

Automatische Referenzextraktion mit PARSCIT

Karima Haddou ou Moussa & Philipp Mayr

1. Motivation

- Nutzung von Zitationsinformationen für Mehrwertdienste (z.B. Ranking)
- Aufbau eines Zitationsindex für die Sozialwissenschaften

2. Fragestellung

- Können Referenzen automatisch extrahiert werden?
- Wie hoch ist die Fehlerquote bei der Extraktion?
- Wie lässt sich der Extraktionsfehler verkleinern?

3. Verwendete Komponenten



PDF-Box

- Open source Java PDF-Bibliothek
- PDF to Text Konverter

ParsCit

- Perl Open source
- automatische Extraktion und Auszeichnung von Referenzen aus Textdateien
- Einsatz des CRF-Modells (CRF++)

4. Problemstellung

Erkennung von Referenzen mittels ParsCit erfolgt nur wenn die Beschriftung der Referenzlisten:

- „References“
 - „Bibliography“
 - oder „References and Notes“
- lautet.

5. Funktionsweise

Referenzen im PDF

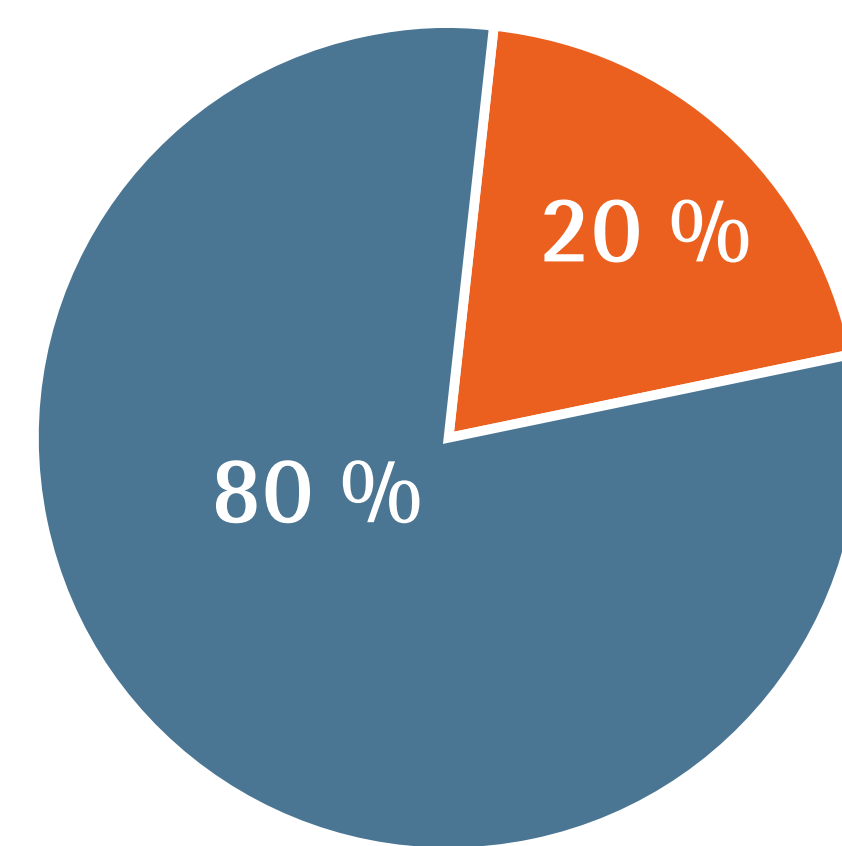
Schütt, P.; Weimer, St. (2005): Beschäftigungsentwicklung in der Region Main-Rhön. München.

PDF-Box

Startpunkt setzen

ParsCit

Erkannte Metadaten der Referenzen



■ Referenzanteil erkannt
■ Referenzanteil nicht erkannt

```
<citation valid="true">
<authors>
<author>P. Schütt</author>
<author>St. Weimer</author>
</authors>
<title>Beschäftigungsentwicklung in der Region Main-Rhön.</title>
<date>2005</date>
<location>München.</location>
<marker>Schütt, Weimer, 2005</marker>
<rawString>Schütt, P.; Weimer, St. (2005): Beschäftigungsentwicklung in der Region Main-Rhön. München.</rawString>
</citation>
```

6. Ergebnisse

Erkennungsquote

Ergebnisse von ParsCit auf 1.000 SSOAR-PDFs

Author	91,53 %	Issue	50,00 %
Titel	87,18 %	location	80,67 %
Booktitle	77,42 %	note	61,19 %
date	94,70 %	pages	88,54 %
editor	68,12 %	publisher	76,55 %
institution	76,26 %	tech	85,71 %
journal	77,40 %	volume	91,04 %

7. Ausblick:

- OCR-Behandlung von PDFs
- Anpassung und Training von ParsCit auf GESIS-Daten sowie andere Zitationsstile
- Evaluation der Qualität des Tools mit größeren Volltextkorpora