étude de la langue [9], mais butent sur la faible quantité de données annotées. Les manuels

scannés présentent une difficulté supplémentaire car leur mise en page n'est pas disponible directement sous forme de métadonnées et la reconnaissance automatique des caractères (OCR) n'est pas suffisante pour l'obtenir. Il faut donc définir une approche pour l'analyse et l'extraction de la mise en page (Document Layout Analysis).

Lei Cui et al. [4] font un panorama des techniques d'apprentissage profond utilisées pour la reconnaissance de la mise en page de documents. On peut retenir plusieurs catégories : Réseaux de Neurones Convolutifs (CNN), Réseaux de Neurones à base de Graph (GNN). Les architectures basées sur Faster R-CNN [14] Mask R-CNN [8] et YOLO [13] ont été largement utilisées dans plusieurs benchmarks pour détecter différents "objets" de la page, tandis que LayoutLM [16] a été la première architecture multimodale basée sur les transformers appliquée à l'analyse de layout de documents. Gemelli et al. [7] ont proposé un GNN pour s'attaquer à l'analyse de la mise en page et à la compréhension des tableaux en même temps. Zhang P. et al. [17] ont proposé un cadre de détection d'objets basé sur Mask R-CNN multimodal qui utilise la vision, le langage et la géométrie, qui a dominé la compétition ICDAR21.

L'évolution des techniques est très rapide, mais la majorité des méthodes utilisées dépend fortement de l'apprentissage supervisé, et la disponibilité de données annotées est un point bloquant.

De « gros » corpus tels que PubLayNet [18, 1] (qui regroupe essentiellement des images annotées issues d'articles de PubMed) ou DocLayNet [11, 6] (qui est plus générique) sont disponibles, mais la mise en page des manuels scolaires est beaucoup plus complexe, et la mise à disposition de tels corpus est compliquée par les restrictions de droit sur les manuels existants.

Outre le fine-tuning de modèles existants, il sera nécessaire de construire de nouveaux corpus, en annotant manuellement ou par des méthodes ad hoc d'IA symbolique (règles sur les fontes, X-Y Cut, etc.).

Une autre solution envisagée face au manque de données étiquetées sera l'augmentation de données par la génération automatique de manuels scolaires. Cela se fait déjà pour d'autres catégories de documents comme les articles scientifiques [12].

Les modèles de génération envisagés pourront, dans un premier temps, être purement textuels et s'appuyer sur de larges modèles de langue (LLM) [5, 15] pré-entraînés et adaptés pour cette tâche. Mais la mise en page est un composant fondamental de tout design graphique, et la génération de mises en page de documents plausibles connaît récemment une explosion à la fois dans la littérature académique [12, 10, 3] et dans les applications. On s'intéressera donc aussi à la génération de manuels scolaires présentant des mises en forme réalistes, qui, outre leur intérêt propre, pourront servir à entraîner les modèles d'analyse de documents.

Enfin, pour pallier le problème de droits, on pourra s'intéresser à des documents dont la mise en forme est assez proche de celle des manuels, notamment via les ressources internes du CNUM (Conservatoire numérique des Arts et Métiers), qui est hébergé par le Cédric.

Échéancier

— Les premiers mois de la thèse seront consacrés à l'étude bibliographique des différents aspects du sujet : méthodes d'analyse de documents, modèles de manuels scolaires, structuration et génération de documents. Ce sera aussi une période de compréhension et d'expérimentation des différents corpus et modèles déjà développés dans le projet.

— La partie principale (30 mois) du travail sera rythmée par des réunions régulières avec l'équipe d'encadrement et les différents partenaires du projet afin de

développer et valider les résultats, coordonner les besoins et faire émerger de nouvelles pistes. Durant cette période, le doctorant sera encouragé à participer et à soumettre ses travaux dans les conférences du domaine et éventuellement à faire participer les outils développés aux compétitions annuelles de la conférence ICDAR (International Conference on Document Analysis and Recognition)[26].

— Les derniers mois de la thèse seront consacrés à la rédaction du manuscrit et à la préparation de la soutenance.

Références

[1] I. R. Australia : Publaynet : A large dataset of document images from pubmed central open access subset, 2022.

[2] G. M. Binmakhashen et S. A. Mahmoud : Document layout analysis : a comprehensive survey. ACM Computing Surveys (CSUR), 52(6):1–36, 2019.

[3] S. Biswas, P. Riba, J. Lladós et U. Pal : Graph-based deep generative modelling for document layout generation. In International Conference on Document Analysis and Recognition, p. 525–537. Springer, 2021.

[4] L. Cui, Y. Xu, T. Lv et F. Wei : Document ai : Benchmarks, models and applications. arXiv preprint arXiv :2111.08609, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee et K. Toutanova : Bert : Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv :1810.04805, 2018.

[6] DS4SD : Doclaynet : A large-scale dataset for document layout analysis, 2022.

[7] A. Gemelli, E. Vivoli et S. Marinai : Graph neural networks and representation embedding for table extraction in pdf documents. In 2022 26th International Conference on Pattern Recognition (ICPR), p. 1719–1726. IEEE, 2022.

[8] K. He, G. Gkioxari, P. Dollár et R. Girshick : Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, p. 2961–2969, 2017.

[9] E. Lincker, O. Pons, C. Guinaudeau, I. Barbet, J. Dupire, C. Hudelot et C. ... Huron : Layout-and activity-based textbook modeling for automatic pdf textbook extraction. In Intelligent Textbooks 2023, vol. 3444 de CEUR-WS.org, p. 37–53, jul 2023.

[10] A. G. Patil, O. Ben-Eliezer, O. Perel et H. Averbuch-Elor : Read : Recursive autoencoders for document layout generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, p. 544–545, 2020.

[11] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar et P. Staar : Doclaynet : a large human-annotated dataset for document-layout segmentation. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, p. 3743–3751, 2022.

[12] L. Pisaneschi, A. Gemelli et S. Marinai : Automatic generation of scientific papers for data augmentation in document layout analysis. Pattern Recognition Letters, 167:38–44, 2023.

[13] J. Redmon, S. Divvala, R. Girshick et A. Farhadi : You only look once : Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, p. 779–788, 2016.

[14] S. Ren, K. He, R. Girshick et J. Sun : Faster r-cnn : Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.

[15] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al. : Llama 2 : Open foundation and fine-tuned chat models. arXiv preprint arXiv :2307.09288, 2023.

[16] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei et M. Zhou : Layoutlm : Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, p. 1192–1200, 2020.

[17] P. Zhang, C. Li, L. Qiao, Z. Cheng, S. Pu, Y. Niu et F. Wu : Vsr : a unified framework for document layout analysis combining vision, semantics and relations. In Document Analysis and Recognition–ICDAR 2021 : 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16, p. 115–130. Springer, 2021.

[18] X. Zhong, J. Tang et A. J. Yepes : Publaynet : largest dataset ever for document layout analysis. In 2019 International conference on document analysis and recognition (ICDAR), p. 1015–1022. IEEE, 2019.