PLOD: An Abbreviation Detection Dataset for Scientific Documents

Leonardo Zilio

Hadeel Saadany

Prashant Sharma

Diptesh Kanojia

Constantin Orasan

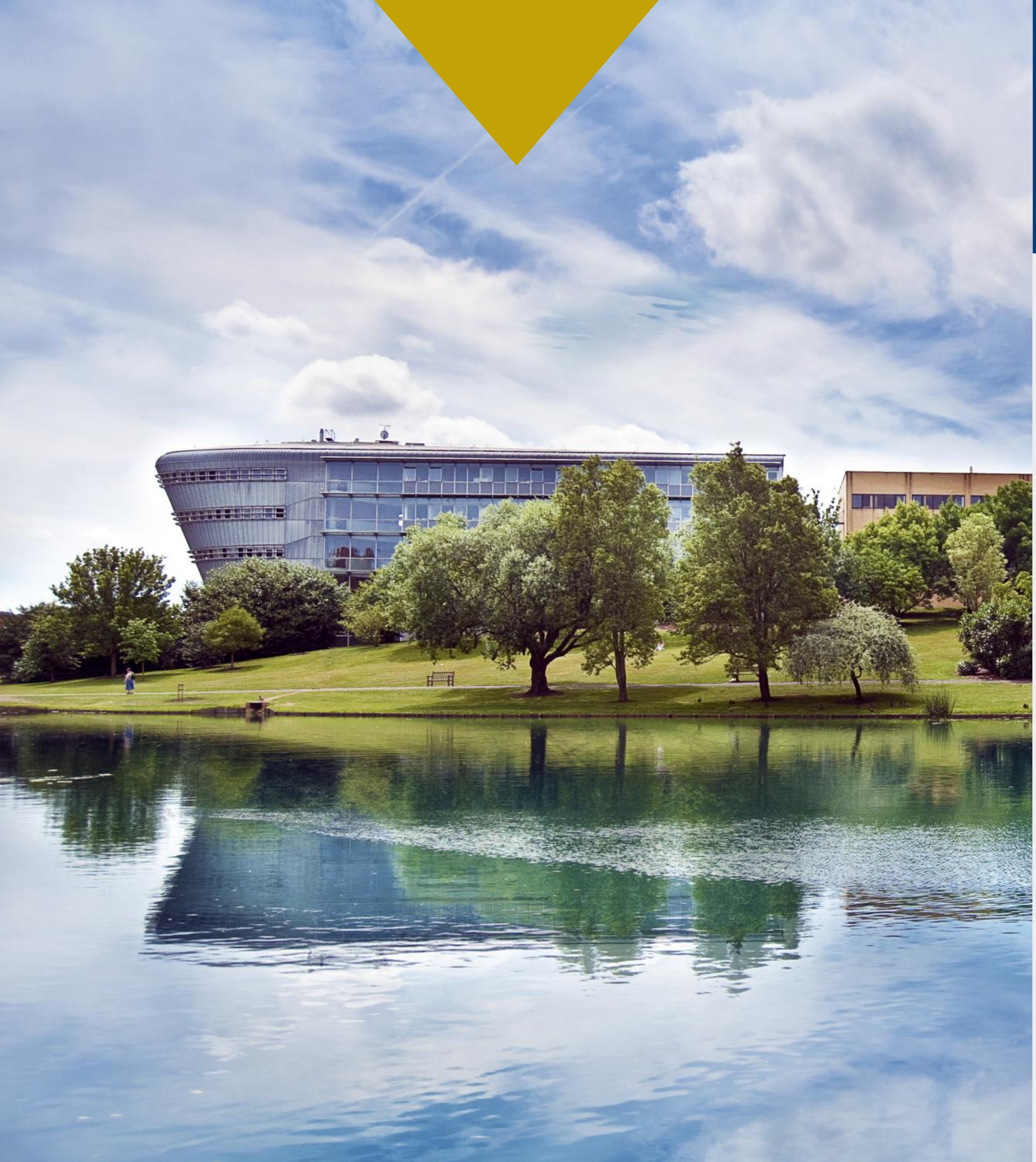
CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

