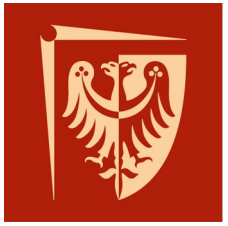
Results of the PolEval 2020 Shared Task 4

Information extraction and entity typing from long documents with complex layouts

Michał Marcińczuk, Michał Olek, Marcin Oleksy, Jan Wieczorek

Department of Computational Intelligence
Faculty of Computer Science and Management
Wroclaw University of Science and Techology
michal.marcinczuk@pwr.edu.pl
michal.olek@pwr.edu.pl



Politechnika
Wrocławska

Task 4 description

The goal of the task is to extract values for a defined set of fields from documents with complex layouts.

The documents' layouts are complex - the documents can contain many (>100) pages and multiple text sections, tables, plots, forms, etc.

Fields to extract:

- company
- drawing_date
- period_from
- period_to
- postal_code
- city
- street
- street_no
- people :
  { name, role, sign }

Dataset

Three datasets:

- training – 1662 documents
- dev – 554 documents
- test – 555 documents

Three formats:

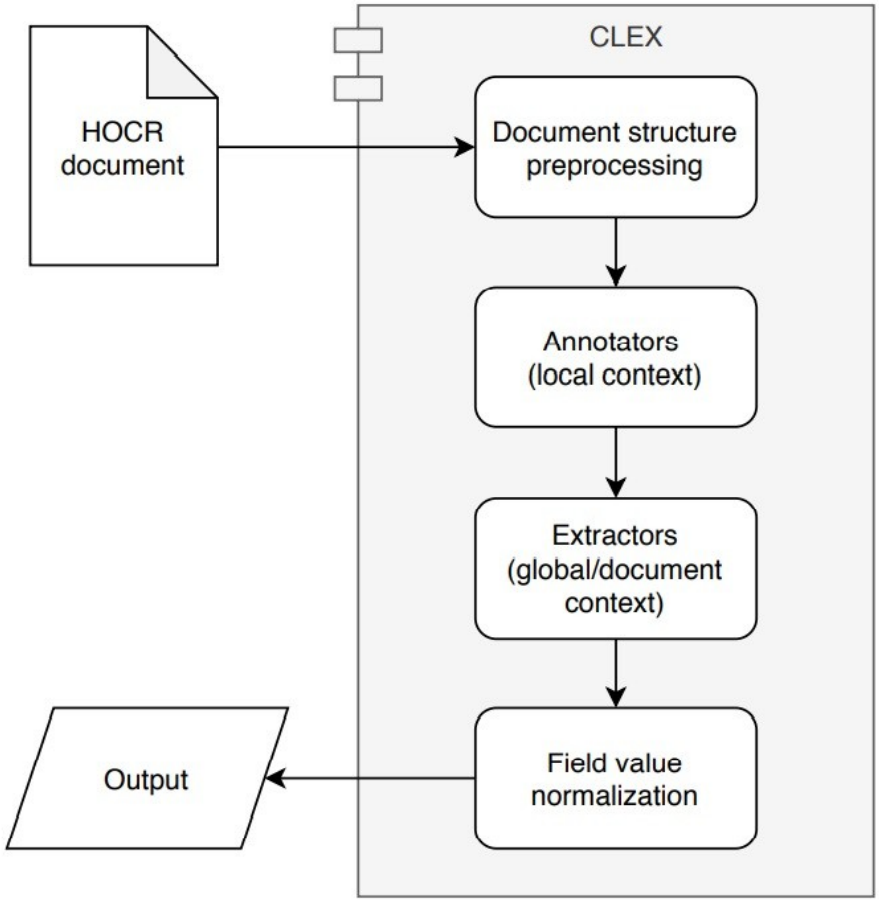
- pdf – original
- txt – only text
- HOCR - text with spatial information

Resources

- company names in base form
  - extracted from training and dev sets
- company addresses
  - extracted from training and dev sets
- person names
  - taken from NELexicon2 [1]
- city names
  - taken from Polish TERYT database

[1] Marcinczuk, M.: NELexicon2 (2014), http://hdl.handle.net/11321/247, CLARINPL digital repository

