Estonian Named Entity Recognition: New Datasets and Models

Kairit Sirts, University of Tartu

23.05.2023, NodaLiDa 2023

Joint work with

- Laura-Katrin Leman
- Chenghan Chung
- Claudia Kittask

Previously in Estonian NER

Datasets

- Estonian NER dataset
  - Tkachenko et al (2013)
  - ca 240K tokens of news texts
  - Annotated with PER, ORG, LOC

Models

- CRF-based model (EstNLTK)
- EstBERT-based model (in Huggingface)

Current work

Datasets

- Main Estonian NER dataset:
  - Reannotation of the Estonian NER dataset
- New Estonian NER dataset:
  - ca 130K tokens of new texts
  - from news and social media domains
  - Estonian Web Corpus 2017

Models

- Based on EstBERT
- Trained on two datasets separately
- Trained on the merged dataset

Annotation scheme

- Person names
- Organization names
- Locations
- Geopolitical Entities (GPE)

- Dates
- Times
- Monetary values
- Percentages

- Events
- Products
- Titles

Geopolitical entities

- Any geographical location related to a political entity
- Used to ease the annotation between ORG and LOC

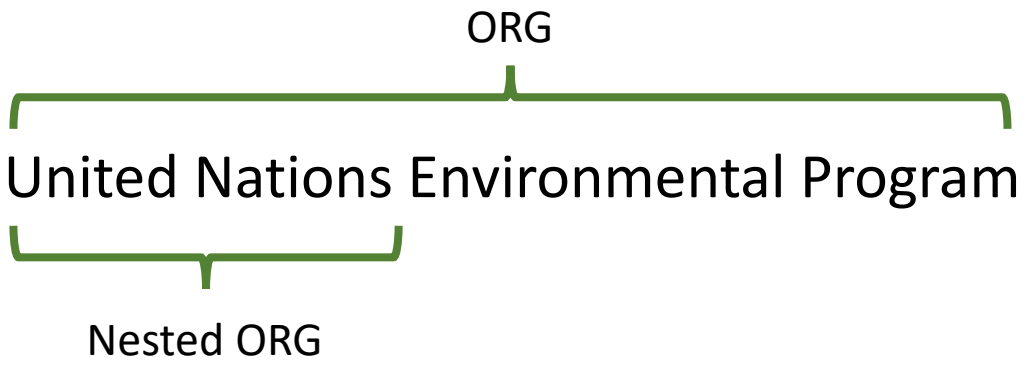
Examples:

- The mayor of Tallinn (ORG aspect of a GPE)
- The President resides in Tallinn (LOC aspect of a GPE)

Nested entities

Addititionally annotated nested entities, up to three levels:

- only ORG can be inside ORG
- other nestings must contain different entity types



Annotation process

Main NER dataset

- Three annotations for each text
  - Three graduate students in linguistics
- Annotated with Label Studio

New NER dataset

- Three annotations for each text
  - One linguistics bachelor student
  - One CS masters student
  - 10 masters level NLP course participants
- Annotated with DataTurks (not available anymore)

Inter-annotator agreement

Main NER dataset

Cohen's kappa

Computed on entities

	1st level	2nd level	3rd level
	0.65	0.23	-0.16
PER	0.95	0.27	0.66
ORG	0.76	0.33	0.19
LOC	0.65	0.35	0.18
GPE	0.84	0.47	-0.08
TITLE	0.63	0.21	0.00
PROD	0.48	0.02	–
EVENT	0.43	0.53	–
DATE	0.72	0.06	–
TIME	0.53	0.00	–
MONEY	0.78	0.00	–
PERCENT	0.90	–	–

Final entity statistics

Entity	Main NER dataset	New NER dataset
PER	4927	3009
ORG	4260	1412
LOC	507	515
GPE	4570	1613
TITLE	1682	772
PROD	1101	801
EVENT	86	271
DATE	1840	887
TIME	525	115
PERCENT	257	87
MONEY	520	131