LREC – 20-25 June 2022

Outline

- Introduction
- PLOD Dataset
  - Methodology
  - Validation steps
  - Evaluation
  - Availability
- Extrinsic Evaluation
  - Pre-trained language models
  - Testsets
  - Results
- Final Remarks



People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

Introduction

Introduction

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

» Automatically detecting abbreviations is important for several tasks

- NLP tasks:
  - Machine translation
  - Information extraction
- Linguistic tasks:
  - Translation
  - Glossary creation
  - Typological studies

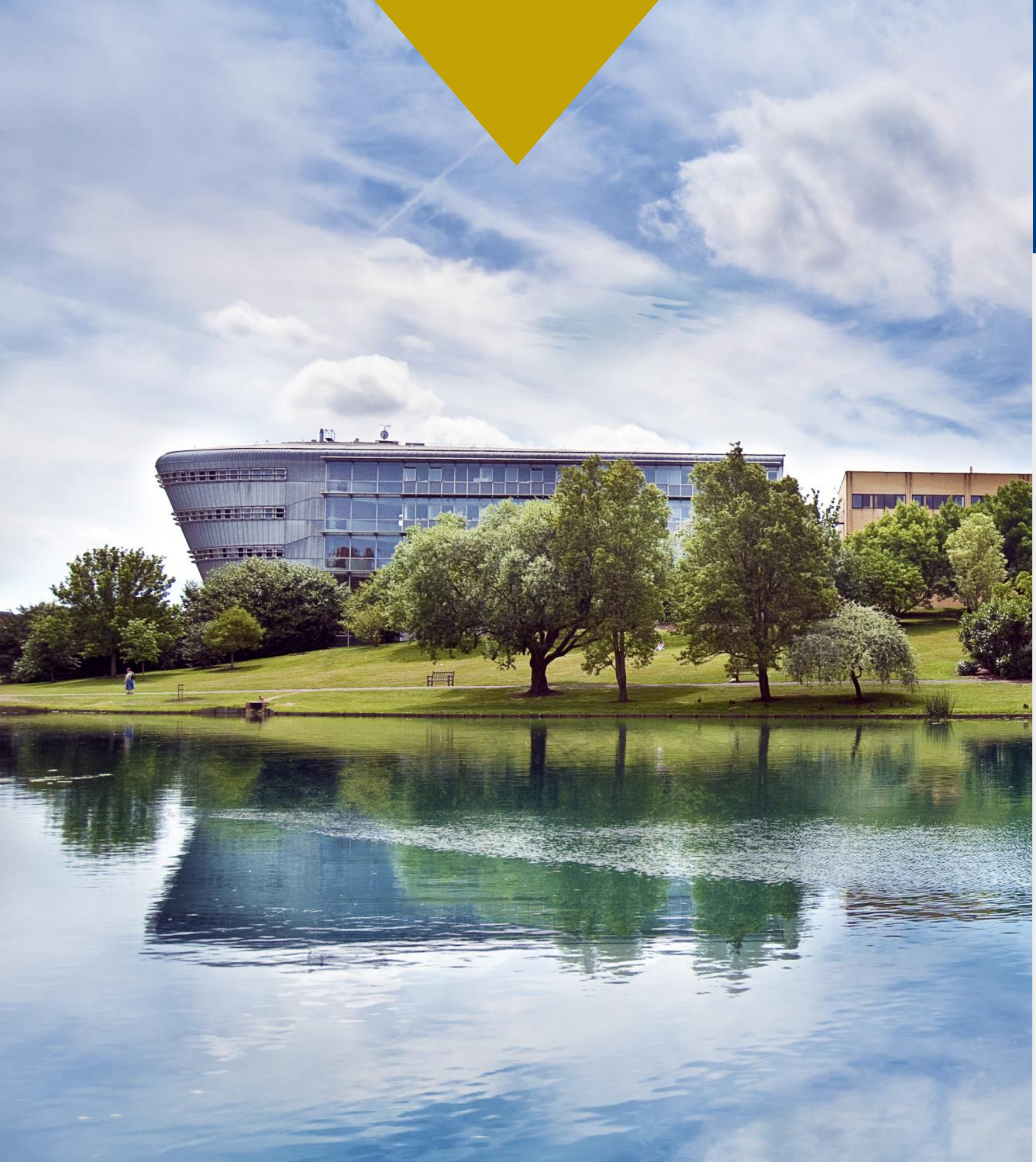
» Contributions of this paper:

