Transformers based Entity Linking in Documents

Information extraction from rich unstructured documents is a complex but key process in various businesses. Within BNP Paribas CIB, Analytics Consulting Datalab has been developing a platform to automatically label documents with NLP models and retrieve user feedback. More than 20 businesses use this platform.

Based on the current state of the art and internal research, a mature information extraction model has been implemented internally [1] [2].
Proper structuring of the information contained in the document requires **extracting the entities from the documents** (dates, company name, specific clauses, etc.) but also **finding the link between them**.

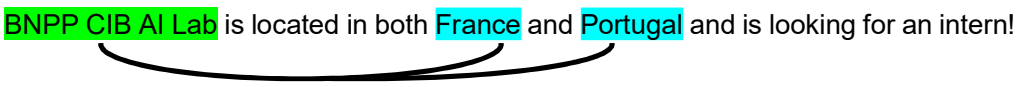

Figure: Illustration of the two tasks: entity extraction and entity linking

The internship will follow the roadmap below:
- Familiarize with the literature on information extraction from visually rich documents
- Implement and train the **neural relation extraction models** [3] [4]
- Extend current methods to **jointly train entity extraction and entity linking** [5] [6]
- Finally, we will explore how to improve the performance of the state of the art in the context where we have only a few positive examples (**few shot entity linking**) and when the entities to be linked are not spatially close in the document. .

The internship will allow the student to gain in expertise in various domain especially: document processing, advanced deep learning, computer vision, natural language processing.

References:

[1] Huang, Yupan, et al. "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking." arXiv preprint arXiv:2204.08387 (2022).
[2] Oussaid, Ismail, et al. "Information Extraction from Visually Rich Documents with Font Style Embeddings." arXiv preprint arXiv:2111.04045 (2021).
[3] Shi, Peng, and Jimmy Lin. "Simple bert models for relation extraction and semantic role labeling." arXiv preprint arXiv:1904.05255 (2019).
[4] Han, Xu, et al. "OpenNRE: An open and extensible toolkit for neural relation extraction." arXiv preprint arXiv:1909.13078 (2019).
[5] Li, Yulin, et al. "StrucTexT: Structured text understanding with multi-modal Transformers." Proceedings of the 29th ACM International Conference on Multimedia. 2021.
[6] Wang, Jiapeng, Lianwen Jin, and Kai Ding. "Lilt: A simple yet effective language-independent layout transformer for structured document understanding." arXiv preprint arXiv:2202.13669 (2022).
Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, Jiawei Han

Practical Information:

Supervisor: RATNAMOGAN Pirashanth, Data Scientist, pirashanth.ratnamogan@bnpparibas.com

Location: Grands Moulins de Pantin, 9 Rue du Débarcadère, 93500 Pantin (Possible partial work from home)
Compensation: Competitive salary
How to apply: Please submit your application at paris.cib.analytics.consulting.careers@bnpparibas.com