LIÈGE université Library

https://lib.uliege.be

https://matheo.uliege.be

Master thesis : Invoice Entity Recognition

Auteur : Hubar, Julien
Promoteur(s) : Louppe, Gilles
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/14584

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

University of Liège
Master Thesis

Invoice Entity Recognition

Author: Hubar Julien
Supervisors: Prof. Gilles Louppe (ULiège)
Rosu Lionel (Billy)

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Data Science. Academic year 2021-2022

Abstract

Today, for Billy and many accounting fiduciaries, invoice information is usually encoded manually by the accountant or by a low-performance software, so a lot of time is wasted on encoding and not on advice. Indeed, During the year 2021, Billy's accountants spent 37% of their time on encoding. Consequently, the recognition of fields in a semi-structured document of variable layout (i.e. invoice) is a growing need for accountants, and especially for Billy, in which the number of new customers increasing every month. Nonetheless, the text and image pre-training strategies of the Transformer architecture model have proven to be efficient in the field of document understanding. Thus, several OCR tools were tested, and the *Azure* OCR tool, which gave the higher performances, was selected to extract text from image invoice of Billy's customers in order to create datasets. Indeed, this allows the elaboration of four datasets partially annotated, named **BTT**, **BTT Star**, **BTT QV**, and **BTT QV Date**, which were created from scanned purchase, sales documents, and their accounting encoding in the accounting Horus software. Then, the fine-tuning of the pre-trained multi-modal models $LayoutLMv2_{BASE}$, $LayoutLMv2_{LARGE}$, and $LayoutXLM_{BASE}$ has been done. In contrast to the previous architectural models of the *LayoutLM* family, these models include, in addition to the text and layout information, information that can be provided by the document image. Thanks to spatial-aware self-attention mechanisms integrated in the Transformer architecture model, it is able to interpret relations through different bounding boxes. According to Billy's accountants, invoice information is recognized by Horus in 70% of the cases. During the experiments conducted in this Master thesis, it was shown that on token classification tasks, higher results were obtained for the the different datasets in terms of F1-score: **BTT** (0.9420), **BTT Star** (0.9553), **BTT QV**(0.9413) and **BTT QV Date** (0.9472). In addition, similar state-of-the-art results were obtained using the open source CORD dataset which gives an F1-score of 0.9354. Moreover, the impact of a pre-trained model on a dataset composed only of English documents ($LayoutXLM_{BASE}$) was studied in comparison with a pre-trained model on a multi-lingual dataset ($LayoutLMv2_{BASE}$) to classify tokens on documents mostly in French. The results show that the pre-trained model does not have a great impact on the final result for this type of task (**BTT** (0.9323 $\rightarrow$ 0.9338), **BTT Star** (0.9354 $\rightarrow$ 0.9468), **BTT QV** (0.9229 $\leftarrow$ 0.8955), and **BTT QV Date** (0.9411 $\leftarrow$ 0.9328). To conclude, since the results of the four datasets were close to each other, the dataset **BTT Star** produced the best results. This dataset has the largest number of labels and is the most widely distributed over the documents, leading to the hypothesis that a more widely distributed set of labels provides better results. Finally, to concretize this work, a web application was developed in parallel in order to use this tool in everyday life for both the accountants and the customers.