Modeling experiments

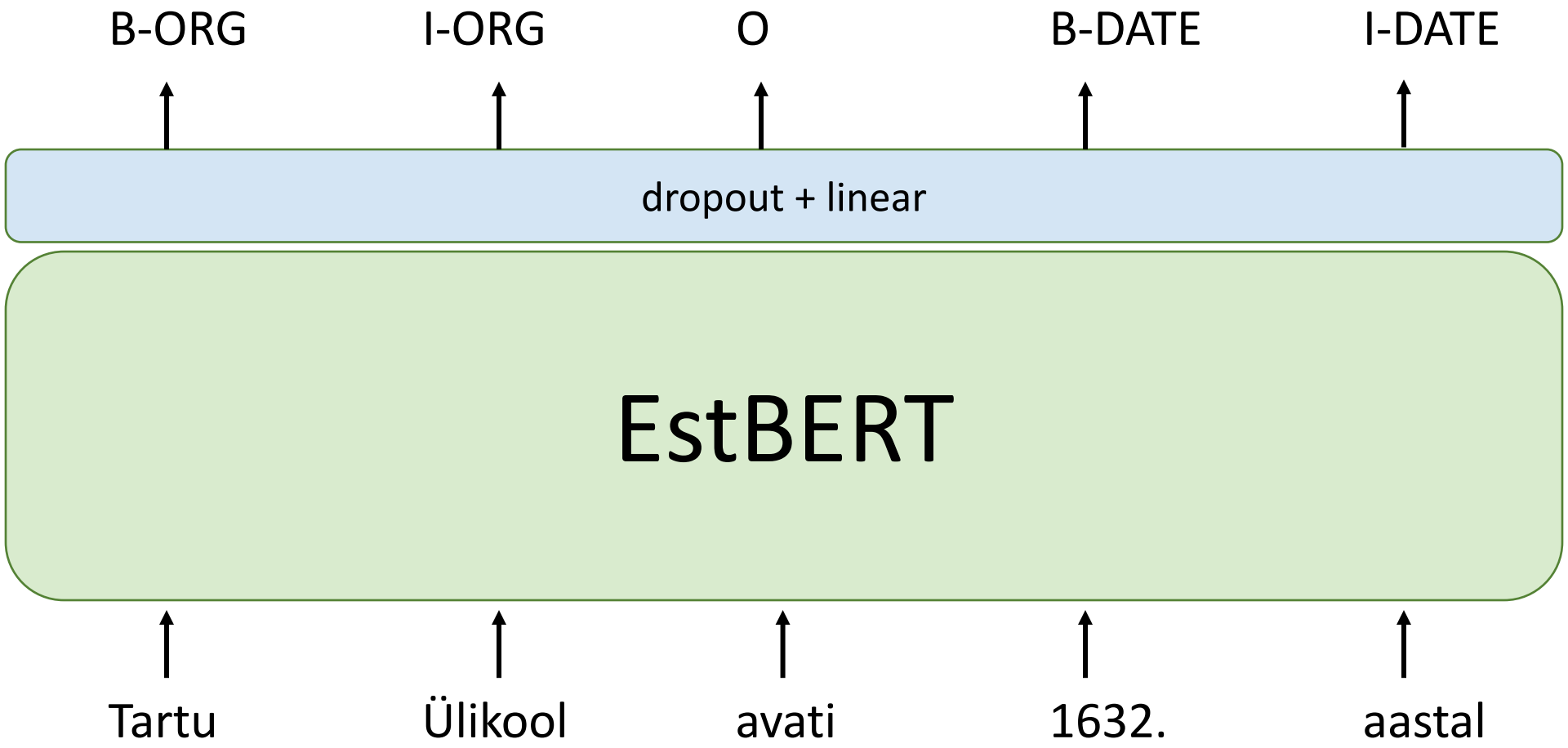
Experimental setup

- BIO tagging model based on EstBERT
- Learning rate tuned on the development set
- Ten models with different random seeds

Models

- Separate models trained on:
  - Main NER dataset
  - New NER dataset
- Joint model trained on merged dataset

