Machine Learning for Data Extraction

12/25/20 8:02 PM

Speaker: Chester Curme (Kensho Technologies, S&P Global)
Date: 2020/10/8

Company overview

- Infrastructure to store and organize data
  - Kensho knowledge graph
- Tools to navigate data
  - Search
  - E.g. S&P market intelligence platform, improve searching algorithms
- Tools to acquire new data
  - Link data
- Tools to analyze data and extract information
  - NLP services

Data Extraction

Investors need structured data

Take unstructured PDF files, extract data and make into tabular data

Automation is crucial

- Tens of thousands of public companies globally, each issuing quarterly and annual reports
- Hundreds of financial line items
- Latency can make or break an investing operation
  - E.g. parsing earning releases
  - Investors want to have it as soon as possible
- Unstructured document formats (PDF) still common globally
  - U.S. largely HTML

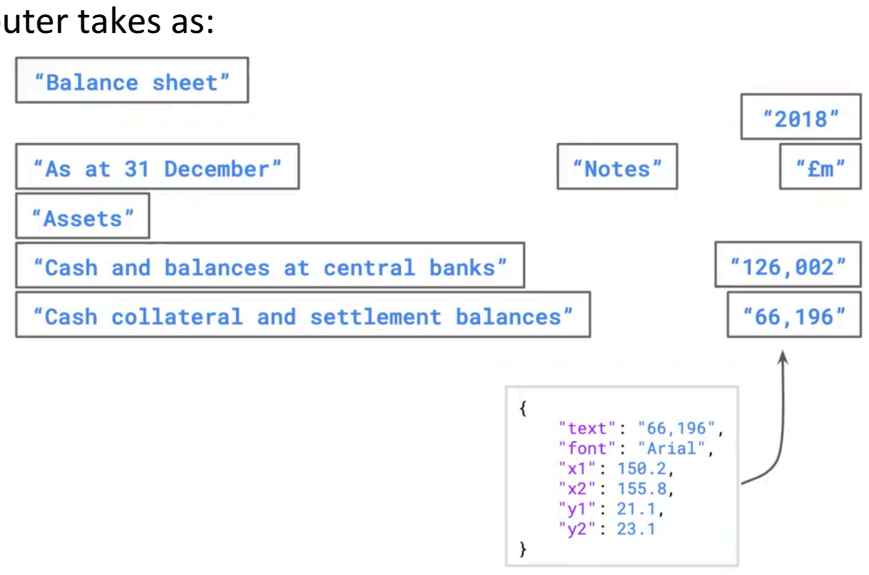
What is so difficult?

PDF:

Balance sheet

As at 31 December	Notes	2018 £m
Assets		
Cash and balances at central banks		126,002
Cash collateral and settlement balances		66,196

Computer takes as:



Area where machine learning is adding value

- CNN for table extraction (bonding box)
- AxCell (FB research, DeepMind): https://arxiv.org/abs/2004.14356
- Table Detection (Graph NN): https://priba.github.io/assets/publi/conf/2019_ICDAR_PRiba.pdf
- TAPAS Information Retrieval (Google Research): https://arxiv.org/abs/2004.02349
  - Question answering models
  - GPT3