EXCITE project: RefExt

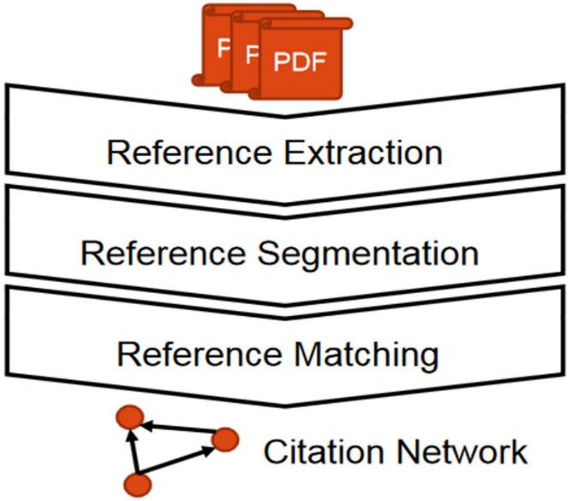
Martin Körner, Behnam Ghavimi, Azam Hosseini
Philipp Mayr, Heinrich Hartmann, and Steffen Staab

EXCITE Objectives

EXtraction of **CIT**ation from PDF Docum**E**nts

Objectives:

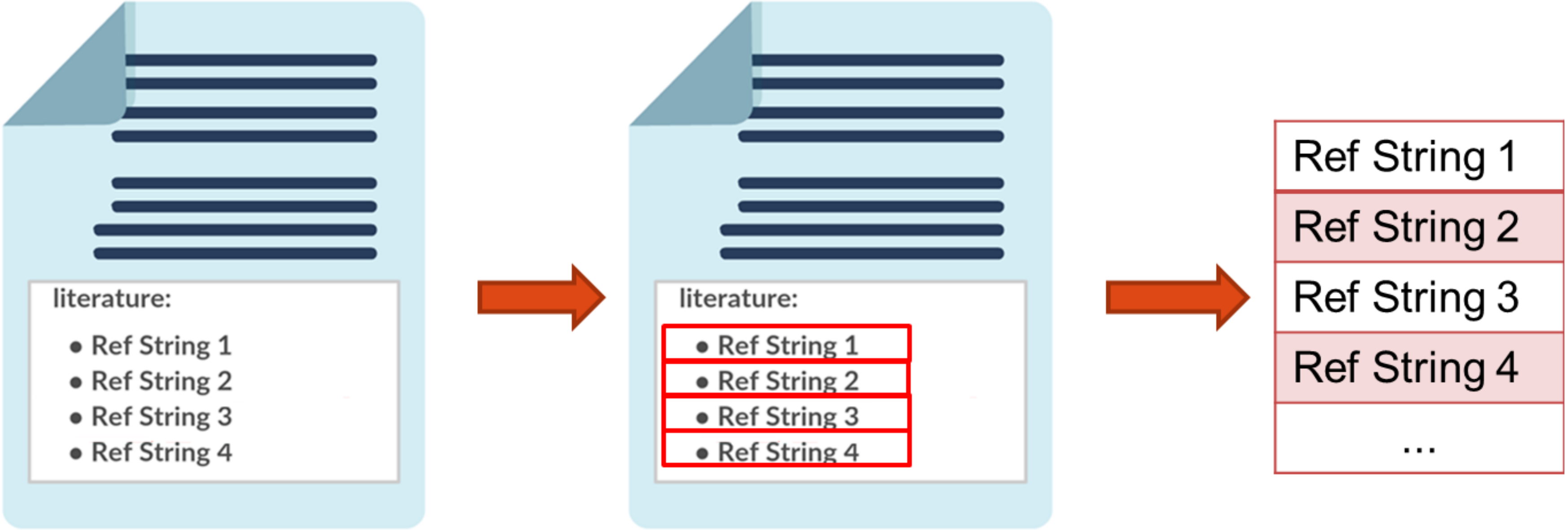
- Developing a toolchain of citation extraction and matching software
- Tools and data will be made available to researchers



http://west.uni-koblenz.de/en/research/excite

Reference String Extraction

- High number of possible reference styles (e.g. https://www.zotero.org/styles/ contains more than 400 different styles – only for social science)
- Large variety of layouts for publications



literature:

- Ref String 1
- Ref String 2
- Ref String 3
- Ref String 4

literature:

- Ref String 1
- Ref String 2
- Ref String 3
- Ref String 4

Ref String 1
Ref String 2
Ref String 3
Ref String 4
…

Related Work

Current solutions that perform reference string extraction:

1. Identify the reference section
2. Segment the reference section into individual reference strings

Thereby, errors that are made during the first step directly impact the accuracy of the reference string extraction.

