EXCITE Development Status

EXCITE Workshop 2017

Behnam Ghavimi and Martin Körner

30.3.2017

gesis Leibniz-Institut für Sozialwissenschaften

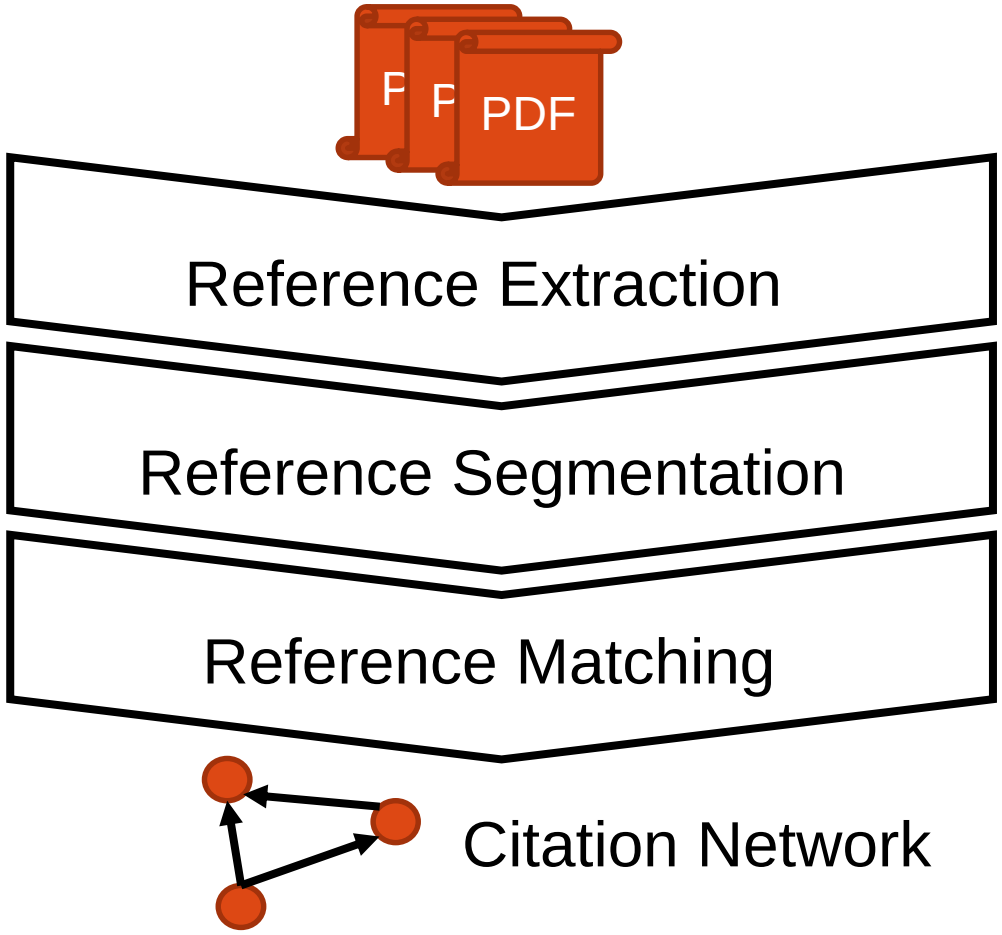
WeST People and Knowledge Networks

Outline

1. Pipeline Overview
2. Reference Extraction (WP 1)
3. Reference Segmentatin (WP 1)
4. Reference Matching (WP 2)
5. Future Work

EXCITE Pipeline Overview

- **EX**traction of **CIT**ations from PDF Docum**E**nts



Reference Extraction

- Typical Approach
  1. Find the reference section
  2. Split the reference section into reference strings
- Our approach
  - No explicit extraction of reference section
  - Every line in text is potentially part of a reference string
  - Usage of layout features and textual features per line for machine learning

Reference Extraction Implementation

- Training of supervised conditional random fields
- 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
  - Special case where lines of reference are spread over two pages with non-reference text (e.g. page number/footer) in between

