Extracting Key-Value Pairs in Business Documents

Eliott Thomas[1,2,*,†], Dipendra Sharma-Kafle[1,†], Ibrahim Souleiman Mahamoud[1,2,†], Aurélie Joseph[2,†], Mickael Coustaty[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

Cet article présente une méthode combinant des règles métiers avec un modèle transformers pré-entraîné pour l'extraction des couples clé-valeur dans les documents administratifs, comme les factures. Cette approche offre une pipeline complète, de l'extraction OCR à la sortie de la liste des couples clé-valeur, et est facilement adaptable aux besoins commerciaux. Cette pipeline est également conçue pour être multilingue, afin de répondre à la diversité des documents. Issu d'une collaboration entre le laboratoire L3i et l'entreprise Yooz, ce travail propose une contribution avec un impact sur des domaines tels que la comptabilité et la prise de décision. Ce résumé traduit une recherche présentée à la conférence ICDAR 2023 [1].

Keywords

Clé-Valeur, Documents Commerciaux, Transformers Pré-Entraînés

1. Introduction

Ces dernières années, le domaine de l'extraction des couples clés-valeurs a connu une évolution significative, avec l'émergence de modèles de pointe tels que BERT[2] et ses variantes, notamment Roberta[3]. Des efforts ont été déployés pour développer des méthodes spécifiquement adaptées aux documents, comme LayoutLM et BROS[4, 5], ainsi que des modèles multilingues et multimodaux, tels que Info-XLM, XLM-Roberta, LayoutLMV2, LayoutLMV3 et GeoLayoutLM[6, 7, 8, 9, 10]. Dans ce contexte, notre étude se focalise sur le développement d'une pipeline complète pour l'extraction des couples clés-valeurs à partir de documents administratifs, tels que les factures. Cette pipeline se distingue par sa légèreté en termes de nombre de paramètres, sa facilité d'adaptabilité, ainsi que sa capacité à traiter des documents multilingues. Nous nous positionnons par rapport aux travaux existants en tant que baselines pour évaluer les performances de notre approche novatrice. En mettant en lumière ces aspects, notre objectif est de contribuer à l'amélioration des performances des systèmes d'extraction des couples clés-valeurs et à leur pertinence dans des contextes d'utilisation réels.

CORIA (COnférence en Recherche d'Information et Applications), Avril 03–04, 2024, La Rochelle
*Corresponding author.
†These authors contributed equally.
eliott.thomas@univ-lr.fr (E. Thomas); dipendra.sharma_kafle@univ-lr.fr (D. Sharma-Kafle); ibrahim.souleiman_mahamoud@univ-lr.fr (I. S. Mahamoud); aurelie.joseph@getyooz.com (A. Joseph); mickael.coustaty@univ-lr.fr (M. Coustaty); vincent.poulaindandecy@getyooz.com (V. P. d'Andecy)
0009-0008-5266-8797 (E. Thomas); 0009-0006-5643-0773 (D. Sharma-Kafle); 0009-0008-2037-4364 (I. S. Mahamoud); 0000-0002-5499-6355 (A. Joseph); 0000-0002-0123-439X (M. Coustaty)

2. Jeux de données

Pour évaluer notre approche d'extraction de paires clé-valeur, nous avons utilisé les jeux de données FUNSD et XFUND [11, 12]. FUNSD, centré sur les formulaires, partage des paires clés-valeurs pertinentes avec les factures, tandis que XFUND couvre sept langues pour démontrer l'adaptabilité linguistique du modèle. En complément, un jeu de données privé de 20 000 factures non annotées a été utilisé à des fins d'inférence, permettant une évaluation qualitative des résultats.

3. Méthodologie

Notre méthode traite efficacement les documents administratifs multilingues avec une faible dépendance à l'égard de la sémantique, la linguistique et la mise en page. L'accent est mis sur l'efficacité computationnelle, avec un temps d'exécution ciblé inférieur à 1 seconde par page

La méthode d'extraction se compose de trois étapes clés : **regroupement** selon des règles métier, **reconnaissance de la sémantique des entités (SER)**, et **extraction de relations (RE)**. La première étape utilise la mise en page pour créer des groupes de mots pertinents. La deuxième étape classe chaque groupe dans des catégories telles que clé, valeur, en-tête ou autre, grâce à un modèle BERT multilingue [2]. Enfin, la troisième étape évalue les relations entre les entités prédites, assurant une correspondance unique entre chaque question et réponse, renforçant la robustesse de l'approche dans la gestion des nuances des documents administratifs.

4. Résultats

Les résultats de nos expériences sur les trois étapes clés (regroupement des mots, reconnaissance de la sémantique des entités et extraction de relations) révèlent une performance prometteuse. La première étape, le **regroupement**, affiche une F1-mesure moyenne de 71%, soulignant ainsi son caractère novateur dans un contexte où, à notre connaissance, aucune méthode existante n'a fourni de résultats chiffrés sur cette tâche spécifique. Pour la deuxième étape **(SER)**, la F1-mesure moyenne de 79% surpasse les scores des modèles XLM-Roberta et InfoXLM, tout en rivalisant avec LayoutXLM. En ce qui concerne la troisième étape **(RE)**, la F1-mesure moyenne de 58% varie fortement selon la langue. Bien que généralement meilleurs que XLM-Roberta et InfoXLM, la variance plus prononcée par langue offre des opportunités d'amélioration significatives.

5. Conclusion

En conclusion, notre approche combine avec succès des méthodes basées sur des règles et des Transformers pré-entraînés pour l'extraction de couples clé-valeur dans les documents administratifs. Bien que des limites subsistent, notamment dans la prise en compte d'informations sémantiques et contextuelles, notre méthode offre une base solide pour des améliorations futures. Ces pistes d'amélioration pourraient inclure une adaptation plus poussée à la diversité des langues et des mises en page, ainsi que l'exploration de nouvelles architectures de modèles (modèles de graphes, modèles multimodaux, ...) pour des performances encore accrues.

References

[1] E. Thomas, D. S. Kafle, I. S. Mahamoud, A. Joseph, M. Coustaty, V. Poulain d'Andecy, Extracting key-value pairs in business documents, in: International Conference on Document Analysis and Recognition, Springer, 2023, pp. 32–46.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[4] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, 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 amp; Data Mining, KDD '20, ACM, 2020. URL: http://dx.doi.org/10.1145/3394486.3403172. doi:10.1145/3394486.3403172.

[5] T. Hong, D. Kim, M. Ji, W. Hwang, D. Nam, S. Park, Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents, 2022. arXiv:2108.04539.

[6] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, M. Zhou, InfoXLM: An information-theoretic framework for cross-lingual language model pre-training, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3576–3588. URL: https://aclanthology.org/2021.naacl-main.280. doi:10.18653/v1/2021.naacl-main.280.

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[8] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, L. Zhou, Layoutlmv2: Multi-modal pre-training for visually-rich document understanding, 2022. arXiv:2012.14740.

[9] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022. arXiv:2204.08387.

[10] C. Luo, C. Cheng, Q. Zheng, C. Yao, Geolayoutlm: Geometric pre-training for visual information extraction, 2023. arXiv:2304.10759.

[11] G. Jaume, H. Kemal 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, 2019, pp. 1–6. doi:10.1109/ICDARW.2019.10029.

[12] Y. Xu, T. Lv, L. Cui, G. Wang, Y. Lu, D. Florencio, C. Zhang, F. Wei, Xfund: A benchmark dataset for multilingual visually rich form understanding, 2022, pp. 3214–3224. doi:10.18653/v1/2022.findings-acl.253.