Model



The University of Tartu was opened in 1632

Training setting

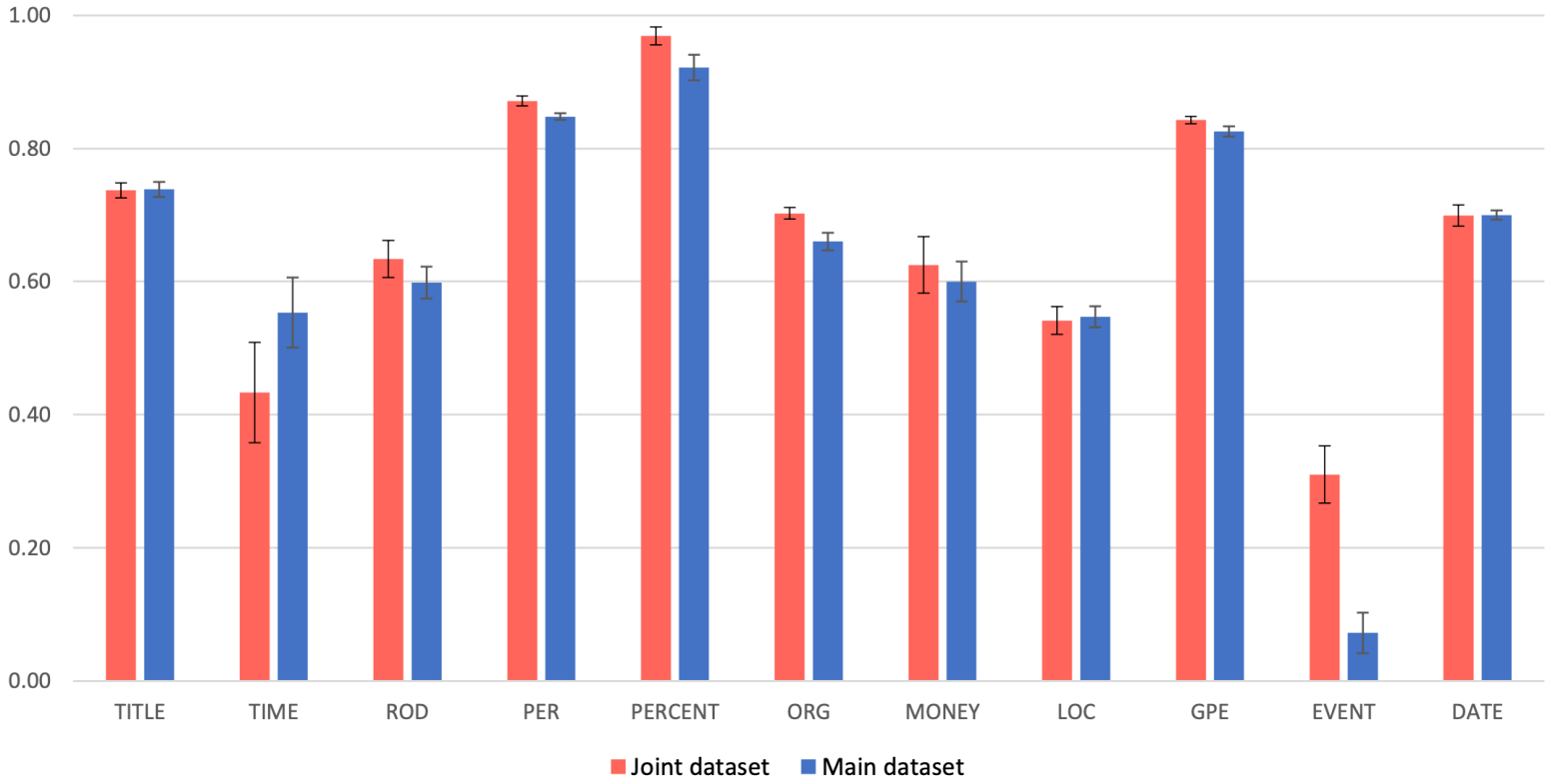
- Trained in a single GPU
- Batch size: 16
- AdamW optimizer
- Trained for max 150 epochs
- Early stopping of 20 epochs on validation set

F1-scores on development set

Evaluated on: / Model	Main NER validation set	New NER validation set	Merged validation set
Separate Main	0.747 (0.004)		
Separate New		0.735 (0.006)	
Joint	0.766 (0.002)	0.752 (0.010)	0.761 (0.004)