Reference Extraction Example

#	BIO	Text
	...	...
1	O	References
2	B-REF	Abbate, Janet Ellen, 1999: Inventing the Inter-
3	I-REF	net. Cambridge/ MA: MIT Press.
4	B-REF	Barber, Benjamin R., 1998: A Place for Us.
5	I-REF	New York: Hill and Wang.
6	B-REF	Berners-Lee, Tim, 1999: Weaving the Web

- Possible features
  - Line 2: LastName, Year, Colon, GapAbove, EndsHyphen
  - Line 3: Indent, City, Colon, EndsPeriod

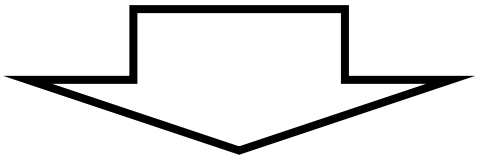
Reference Extraction Results

- Current results for German social science papers from SSOAR
- Trained on 63 PDFs (2089 References)
- Tested on 18 PDFs (993 References)

Label	Precision	Recall	F1 Score
B-REF	0.9757	0.9558	0.9656
I-REF	0.9341	0.9911	0.9617
O	0.9992	0.9978	0.9985

Reference Extraction Result Example

Cohen, Josua/ Joel Rogers (Hrsg.), 1995: *Associations and Democracy.* London: Verso.
Conca, Ken, 1996: Greening the UN: Environmental Organisations and the UN System. In: Thomas G. Weiss/ Leon Gordenker (Hrsg.), *NGOs, the UN, & Global Governance.* London: Lynne Rienner Publishers, 102-119.
CSTB (Computer Science and Telecommunications Board, National Research Council), 1999: *Funding a Revolution. Government Support for Computing Research.* Washington, D.C.: National Academy Press.



5. Cohen, Josua/ Joel Rogers (Hrsg.), 1995: Associations and Democracy. London: Verso.
6. Conca, Ken, 1996: Greening the UN: Environmental Organisations and the UN System. In: Thomas G. Weiss/ Leon Gordenker (Hrsg.), NGOs, the UN, & Global Governance. London: Lynne Rienner Publishers, 102-119.
7. CSTB (Computer Science and Telecommunications Board, National Research Council), 1999: Funding a Revolution. Government Support for Computing Research. Washington, D.C.: National Academy Press.

