Extraction d'information induite par sous-graphes (SETI) appliquée aux documents administratifs

Dipendra Sharma-Kafle[1,*,†], Eliott Thomas[1,2,†], Mickael Coustaty[1,†], Aurélie Joseph[2,†], Antoine Doucet[1,†] and Vincent Poulain d'Andecy[2,†]

[1]*Université de La Rochelle, L3i, Avenue Michel Crépeau, 17042 La Rochelle, France*
[2]*Yooz, Immeuble le sequoia, Parc d'Andron, 30470 Aimargues, France*

Abstract

L'extraction d'informations joue un rôle clé dans le traitement automatique des documents administratifs. Cependant, la variété dans la mise en page et la langue constitue toujours une tâche difficile. D'autre part, il est rare de trouver d'importants ensembles de données d'entraînement publics liés aux documents administratifs tels que les factures. Dans ce travail, nous utilisons le modèle Graph Attention Network pour l'extraction d'informations. Ce type de modèle facilite la compréhension du mécanisme de classification par rapport aux réseaux neuronaux classiques en raison de la visualisation du lien entre les entités dans le graphe. De plus, il maximise la récupération de la mise en page et de la structure, ce qui constitue un avantage crucial dans les documents administratifs. À partir du même graphe, notre modèle apprend à différents niveaux du graphe pour encapsuler des connaissances dynamiques et plus riches dans chaque lot, maximisant ainsi la généralisation sur des ensembles de données plus petits. Nous montrons comment le modèle apprend à chaque niveau du graphe et comparons les résultats avec des bases sur des ensembles de données privés ainsi que publics. Notre modèle permet de améliorer les scores de rappel et de précision pour certaines classes dans notre ensemble de données privé et produit des résultats comparables pour les ensembles de données publics destinés à la compréhension des formulaires et à l'extraction d'informations. Cette soumission est le résumé traduit d'un article publié à la conférence ICDAR 2023 [1].

Keywords

Extraction d'informations, Facture, Réseau neuronal graphe

1. Introduction

L'extraction d'informations est cruciale pour automatiser le traitement des documents administratifs, mais la diversité dans la mise en page et la langue reste un défi. De plus, il est rare de disposer de vastes ensembles de données d'entraînement publics liés aux documents administratifs, comme les factures. Dans cette étude, nous employons le modèle Graph Attention Network pour l'extraction d'informations à différents niveaux de graphes.

CORIA (COnférence en Recherche d'Information et Applications), Avril 03–04, 2024, La Rochelle
*Corresponding author.
†These authors contributed equally.
dipendra.sharma_kafle@univ-lr.fr (D. Sharma-Kafle); eliott.thomas@univ-lr.fr (E. Thomas); mickael.coustaty@univ-lr.fr (M. Coustaty); aurelie.joseph@getyooz.com (A. Joseph); antoine.doucet@univ-lr.fr (A. Doucet); vincent.poulaindandecy@getyooz.com (V. P. d'Andecy)
0009-0006-5643-0773 (D. Sharma-Kafle); 0009-0008-5266-8797 (E. Thomas); 0000-0002-0123-439X (M. Coustaty); 0000-0002-5499-6355 (A. Joseph); 0000-0001-6160-3356 (A. Doucet)

2. Méthodologie

Notre approche, distincte des avancées récentes [2, 3, 4, 5], montre la nécessité de connaissances spécifiques par classe. Nous utilisons des réseaux de neurones sur des graphes (GNN) pour extraire des informations textuelles et de mise en page. Optant pour un réseau neuronal à attention pour les graphes (GAT) [2], combinant les avantages des transformers [6] et des graphes. Notre approche convertit les documents d'entrée en graphes, traitant chaque mot comme un nœud, et les arêtes sont formées en établissant des voisinages locaux [7]. Pour la connaissance bidimensionnelle, nous utilisons les coordonnées des tokens, améliorant les embeddings de BERT avec des caractéristiques booléennes. La vectorisation du texte est réalisée à l'aide de Camembert[8], modèle pré-entraîné en français. Enfin, et afin de capturer une connaissance approfondie du document, nous avons utilisé un modèle fonctionnant à trois échelles : **Niveau graphe global** (pour capturer des informations au niveau du document), **Niveau sous-graphe local** (pour capturer des sous-parties de documents), et **Niveau sous-graphe global** (pour capturer les sous-parties de graphes communes à l'ensemble de documents).