CLEX – system architecture



Document preprocessing

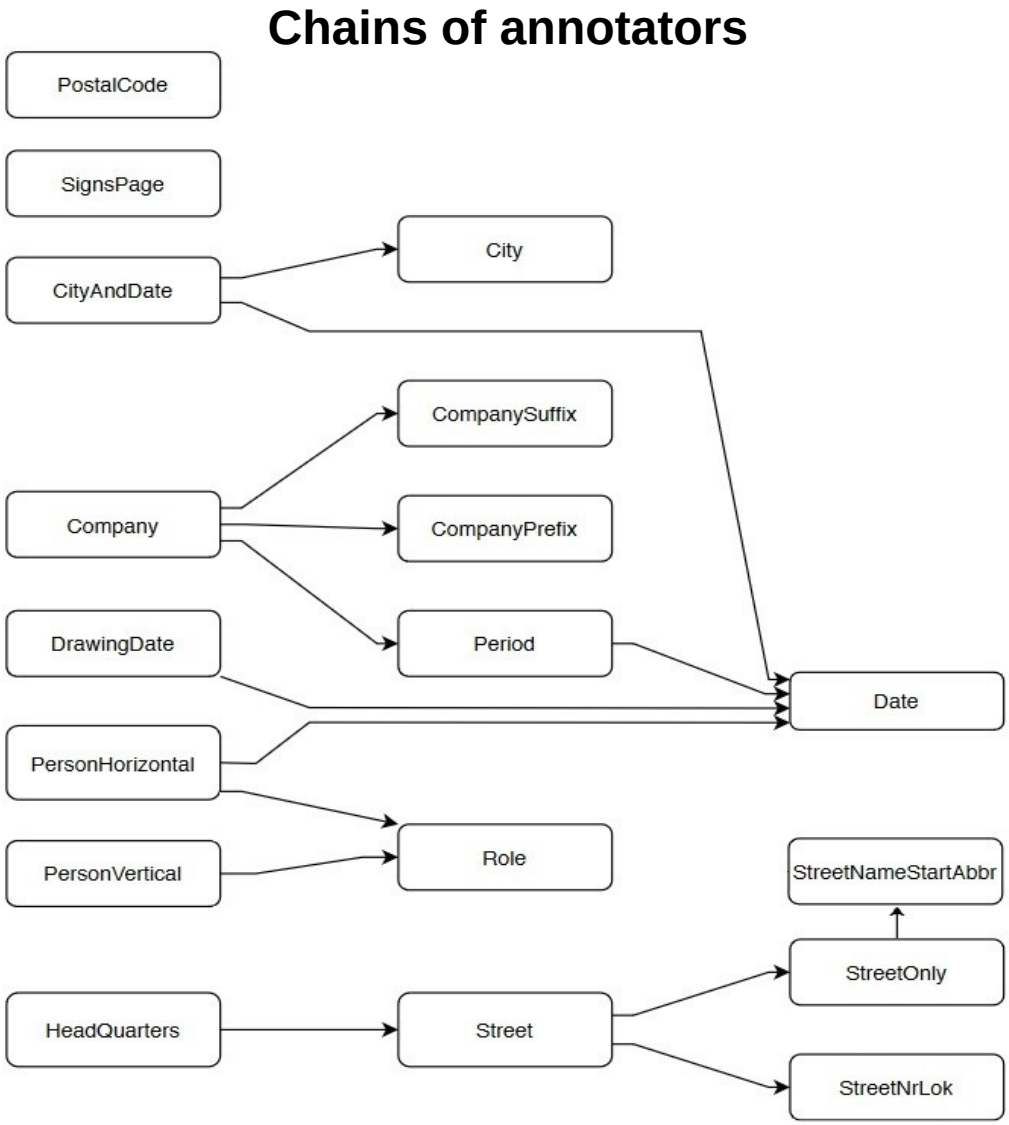
- Restoring order
- Separation of punctuation marks
- Handling empty leading pages
- Calculating line heights and its distribution
- Separating headers and footers from document

Annotators - local context

At the lowest level, we have a set of *matchers* that match the given text or the given regular expression

The *matchers* are combined to define more elaborated *patterns*.

Annotators use the *patterns* to create rules that allow extracting interesting fragments of the text



Extractors - global view

Extractors - try to select the best result from the ones found and stored so far.

Extractors have a defined behavior depending on the type of field they are trying to find

One special : finds the page with signatures

For some fields additional lematization and normalization is performed:

- names of companies
- names of streets

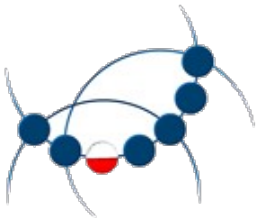
Results

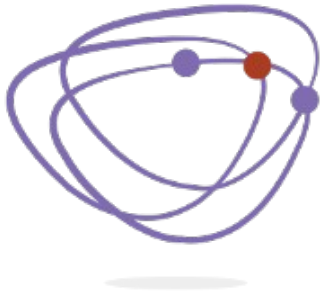
Evaluation results on the test set

Submission	F1(UC) on test
CLEX	**0.651** ±0.019
double_big	0.606 ±0.017
300_xgb	0.592 ±0.015
double_small	0.588 ±0.018
300_RF	0.587 ±0.015
middle_big	0.585 ±0.016
100_RF	0.584 ±0.016
Multilingual BERT + Random Forest	0.440 ±0.014

Thank you for your attention

CLARIN-PL
Common Language Resources and Technology Infrastructure





CENTRUM TECHNOLOGII JĘZYKOWYCH CLARIN-PL