- A dataset annotated with abbreviations and their corresponding long forms
- Several pre-trained baseline models for abbreviation detection

Abbreviations: Terminology

» Abbreviations, acronyms, initialisms, blended forms, short forms etc.
  - Different typologies define these terms differently

» We use “abbreviations”, “short forms” or “abbreviated tokens” as umbrella terms



People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

PLOD Dataset

PLOD: Methodology

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

» PLOD uses the PLOS open journals as basis (https://plos.org/)

Journal	Publication Period	Number of Files
PLOS Biology	2003-present	6072
PLOS Medicine	2004-present	4494
PLOS Computational Biology	2005-present	8473
PLOS Genetics	2005-present	9251
PLOS Pathogens	2005-present	9148
PLOS Clinical Trials*	2006-2007	68
PLOS ONE	2006-present	257854
PLOS Neglected Tropical Diseases	2007-present	9388
PLOS Currents	2009-2018	697

*Later merged with PLOS ONE.

This table is based on data downloaded on 16 October 2021.

PLOD: Methodology

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

» Extraction of abbreviations from XML files

- Only files identified as “Research Articles” were used (most of the corpus)

- “<abbreviation>” tag was identified and parsed
  - Extraction of a list of abbreviations associated to their long forms

- “<p>” tags were selected and parsed:
  - Application of simple and fast regex sentence splitter
  - Matching of abbreviations in each segment:
    - If abbreviation was found, then matching of its long form in the segment

» Application of several validation methods

- Both manual and automatic

Validation: First Impression

» Raw extraction:

- >1.3mi annotated segments
- Lots of segments with no long form at all

» Filtered for only segments with long forms:

- >162k segments
- >56k combinations of abbreviations and long forms

» 500 random segments used as first-impression validation

- Detection of main issues

Main Issues in the Raw Extraction

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

- One-character abbreviations

- Missing annotations

Example 1

The reaction of an oligonucleotide substrate bearing a S P-phosphorothioate at the cleavage site (SSp, Table 1) also experiences Cd2+ stimulation with the WT ribozyme.

S = oligonucleotide substrate

SSp and P = not annotated

Main Issues in the Raw Extraction

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

» One-character abbreviations

- Removed all annotations of one-character abbreviations
- A total of 705 unique long forms
- Almost 1.7k segments removed + several annotations in existing segments

» Missing annotations

- Accepted as a minor issue, considering that most segments have several abbreviations

Extra Annotation and Validation

- spaCy
  - Simple language model: stop-words
  - Segments with long forms that start or end with stop-words were annotated
  - Segments with long forms that are longer than 12 words:
    - Manual validation: 36 instances removed

- Validation of long abbreviations
  - >15 characters
  - 11 incorrect abbreviations out of 141

Result of the Validated Extraction

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

After removal of one-character abbreviations and after removal of segments from the previous validation steps

Journal	Number of Segments	Annotated Abbreviations	Annotated Long Forms
PLOS Biology	50975	165099	97002
PLOS Medicine	33036	83549	54237
PLOS Computational Biology	2124	4380	2540
PLOS Genetics	2740	5659	3152
PLOS Pathogens	2394	6225	2814
PLOS Clinical Trials	325	709	410
PLOS ONE	69217	183358	106031
PLOS Neglected Tropical Diseases	121	287	165
Total	**160932**	**449266**	**266351**