3. Expériences et Résultats

Nous avons mené des expériences sur notre ensemble de données privé, afin d'extraire des informations spécifiques telles que la date, le montant de la taxe, le montant total/net, le numéro du document, le type de document, etc. Nous avons également testé le modèle sur des jeux de données publics comme SROIE (la société, la date, le montant, l'adresse) [9] et FUNSD (l'en-tête, la question, la réponse, l'autre)[10] pour la tâche de compréhension de formulaires. Les scores f1 pour la prédiction des balises ont augmenté de 15,38 % en moyenne lorsque l'on passe du niveau du graphe au niveau du sous-graphe. Cependant, le niveau global reste crucial, puisqu'il nous permet d'obtenir une information de contexte nécessaire pour les différentes informations à extraire. Plus précisément, la combinaison globale/locale est bénéfique avec une capacité à extraire des informations spécifiques au niveau local, et des informations s'appuyant sur un contexte au niveau global. La prédiction de 7 étiquettes sur 8 a été améliorée par l'utilisation de graphes à la place de BERT. Notre méthode a dépassé de 9% le benchmark FUNSD, avec un score de 90,29% dans SROIE, proche des 90,49% de l'équipe gagnante.

4. Conclusion

Avec des connaissances renforcées provenant des sous-graphes, nous avons amélioré les scores de rappel et de précision pour la plupart des balises de notre ensemble de données privé par rapport à la méthode des transformers. Notre approche basée sur les graphes s'est avérée plus performante pour les informations structurées, exploitant la nature structurée des données et capturant les relations entre les entités. Pour les ensembles de données publics, les résultats de notre méthode sont comparables à ceux des modèles basés sur le texte. Les travaux futurs visent à améliorer notre approche de contexte sélectif et à mettre en œuvre un pipeline en cascade pour une extraction améliorée, capturant une gamme plus large de motifs et de relations dans les données.

Acknowledgments

Ce travail a été soutenu par le gouvernement français et par l'Union européenne dans le cadre du programme France Relance. Nous tenons également à remercier Guenael Manic, Mohammed Saadi, Jonathan Ouellet et Jerome Lacour de chez Yooz pour leur soutien.

References

[1] D. Sharma Kafle, E. Thomas, M. Coustaty, A. Joseph, A. Doucet, V. Poulain d'Andecy, Subgraph-induced extraction technique for information (seti) from administrative documents, in: International Conference on Document Analysis and Recognition, Springer, 2023, pp. 108–122.

[2] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903 (2017).

[3] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, Layoutlmv3: Pre-training for document ai with unified text and image masking, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4083–4091.

[4] M. Carbonell, P. Riba, M. Villegas, A. Fornés, J. Lladós, Named entity recognition and relation extraction with graph neural networks in semi structured documents, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 9622–9627.

[5] C.-Y. Lee, C.-L. Li, C. Wang, R. Wang, Y. Fujii, S. Qin, A. Popat, T. Pfister, Rope: reading order equivariant positional encoding for graph-based document information extraction, arXiv preprint arXiv:2106.10786 (2021).

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

[7] L. E. Peterson, K-nearest neighbor, Scholarpedia 4 (2009) 1883.

[8] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, arXiv preprint arXiv:1911.03894 (2019).

[9] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, C. Jawahar, Icdar2019 competition on scanned receipt ocr and information extraction, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1516–1520.

[10] G. Jaume, H. K. Ekenel, J.-P. Thiran, Funsd: A dataset for form understanding in noisy scanned documents, in: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), volume 2, IEEE, 2019, pp. 1–6.