Related Work (Cont.)

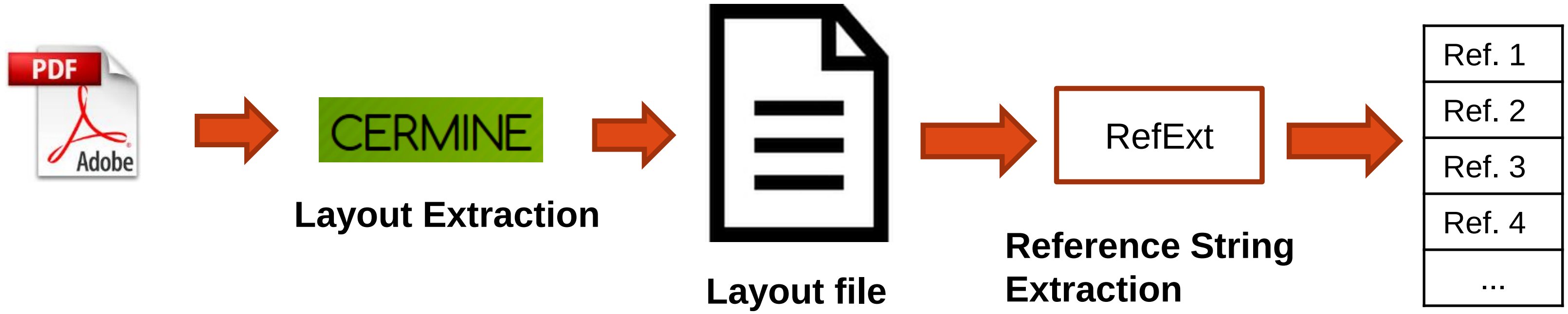
- Step1 - Identify the reference section:
  - ParsCit[1] uses a set of regular expressions
  - Cermine[2] uses a trained SVMs model
  - GROBID[3] uses a CRF model

- Step 2 - Segment the reference section into individual reference strings
  - ParsCit applies regular expression
  - Cermine uses k-means learning algorithm
  - GROBID uses a CRF model

Sources:

1. https://github.com/knmnyn/ParsCit
2. https://github.com/CeON/CERMINE
3. https://github.com/kermitt2/grobid

Our Approach (RefExt)



- Every line in text is potentially part of a reference string
- Use of layout features and textual features per line for machine learning
- Training of supervised conditional random fields
- This CRF model tags every line with our BIO-annotation

Source code: https://github.com/exciteproject/refext

BIO-annotation

- Target variables to predict are for each line one of:
  - B-REF: Beginning of a reference (first line)
  - I-REF: Intermediate reference (second+ line)
  - O: Other (not part of a reference string)

#	BIO	Text
529	**…**	
530	**O**	appear in the footnotes.This could present an
531	**O**	interesting use case of our approach.
532	**O**	References
533	**B-REF**	Abbate, Janet Ellen, 1999: Inventing the inter-
534	**I-REF**	net. Cambrige/ MA: MIT Press.
535	**B-REF**	Barber, Benjamin R., 1998: A Place for US.
536	**I-REF**	New York: Hill and Wang.

Features

Textual features:

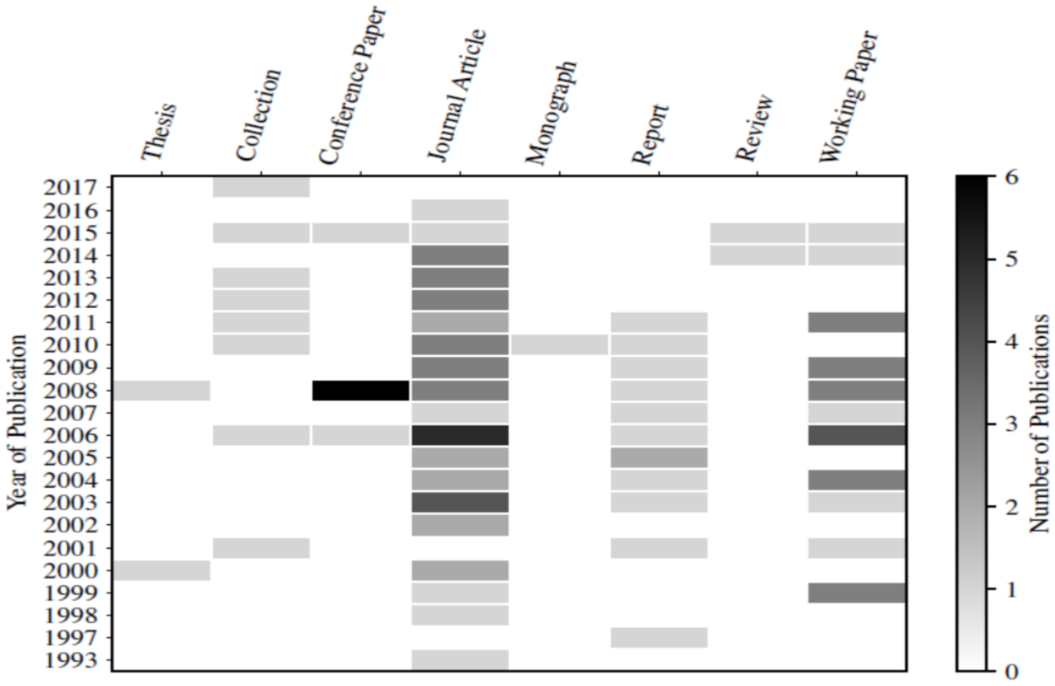
1. Line starts with a capitalized letter, ends with a period, a comma, contains a year, a page range, a quotations mark
2. counts the occurrences of numbers, words, periods, commas, and words that only consist of one capitalized letter

Layout features:

1. current line is indented when compared to the previous line
2. gap between the current and previous lines
3. line contains less characters than the previous one
4. the position of a given line in the whole document

Reference Extraction Evaluation

- 100 Random full-text Paper (https://github.com/exciteproject/ssoar-gold-standard)
  - From SSOAR(http://www.ssoar.info/) corpus
  - Text PDF
  - Contain Reference Section
  - German
- We manually annotated them. They contain 5,355 reference strings.



Evaluation

- Most existing tools focus on English publications
- Adapt these tools to German language publications.
- We retrain CERMINE and GROBID
- We Modified ParsCit (e.g. adding “Literatur” and “Anhang”.)
- We exclude other tools that we could adapt them such as PDFX and pdfextract

Evaluation (Cont.)

- Macro-metrics of BIO-annotated reference lines (10-fold cross-validation)

Metric	CER-D	CER-T	Pars-D	Pars-M	GRO-D	GRO-T	RefExt-T
B-REF Precision	0.719	0.734	0.683	0.769	0.692	0.871	**0.916**
B-REF Recall	0.600	0.557	0.620	0.688	0.789	0.865	**0.952**
B-REF F1-Score	0.616	0.589	0.616	0.689	0.712	0.861	**0.922**
I-REF Precision	0.729	0.755	0.577	0.678	0.664	0.857	**0.882**
I-REF Recall	0.340	0.313	0.809	0.843	0.839	0.871	**0.944**
I-REF F1-Score	0.432	0.415	0.647	0.716	0.703	0.855	**0.902**

- CER:CERMINE, GRO:GROBID, and Pars:ParsCit
- D: Default , T: Trained , and M: Modified
- Pars-M (V. 101101), Pars-D (May 31, 2017)
- CERMINE (V. 1.13) and GROBID (V. 0.4.1.)
- RefExt (V. 0.1.0)

Evaluation (Cont.)

- Macro-metrics of reference string extraction using 10-fold cross-validation

Metric	CER-D	CER-T	Pars-D	Pars-M	GRO-D	GRO-T	RefExt-T
Precision	0.296	0.303	0.558	0.617	0.627	0.847	**0.879**
Recall	0.233	0.220	0.552	0.595	0.718	0.839	**0.906**
F1-Score	0.245	0.235	0.542	0.590	0.650	0.837	**0.885**

- GROBID had a recall of zero in 7 publications
- RefExt had a recall of zero for 2 publications
- But recall of 1.0 and 0.662 in GROBID

Future Work about RefExt

- Evaluation on English papers
- It remains to be evaluated how the performance improves when extending the training data
- Improvements might also be possible by adding more domain-specific features

Paper about RefExt

- Körner M., Ghavimi B., Mayr P., Hartmann H., Staab S. (2017) Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications. In: Kirikova M. et al. (eds) New Trends in Databases and Information Systems. ADBIS 2017. Communications in Computer and Information Science, vol 767. Springer, Cham DOI: https://doi.org/10.1007/978-3-319-67162-8_15