Manual Evaluation of PLOD

» 1k random segments

- 55 segments contained at least one wrong annotation
- 267 segments were missing the annotation of at least one abbreviation or long form

PLOD: Availability

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

» PLOD is readily available for download from this GitHub repository:

- https://github.com/surrey-nlp/PLOD-AbbreviationDetection
  - Unfiltered version: raw extraction, all segments have at least one long form
  - Filtered version: validated data, no one-character abbreviations

https://github.com/surrey-nlp/PLOD-AbbreviationDetection

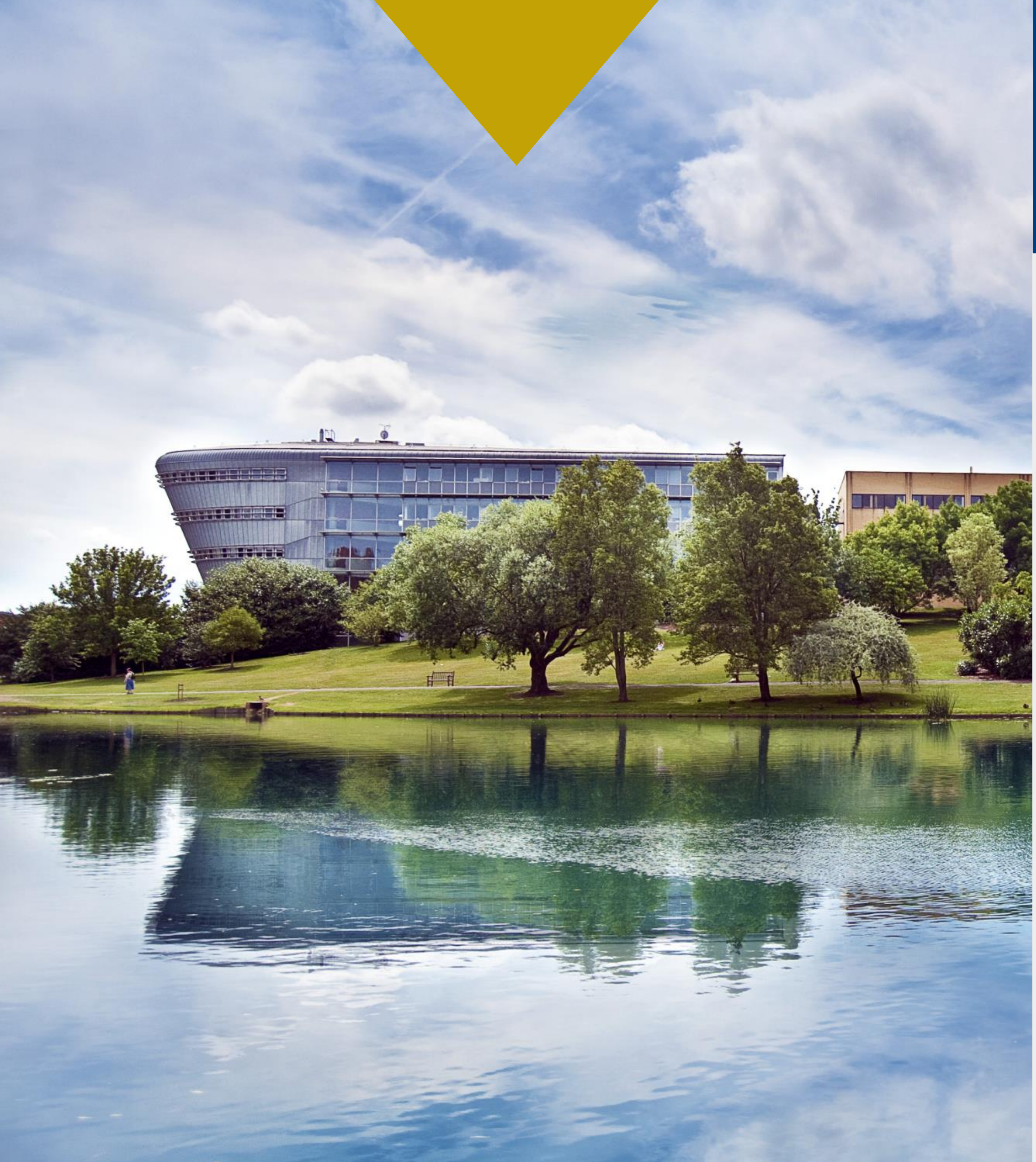
(Code / Documentation)