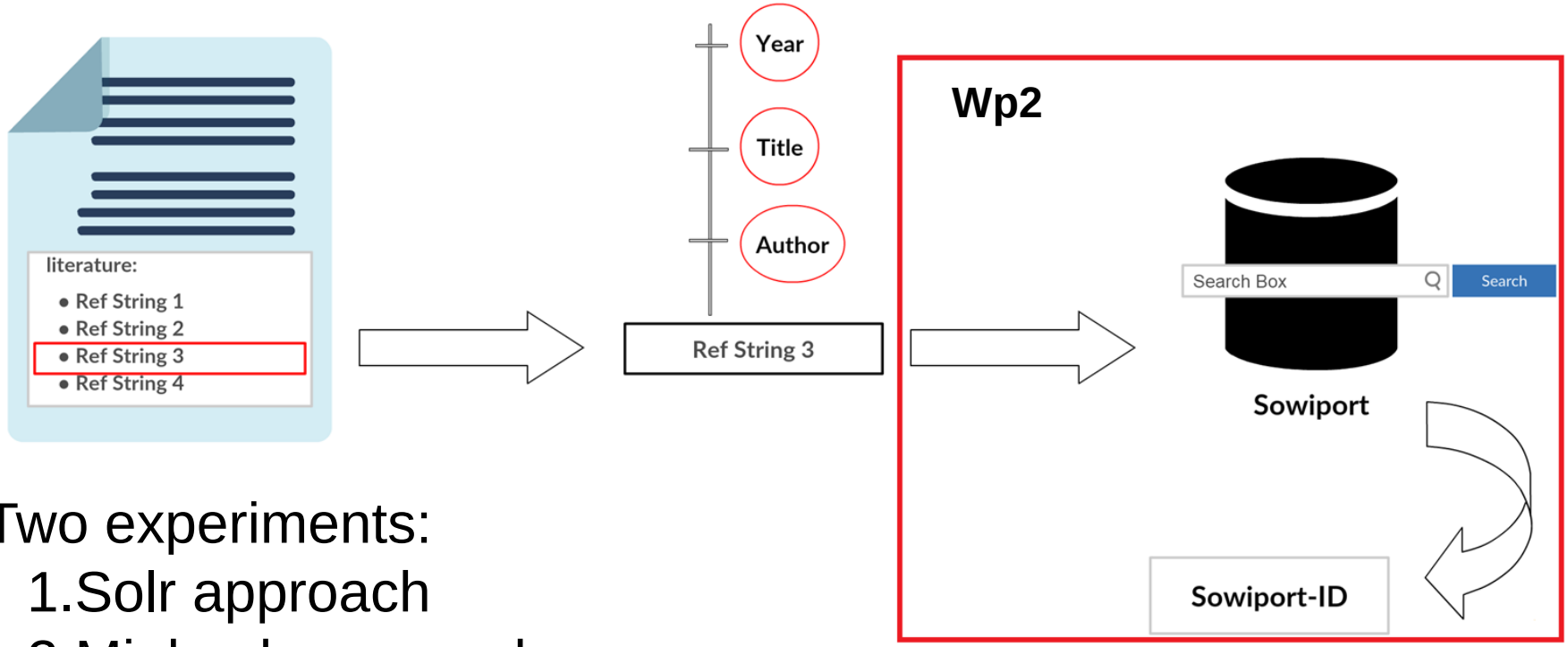
Paper ID: 23446-36368

Reference Segmentation

- Split a given reference string into its components:
  - Author
  - Year
  - Title
  - Journal
  - …
- Currently used for this task: CERMINE
- Focus on author, title, and year because of their importance for matching

Reference Matching

Match reference strings in PDF documents to corresponding items in a database of publications (e.g. Related-work.de or Sowiport.de)



Two experiments:

1. Solr approach
2. Minhash approach

Solr Approach

- Sowiport is one of the repositories of publication which will be used as the target of our matching reference strings.
- Sowiport uses Solr for search function and indexing.
- We easily can pass a query to solr like the below query and then receive lists of items:

```
(title:aspirin~0.5) AND (author:lewis OR author:blume) AND (Year:1983)
```

Levenshtein-distance:

