Bootstrapping Multilingual Metadata Extraction: *A Showcase in Cyrillic*

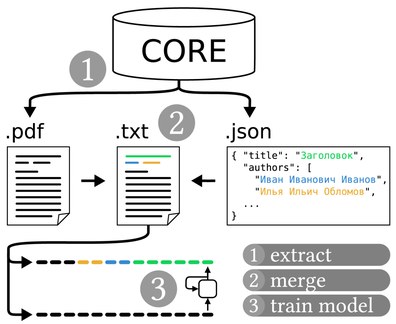
PRESENTER:
Johan Krause

Motivation
Lack of NLP applications for non-English content leads to underrepresentation in data.

Goal
Create a data set of Cyrillic script scholarly documents. Use it to re-train an existing tool and a custom neural network for metadata extraction of author and title labels.

APPROACH

1. Collection of Cyrillic script documents from the CORE data set
2. Selection of data and pre-processing, resulting in ~15,000 documents
3. Re-training the widely used GROBID tool
4. Training a BiLSTM network for comparison



RESULTS

Model	Precision	Recall	F1
GROBID$_{vanilla}$	0.06	0.06	0.06
GROBID$_{re\text{-}trained}$	**0.85**	0.81	0.83
BiLSTM	0.84	**0.96**	**0.90**

Using our high-quality data set of **Cyrillic script** scholarly documents greatly improved performance in **metadata extraction**.

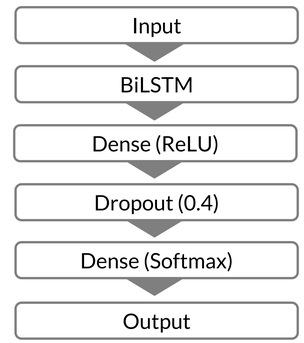
Contributions

- Showcase of effective creation of high-quality data for training and evaluating metadata extraction models
- Creation of a data set comprising 15k documents
- Creation of a sequence labeling model that outperforms available methods

Detailed Model Scores

Model	Precision	Recall	F1
GROBID$_{r.\|title}$	**0.90**	0.90	0.90
BiLSTM$_{title}$	0.88	**0.96**	**0.92**
GROBID$_{r.\|author}$	**0.81**	0.74	0.77
BiLSTM$_{author}$	0.80	**0.95**	**0.87**

BiLSTM Model Structure



Johan Krause, Igor Shapiro, Tarek Saier, Michael Färber

Karlsruhe Institute of Technology