https://huggingface.co/datasets/surrey-nlp/PLOD-unfiltered

(PLOD: Unfiltered Dataset)

https://huggingface.co/datasets/surrey-nlp/PLOD-filtered

(PLOD: Filtered Dataset)



People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

Extrinsic Evaluation

Pre-trained Language Models

» We used PLOD to fine-tune several language models in the task of abbreviation detection:

- ALBERT (base and large)
- BERT (base and large, both cased)
- DeBERTa (base)
- DistillBERT (base)
- MPNet (base)
- RoBERTa (base and large)

» Random split based on number of segments: 70-15-15

» Models fine-tuned both on the unfiltered and filtered versions

Test Set

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

» All models were tested against both PLOD and the SDU Acronym Extraction* dataset:

- PLOD: Random 15% segments of the dataset (as per training/validation/test split)
- SDU: combined both train and validation sets and used them for testing

*https://sites.google.com/view/sdu-aaai22/home

Results: Unfiltered PLOD

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

	$\text{PLOD}_{\text{test-unfiltered}}$						$\text{SDU@AAAI-22 Shared Task}_{\text{train + dev}}$					
	Abbreviations			Long-forms			Abbreviations			Long-forms		
	P	R	F	P	R	F	P	R	F	P	R	F
$\text{ALBERT}_{\text{base}}$	0.845	0.898	0.871	0.758	0.812	0.784	0.682	0.638	0.659	0.462	0.154	0.231
$\text{BERT}_{\text{base-cased}}$	0.855	0.906	0.880	0.781	0.826	0.803	0.691	0.650	0.670	0.461	0.151	0.228
$\text{DeBERTa}_{\text{base}}$	0.877	0.910	0.893	0.817	0.874	0.845	0.682	0.638	0.659	0.462	0.154	0.231
$\text{DistillBERT}_{\text{base}}$	0.845	0.900	0.872	0.772	0.798	0.785	0.700	0.641	0.670	0.467	0.139	0.214
$\text{MPNet}_{\text{base}}$	0.846	0.899	0.872	0.782	0.823	0.802	0.691	0.606	0.645	0.466	0.145	0.221
$\text{RoBERTa}_{\text{base}}$	0.860	0.919	0.889	0.805	0.862	0.833	**0.707**	**0.641**	**0.672**	0.516	0.163	0.248
$\text{ALBERT}_{\text{large}}$	0.895	0.920	0.907	0.848	0.898	0.872	0.476	0.607	0.534	0.397	0.160	0.228
$\text{RoBERTa}_{\text{large}}$	**0.911**	**0.935**	**0.922**	**0.876**	**0.921**	**0.898**	0.515	0.650	0.575	**0.423**	**0.191**	**0.264**
$\text{BERT}_{\text{large-cased}}$	0.899	0.928	0.913	0.866	0.909	0.887	0.532	0.645	0.583	0.362	0.173	0.234

