Information extraction from contracts

Thi Hien Ha

May 2020

Example

Krajská zdravotní, a.s.

Smlouva o poskytování služeb

*uzavřená dle § 1746 a násl. zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisů a na základě veřejné zakázky s názvem „**Dodávka a instalace aplikace pro využití revizí a pasportizace**“*

mezi

TESCO SW a.s.
se sídlem: tř. Kosmonautů 1288/1, Hodolany, Olomouc, PSČ 779 00
IČO: 25892533
DIČ: CZ699000785
zastoupená: RNDr. Josefem Tesaříkem, předsedou představenstva
zapsána v obchodním rejstříku vedeném Krajským soudem v Ostravě, oddíl B, vložka 2530
[redacted]
(dále jako „***poskytovatel***“)

a

Krajská zdravotní, a.s.
se sídlem: Sociální péče 3316/12A, Ústí nad Labem, PSČ 401 13
IČO: 25488627
DIČ: CZ25488627
zastoupená: Ing. Petrem Fialou, generálním ředitelem společnosti na základě pověření představenstvem společnosti ze dne 17. 12. 2015
zapsána v obchodním rejstříku vedeném Krajským soudem v Ústí nad Labem, oddíl B, vložka 1550
[redacted]
(dále jako „***uživatel***“)

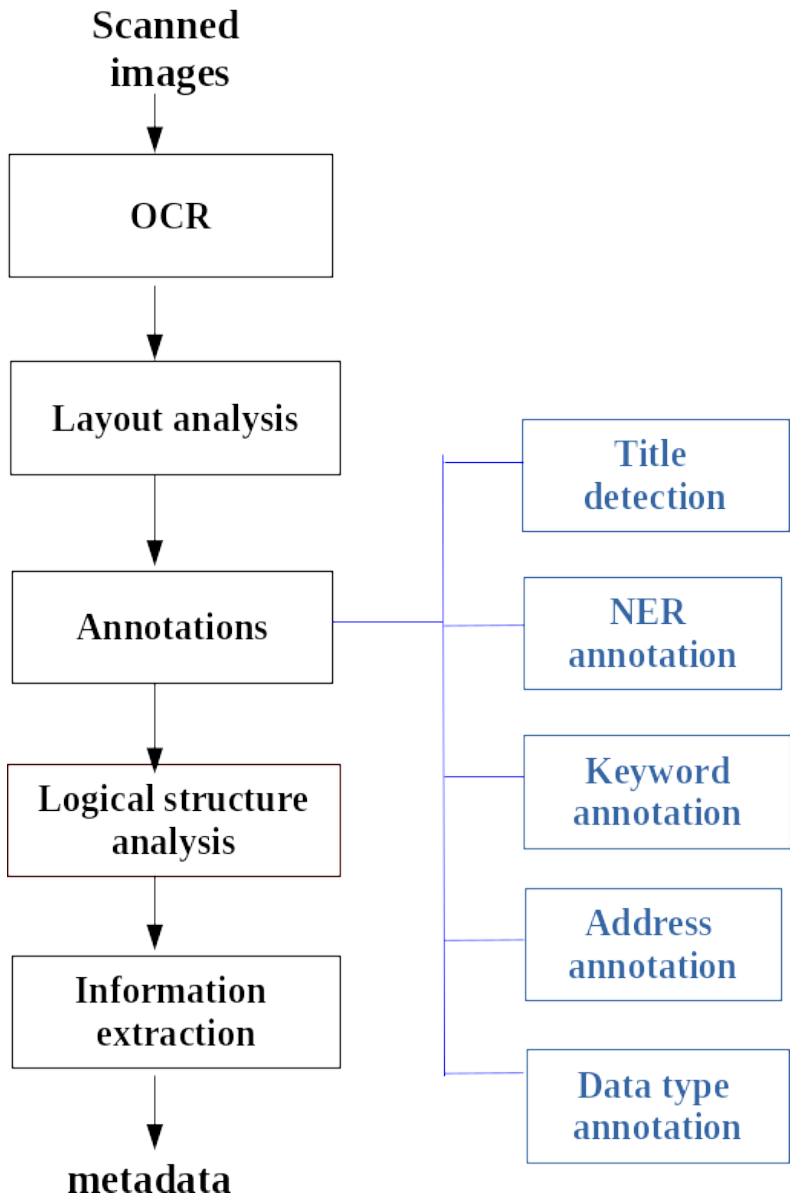
tuto
smlouvu
(dále jen „smlouva“)