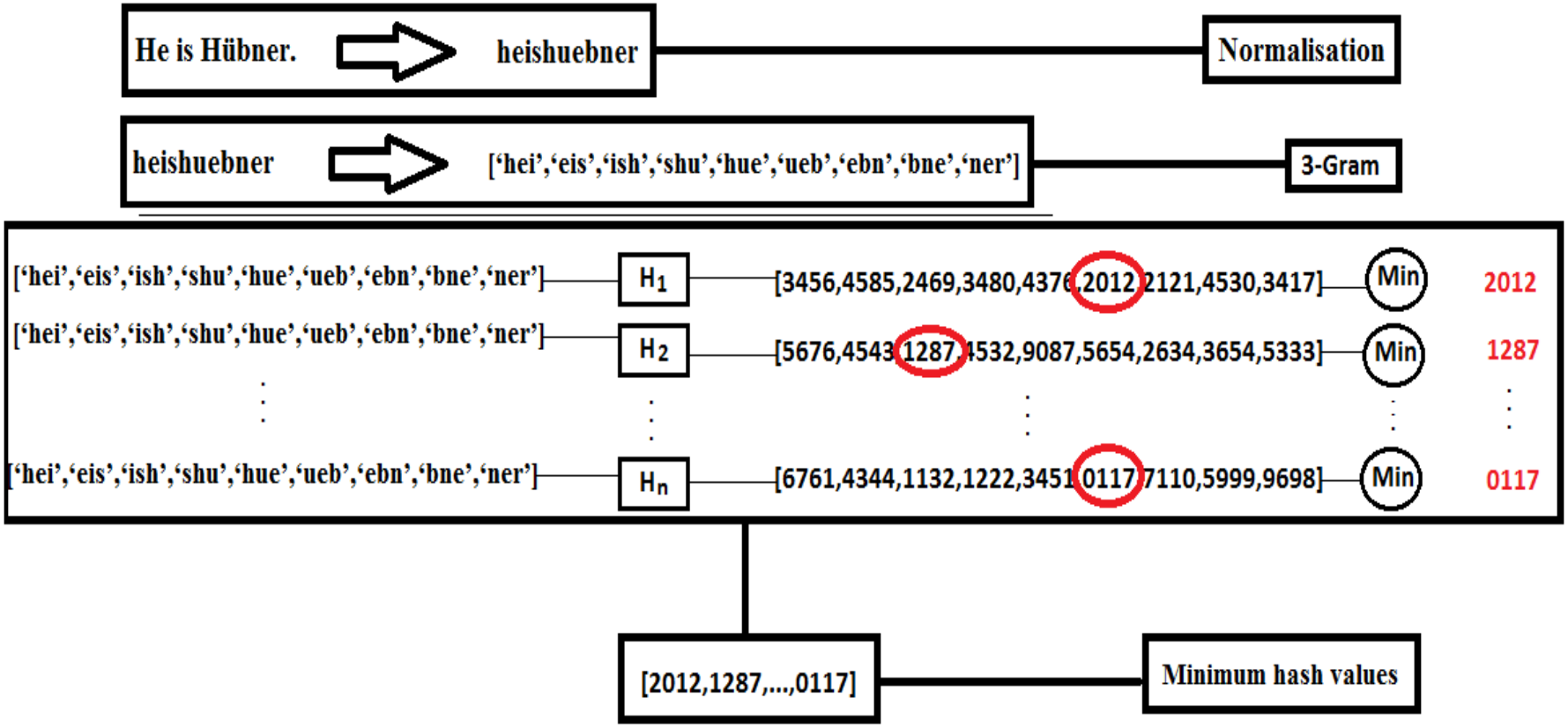
search(title:aspirin~0.71)=> spirit
aspiri**n**-> (delete,replace)=2 , Max(len(aspirin),len(spirit))=7
similarity=1-(2/7)=0.71

Minhash Approach

1. Generate min hash values for titles in Sowiport (once).
2. Generate min hash values for extracted titles of reference string
3. Filter items in sowiport by extracted year and authors' names of reference string
4. Apply Jaccard comparison on min hash values of titles of reference string and filtered items in sowiport
5. Rank the sowiport items regarding their Jaccard score
6. The match item to the reference string is the top item in the ranking which is higher than our predefined threshold

Minhash Approach

Generating minhash value:



Evaluation

- 230 reference strings
- Randomly picked from our PDF corpus
- The metadata extracted from the reference strings in two ways:
  1. Human assessor (Data without noise)
  2. Cermine v. 1.12

Opt. Precision: year and Title~0.9 and Author

Opt. Recall: year and Title~0.35 and Author

Solr	Opt. Precision		Opt. Recall	
	Precision	Recall	Precision	Recall
Human Data	1.0	0.69	0.85	0.75
Cermine Data	1.0	0.47	0.83	0.52

Minhash	Opt. Precision		Opt. Recall	
	Precision	Recall	Precision	Recall
Human Data	1.0	0.62	0.95	0.84
Cermine Data	1.0	0.55	0.87	0.82

Future Work

- Reference Segmentation
  - Own model or retrain CERMINE
  - Focus on fields that are important for matching
- Overall optimization
  - Pass confidence values between steps
  - Allow multiple possible extractions/segmentations and decide in later step which one to use
- Integration of data into platforms
  - related-work.net
  - sowiport.gesis.org

Thank you for your attention!

Backup Slides

Foundation of My Work: Conditional Random Fields

- Probabilistic graphical model by Lafferty et al. (2001)
- Conditional prob. distr. of target variables $\mathbf{Y}$ given observed variables $\mathbf{X}$
- E.g. $\mathbf{Y}$ = Assigned POS-Tags for words in sequence,
  $\mathbf{X}$ = Set of features per word (capitalized, punct.)