23384: How AI Revolutionizes Data & Document Management

LEVERTON
Your Data. Your Value.

Florian Kuhlmann (CTO & Co-founder)
LEVERTON - Company overview

- **100+** Global Corporate Clients
- **20+** Languages supported
- **100%** Data Transparency
- **75+** Employees in offices in Berlin, London and New York
- **4+** Years using Deep Learning Technology
- **Highest** Data Security – ISO 27001 / 9001
“We enable smarter decisions by structuring the world’s data”
From documents to data
Which problem does LEVERTON solve?

Manual data input
- Inefficient
  - Use AI to classify documents and extract data
- Non-transparent
  - Keep link to position in document
- Error prone
  - Control using efficient 4-eyes principal
Data input with LEVERTON
Three steps from document to data

- Upload scanned PDFs
- Automated extraction
- Review and correct

OCR – convert images into machine readable text

Document Classification – categorize documents

Information Extraction – extract relevant data

DATA CORE

Train

Deep Learning
Information Extraction

Automatically extract structured information from un- or semi-structured machine-readable documents
(Wikipedia)

1. Define the **type of information** (e.g. as part of a **data model**)  
2. Find **location** (start + end, coordinates) of information **in document**  
3. **Extract** the information (parse numbers, dates, …)
Information Extraction: Define Types

Data model as abstraction of real world entities

- Example data model for rent charges of a lease -
From 1 October 2006 to 31 December 2010: The rent (inclusive of management fees and air-conditioning charges during normal office hours) for the period from the Commencement Date to the Expiration Date of the said term shall be USD ONE THOUSAND EIGHT HUNDRED AND THIRTY FOUR (USD 1,834.00) per calendar month.
Information Extraction

**Rule based**

Extraction based on handwritten rules

1. Apple is bad, if it has brown spots
2. Apple is bad, if it has wrinkles
3. Apple is bad, if it is brown
4. ...

**Machine & Deep Learning**

Learn from examples

- Bad apple
- Bad apple
- Good apple
Rule based Information Extraction

Example: Extract phone or fax number from text

Rule: Phone1

```plaintext
((Token.string=="Telefon")
 | ({Token.string=="Tel"}(Token.string==".")?)
 | ({Token.string=="Fax"}))
(Token.kind==punctuation)?
(Token.string=="(")?
(Token.numtype==ordinal)[1,3]
(Token.string==")")|{Token.string=="/"})?
(Token.numtype==ordinal)[1,6]
(Token.string=="-"{Token.numtype==ordinal})?

:contact

--> 

:contact.Phone = {rule = "Phone"}
```
Information Extraction with Deep Learning
Feed annotated examples and wait
Information Extraction
Rule based vs. Deep Learning

Rule based

PROS
• Start without labeled data
• Can be debugged and adjusted easily

CONS
• Rule engineers need domain knowledge
• Rule engineers need language knowledge
• Rules must be adopted to different languages
• Rules can become quite complex and hard to understand
• Rules never cover all possible cases, esp. in complex languages

Machine & Deep Learning

PROS
• Same algorithms for all languages / domains
• No need for engineers to understand the language
• No need for engineers to understand the domain
• Robust, generalizes to different verbalizations
• Fast

CONS
• Labeled data needed
• Cannot be debugged
• Hard to understand and to tune
• Training might take long time...

... So what is the better choice?
Rules vs. Deep Learning @ LEVERTON

What to use?

• Approx. one person year into writing **rules** led to performance (F1 score) of 20%-65% on 20 different data points in 1 language.

• Current **Deep Learning** led to performance (F1 score) of 70% to near 100% on 700+ different data points in 4 languages.

→ This sounds good, but how can we get better?
Challenges: Layout

Different layouts might need different extraction strategies

Structured

Semi structured

Unstructured
Challenge in Layout:
Determine the correct reading order

- Visual features alone often not sufficient!
- Impossible to determine the reading order without reading the text. We need to guess.
Determine the correct reading order

- Combination of visual features and interpretation of extracted text leads to correct reading order
- Human brains are able to combine multiple steps (Visual separation, recognizing of characters and words, interpreting words to form sentences, which leads to correct separation / layout recognition)
- Deep Learning is focused on one problem at a time
- Joint approaches necessary to solve such problems
Challenges: Data model

Abstraction comes with a price

surplus earth PROVIDED that in so doing the Company shall cause as little damage as may be to such piece of land and shall make good and restore the surface thereof TO HOLD the same unto the Company for the term of ninety-nine years from and including the date hereof YIELDING AND PAYING therefor YEARLY during the said term in advance the rent of one peppercorn (if demanded) on each anniversary of the date of commencement of the term created by this Lease.
More challenges ahead...

How can we improve in the future

• Co-Referencing (which information belongs to what)

• Scan quality

• **Scaling** out: Training already takes 6 weeks of one core

• ...
... But we are on a good track
Achievements of LEVERTON AI

- **Simple documents:** >95% automation for >50 different data points
- **Complex documents:** >70% automation for >200 different data points
- Available in >20 different languages
- Build **own OCR engine** in < 1 year, competitive to all known OCR engines, more robust on layout
Example Use Case: IFRS16
Changes in balancing standards forces corporates to revisit leasing data

- Global consolidation
- Extract data
  - IFRS 16 DATA
    - Options
    - Valuation
    - Indexation
    - Payment types
- System Integration
  - LEVERTON can be integrated into ERP systems at customers

Real estate leases
Machinery leases
Car leases & more

LEVERTON
LEVERTON CORE
Example complex document >70% automation
LEVERTON CORE

Example simple document, >95% automation
LEVERTON CORE

Enables decisions using data based on legally binding documents
Recap

*How AI Revolutionizes Data & Document Management*

- Breakthrough with switch to **Deep Learning**
- More **challenges** for **100% automated** information extraction on **complex docs**
- To achieve this, we probably need to **combine multiple steps** into one large network
- We’ll need lots of **computational resources** to do so
- If you have ideas about any of these: **We are hiring.**
LEVERTON

Florian Kuhlmann  
CTO & Co-founder  
florian.kuhlmann@leverton.ai