Poskytovatel a uživatel jsou dále označeni rovněž jako „**smluvní strana**“ či společně jako „**smluvní**

Fields to be extracted

- General info: contract number, contract date, legislation
- Parties: ORG, address, company id, vat number, representative, bank info, role

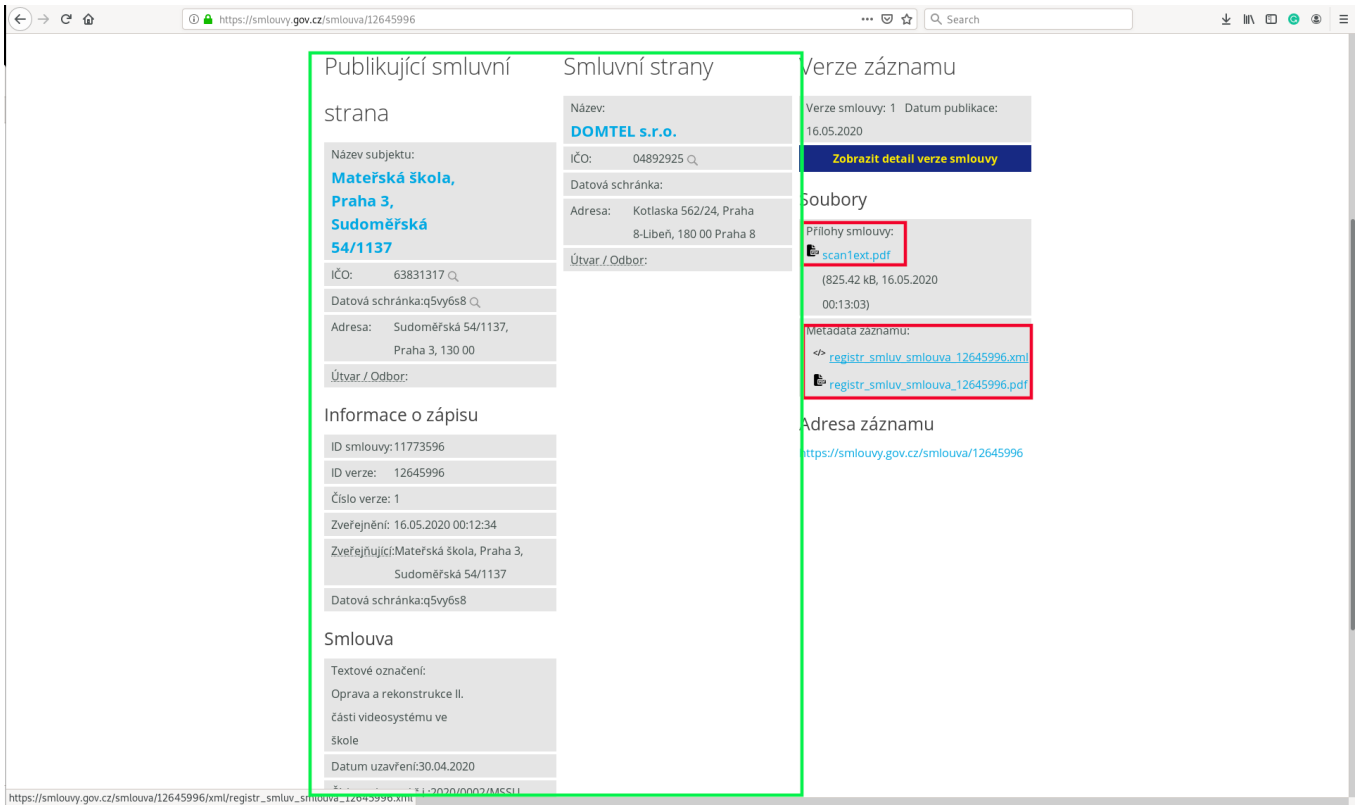
Pipeline


Scanned images
OCR
Layout analysis
Annotations
Logical structure analysis
Information extraction
metadata
Title detection
NER annotation
Keyword annotation
Address annotation
Data type annotation

what is new in the annotation

- Process all pages of the contract
- Title: based on 3 criteria: font size, alignment and keyword
- NER: using Slavic-Bert-Ner: https://github.com/deepmipt/Slavic-BERT-NER

Data source: https://smlouvy.gov.cz



Data preparation

Prepare training set and test set

- Down load data from website
- Apply 2-step filters on downloaded data
- Randomly select 60 files for the experiment, than random 10 of 60 is used as training set and 50 others as test set.