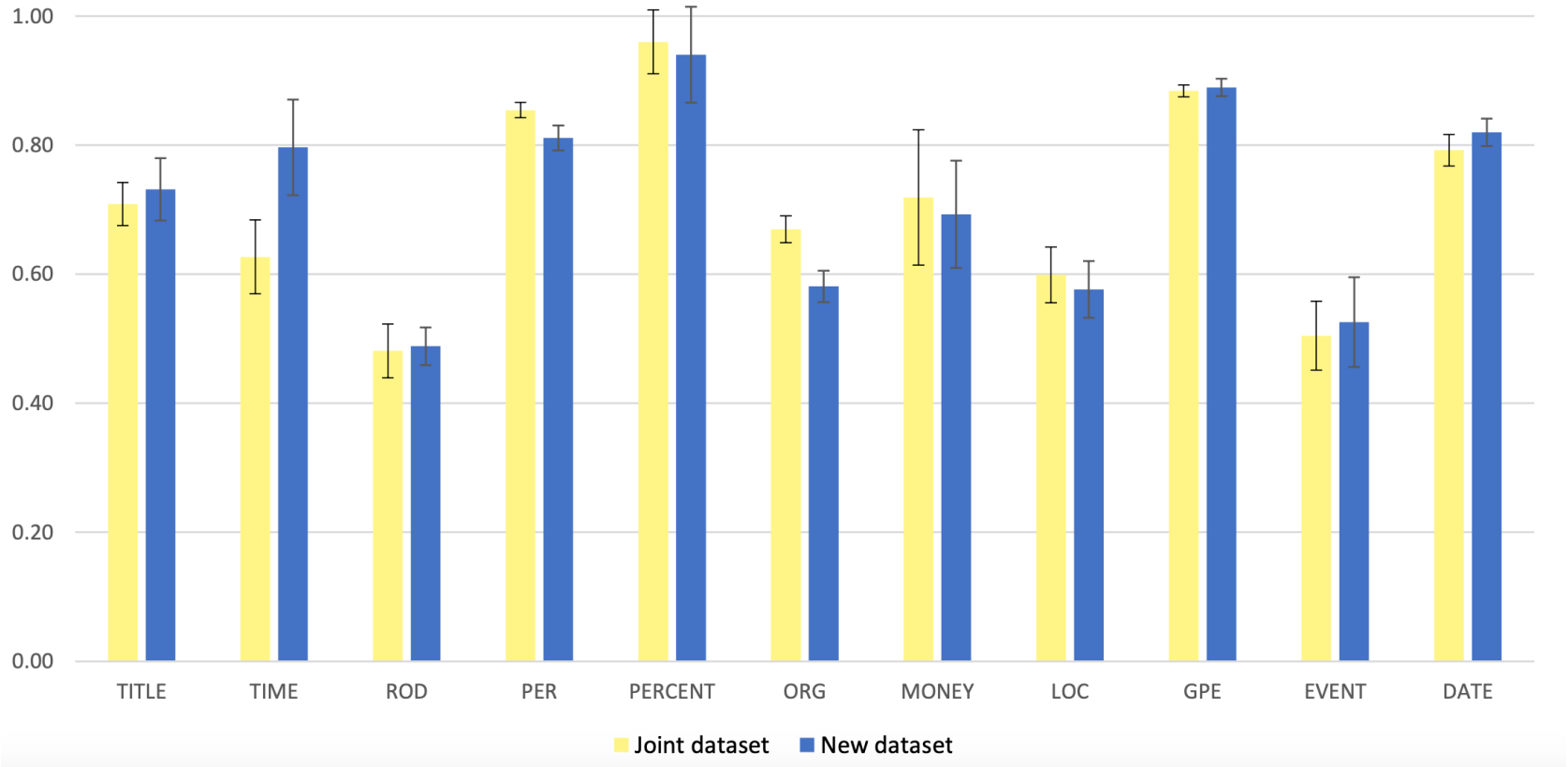
Main vs Joint model on the Main dev set

- Joint model better on most entities
- worse on TIME
- better on EVENT



New vs Joint model on the New dev set

- Joint model better on some entities
- Slightly lower on others
- considerably worse on TIME



Best model on the test set: F1-score

- Best model based on validation F1-score
- Comparison to model trained on the old Estonian NER dataset(Tanvir et al., 2021)
- Lower results on new annotations → New annotations more complex

	Joint test set	Old Estonian NER test set
Overall	0.774	0.901
PER	0.882	0.953
ORG	0.696	0.805
LOC	0.517	0.907
GPE	0.828	-

Comments on data annotation

- PER, GPE and PERCENT entities most reliably annotated
  - Further analysis needed to identify the sources of confusion in annotations
- The reliability of annotation of the nested entities overall low
  - Might not be good idea to use to train predictive models
- Potential problems with the selection of annotators:
  - Limited diversity
  - Limited motivation

Comments on modeling results

- The presented models are baselines
- Better results might be obtained with base models other than EstBERT
- Based on the modeling results, it seems that new annotations are more complex (even for PER, ORG, LOC)
- More detailed error analyses need to be done

Summary

- Annotated two NER datasets for Estonian
  - Reannotation of the existing Estonian NER dataset
  - Newly annotated 130K tokens from news and social media domains
- Rich set of labels covering 11 different entities
- Also nested annotations with up to three levels
- Experiments with EstBERT-based models
  - on both datasets separately
  - on the merged dataset
- Recommendation: adopt joint model trained on the merged dataset

Thank you!

- Datasets are available (CC-BY-4.0):
  - Main: https://github.com/TartuNLP/EstNER
  - New: https://github.com/TartuNLP/EstNER_new
- The best joint model is available (CC-BY-4.0):
  - https://huggingface.co/tartuNLP/EstBERT_NER_v2