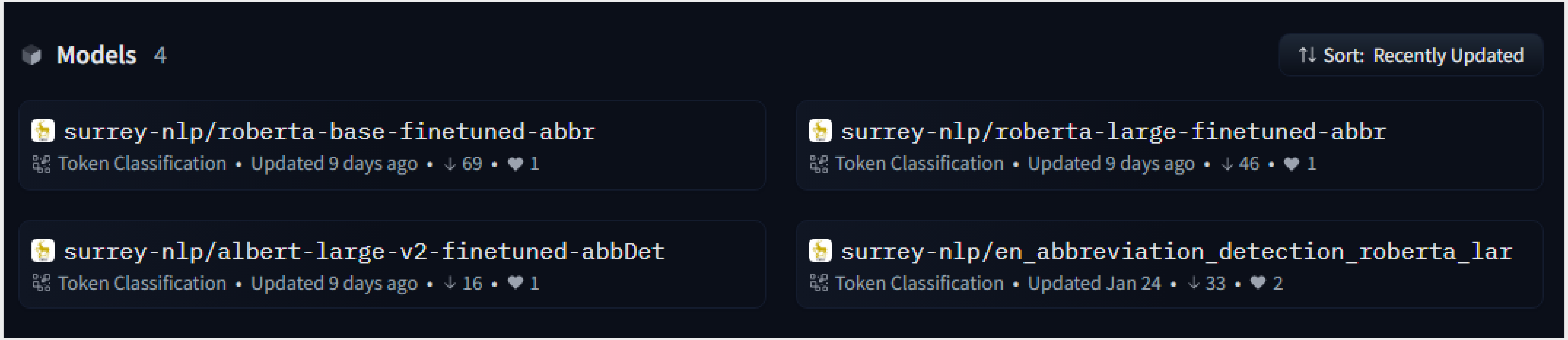
Results: Filtered PLOD

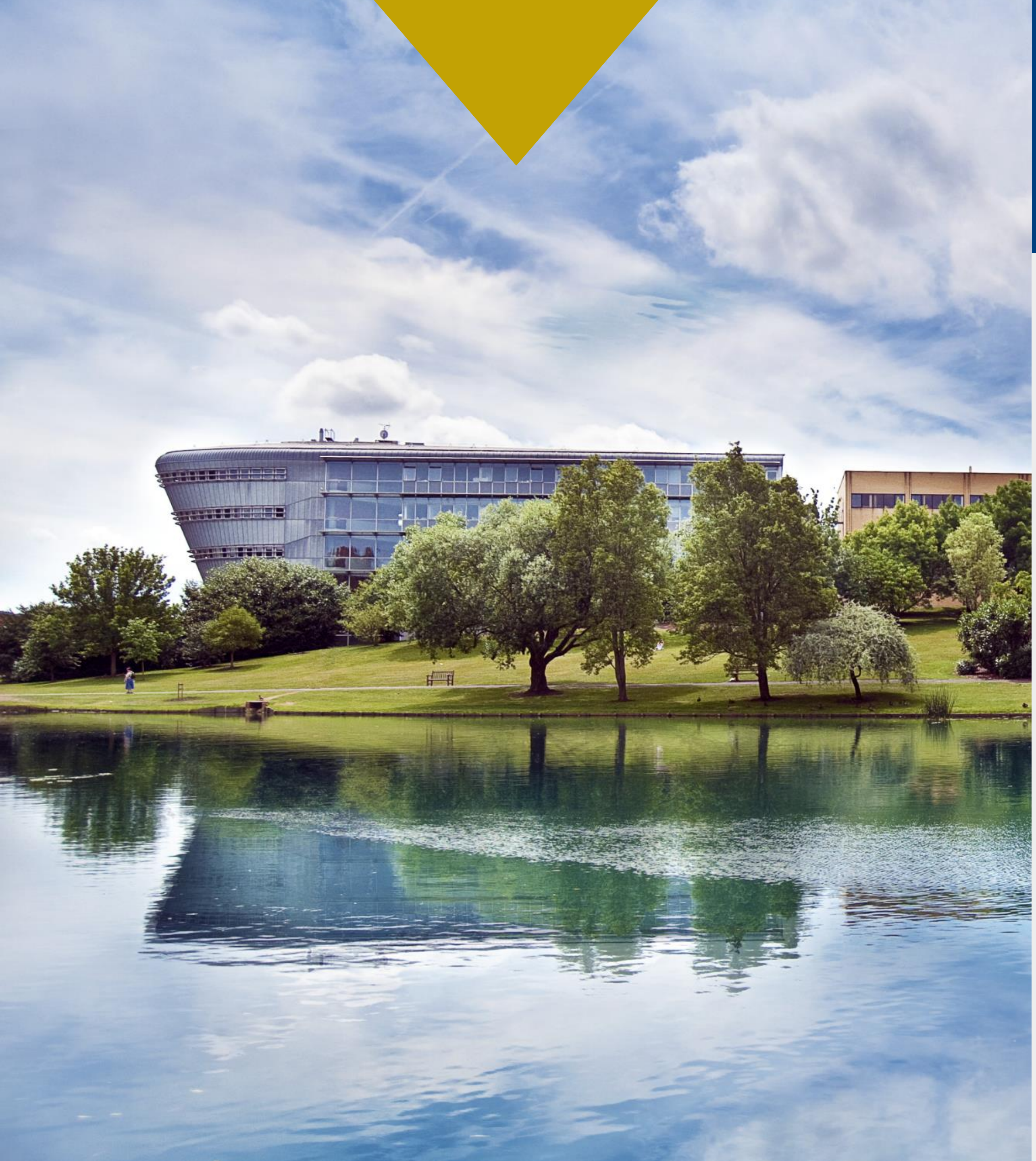
	PLOD$_{\text{test-filtered}}$						SDU@AAAI-22 Shared Task$_{\text{train + dev}}$					
	Abbreviations			Long-forms			Abbreviations			Long-forms		
	P	R	F	P	R	F	P	R	F	P	R	F
ALBERT$_{\text{base}}$	0.842	0.899	0.870	0.734	0.819	0.774	0.716	0.629	0.670	0.485	0.146	0.225
BERT$_{\text{base-cased}}$	0.853	0.902	0.877	0.766	0.834	0.799	0.723	0.628	0.672	0.471	0.150	0.228
DeBERTa$_{\text{base}}$	0.852	0.937	0.893	0.803	0.881	0.840	0.691	0.606	0.645	0.466	0.145	0.221
DistilBERT$_{\text{base}}$	0.842	0.904	0.872	0.763	0.805	0.783	0.709	0.642	0.674	0.456	0.140	0.215
MPNet$_{\text{base}}$	0.852	0.888	0.870	0.777	0.824	0.800	0.711	0.586	0.642	0.472	0.147	0.224
RoBERTa$_{\text{base}}$	0.857	0.918	0.886	0.798	0.867	0.832	**0.728**	**0.643**	**0.683**	**0.520**	**0.169**	**0.255**
ALBERT$_{\text{large}}$	0.840	0.918	0.877	0.770	0.830	0.799	0.532	0.651	0.585	0.373	0.174	0.237
RoBERTa$_{\text{large}}$	**0.906**	**0.935**	**0.920**	**0.874**	**0.925**	**0.898**	0.502	0.645	0.564	0.427	0.181	0.254
BERT$_{\text{large-cased}}$	0.892	0.931	0.911	0.858	0.912	0.884	0.532	0.651	0.585	0.373	0.174	0.237

Fine-tuned Models: Availability

» The best fine-tuned models are readily available in our Huggingface repository:

- https://huggingface.co/surrey-nlp





Final Remarks

Final Remarks

People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

» We introduced PLOD, a new dataset with annotated abbreviations and their long forms
  - With more than 160k annotated segments, this dataset is large enough to have both linguistic and computational value

» We performed several validation steps, both manual and automatic
  - Unfiltered and filtered version

» We fine-tuned several pre-trained language models for abbreviation detection and tested them against our own dataset and against the SDU Acronym Detection dataset

» The dataset and the best-performing fine-tuned models were made available in GitHub and Huggingface repositories

PLOD: An Abbreviation Detection Dataset for Scientific Documents

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

Thank you!

https://www.surrey.ac.uk/centre-translation-studies/
l.zilio@surrey.ac.uk

LREC – 20-25 June 2022