Create GT from metadata

- Read html files to map each contract with metadata file
- convert metadata file into desired format

Problem with the metadata

- Info is not in the pdf file but in the metadata
- Info in different formats, e.g: date, address
- Lack of info: VAT, representative, role, bank info, legislation

Parties' information

- All potential parties are extracted
- The more info, the higher confidence.
- Evaluation: Find the party which has highest number of common fields and compare with that party only.

Result on the training set

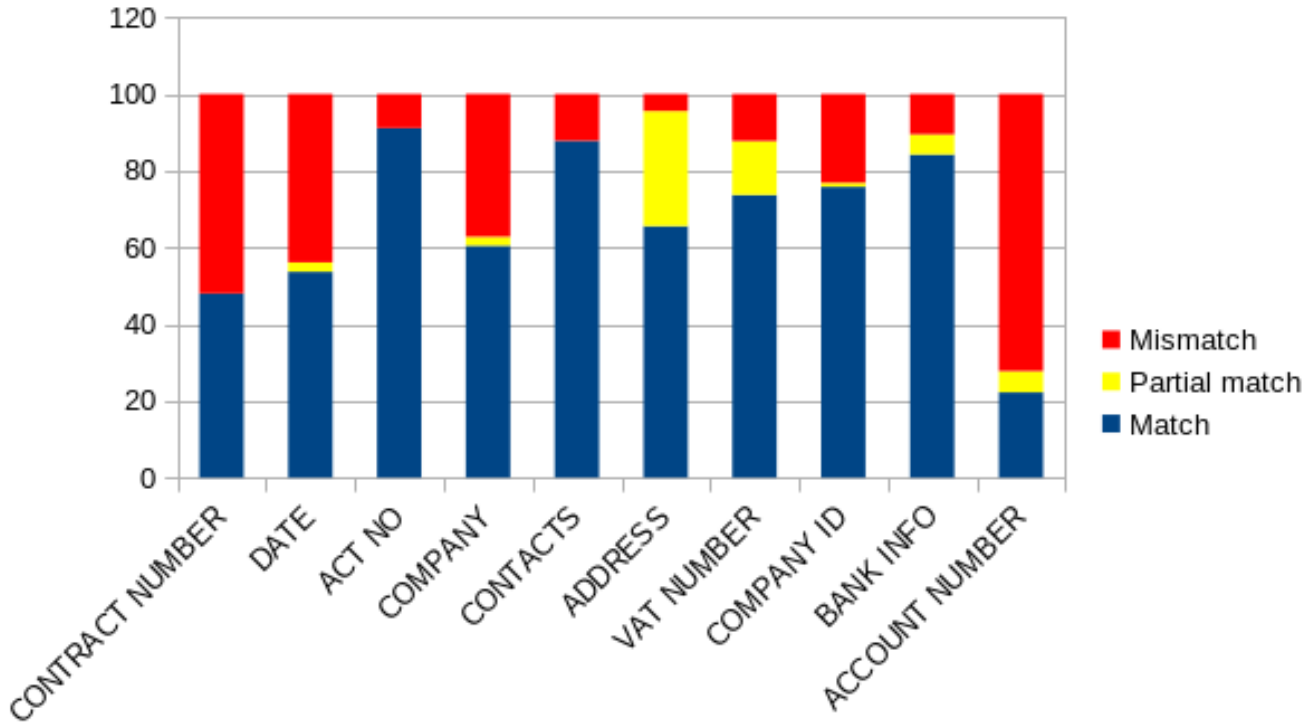
Match	104	83.87%
Patital match	8	6.45%
Mismatch	12	9.68%
Total	124	100%

Mismatch analysis:

- ORGs are marked as LOCs by both NER and libpostal (4). E.g: "HLAVNÍ MĚSTO PRAHA"
- OCR errors: e.g dates are hand-written (2)
- info is spread more than 1 line, e.g law (1)
- Party's info is in different blocks (role: 4, bank: 1)

Result on the test set

Match	433	66.31%
Patital match	41	6.28%
Mismatch	179	27.41%
Total	653	100%



Error analysis on the test set

- OCR errors: hand-written dates, low quality images, covered areas
- Keywords are not appeared in the training set, e.g "Smlouva č." (8/13 missing contract number), 13/13 missing account number (Bankovní spojen, č.ú)
- ORG and LOC mis-classified
- Party info in different blocks