Automatische Referenzextraktion mit PARSCIT

Karima Haddou ou Moussa & Philipp Mayr

1. Einleitung

Meist werden am Ende eines wissenschaftlichen Dokuments Referenzlisten angegeben, die die verwendeten Quellen und Hinweise auf Arbeiten anderer Wissenschaftler beinhalten.
In unserem Projekt haben wir uns die Frage gestellt, wie man auf diese Referenzlisten von wissenschaftlichen Artikeln automatisiert zugreifen kann. Zur Extraktion der Literaturangaben soll eine automatisierte Prozesskette entwickelt werden, um mit Open Source-Tools Referenzen automatisiert aus Volltextdokumenten zu extrahieren. Eine nutzbringende Weiterverarbeitung von Referenzen erfordert eine möglichst eindeutige Erkennung von typischen Metadaten.
Die Auflistung der zitierten Literatur liegt häufig in sehr heterogenen Zitationsstilen vor. Dies erschwert die Extraktion der einzelnen Referenzbestandteile wie (z.B. Titel, Autor, Jahr) erheblich.
Eine gute Basis für die Lösung der oben genannten Schwierigkeiten zur Verarbeitung von Referenzen stellt das in diesem Projekt eingesetzte Open Source-Programm Parscit dar. Für die Vorverarbeitung der PDF-Dateien haben wir PDFBox verwendet, das sich wegen der einfachen Einbindung von vorhandenen Bibliotheken und der Möglichkeit, viele Dateien automatisiert zu verarbeiten, angeboten hat.

2. Verwendete Komponenten

Parscit[1] ist eine freiverfügbare Software, die darauf spezialisiert ist, Referenzen aus Textdateien auszulesen und zu analysieren. Die Software stellt Funktionen zur Verfügung, um logische Strukturen aus einer gegebenen Veröffentlichung zu extrahieren. Dazu wird das Parscit zugrunde liegende CRF-Modell (Conditional Random Field) mit zusätzlichen heuristischen Funktionen kombiniert. CRF ist ein ungerichtetes graphisches Modell, das zur Segmentierung und Kennzeichnung sequentieller Daten sowie der Informationsextraktion natürlicher Sprache genutzt wird. Bei Parscit wird die CRF-Implementierung „CRF++“[2] eingesetzt.

PDFBox[3] ist eine Open Source Java PDF-Bibliothek, mit der sich PDF-Dateien anlegen, öffnen, verändern und Inhalte auslesen lassen. PDFBox bietet neben der Konvertierung von PDF in Text einige Vorteile, z.B. Zusammenführung mehrerer PDF-Dokumente, Erstellung einer PDF-Datei aus einer Textdatei und Bildererstellung aus PDF-Seiten.

3. Funktionsweise

Damit Referenzen mittels Parscit automatisch extrahiert werden können, wird hier ein Verfahren umgesetzt, das mehrere PDF-Dateien gleichzeitig konvertiert und Zitate aus den konvertierten Texten extrahiert. Im folgenden Abschnitt wird der Extraktionsprozess in groben Zügen beschrieben.

[1] http://aye.comp.nus.edu.sg/parsCit/
[2] http://crfpp.sourceforge.net/
[3] http://pdfbox.apache.org/

a. PDF-Konvertierung

Der erste Schritt ist die Vorverarbeitung der rohen PDF-Dateien. Die Referenzextraktion mit Parscit erfolgt erst, nachdem die PDF-Dokumente in UTF-8 kodierten Reintext konvertiert worden sind.

b. Pre-Processing

Um die Referenzen eines Artikels mittels Parscit identifizieren zu können, ist es zuerst erforderlich, den Textabschnitt mit Literaturangaben innerhalb dieses Artikels zu finden. Dies erfolgt durch Anwendung heuristischer Verfahren. Hierbei wird zunächst der Start- und Endpunkt der Referenzliste bestimmt. Der Startpunkt wird innerhalb von Parscit durch die folgenden Bezeichnungen der Zitate: „References“, „Bibliography“, “References and Notes“ definiert. Der Endpunkt wird durch die Suche von Abschnittsbezeichnungen wie z. B. „tables“, „figures“ oder das Ende des Dokumentes festgelegt.

Nachdem der Literaturabschnitt eingegrenzt wurde, ist es wichtig, die Markierungsart für die Zitatliste zu erkennen. Diese Erkennung erfolgt zuerst durch Segmentierung der Referenzen. Anschließend wird der Haupttext mittels regulärer Ausdrücke nach Zitaten durchsucht. Dabei sind drei Typen von Formatierungen zu berücksichtigen:

1. Referenzstrings beginnen mit Klammern (z. B. [1], (1));
2. Referenzstrings haben eine einfache Nummerierung (z. B. 1 oder 1.);
3. unmarkierte Zitierungen.

c. Post-Processing

Basierend auf der Ausgabe des CRF++-Modells müssen einige weitere Analyseschritte durchgeführt werden. Die extrahierten Referenzstrings werden untersucht und nach Trennstellen (Komma, Semikolon etc.) aufgesplittet. Danach wird jedes markiertes Feld in eine Standard-Darstellung normalisiert. Autorennamen können zum Beispiel in verschiedenen Formaten “M.-Y. Kan and I. G. Councill” oder “Kan, M.-Y. & Councill, I. G.” auftreten, werden allerdings in ein einheitliches Format wie “M-Y Kan” und “I G Councill” überführt. Diese Normalisierung wird nicht nur bei Autorennamen, sondern auch bei Nummern („vol. 7“) sowie Seitenzahlen („pp. 13 - 42“) durchgeführt.

4. Beispielhafte Umsetzung

Zum Testen der zuvor beschriebenen Prozesskette wurde Parscit auf ein Sample von ca. 3.000 PDF-Dokumenten aus SSOAR[4] (Social Science Open Access Repository) angewendet.

Beim ersten Testen des Programmes Parscit kann man feststellen, dass die unterschiedlichen nationalen Bezeichnungen für Referenzen, wie z. B. „Literatur“, zunächst von Parscit nicht verarbeitet werden können. Daher wurde ein Ersetzungsprozess mit Hilfe einer Liste von Heuristiken entwickelt.

[4] http://www.ssoar.info/

Nach einem Parscit-Durchlauf soll das Ergebnis als XML-Datei ausgegeben werden. Ein Beispiel für einen Referenzstring vor und nach der Verarbeitung mit Parscit findet sich in Abb. 1 und 2.

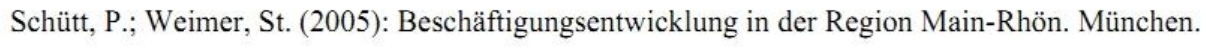

Abb. 1: Beispiel einem unbehandelten Referenzstring

```
<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>
```

Abb. 2: Referenz aus Abb. 1 nach der Verarbeitung mit Parscit in XML

5. Zusammenfassung

Die automatische Extraktion von Referenzen aus Textdokumenten ist keine triviale Aufgabe. Parscit konnte erfolgreich mit weiteren Tools innerhalb einer Prozesskette gekoppelt werden und zeigte in unserem Sample eine zufriedenstellende Extraktions- und Erkennungsleistung für Referenzen. Insgesamt konnten aus 80 % der Dokumente automatisiert Referenzen extrahiert werden. Es ist weiterhin geplant, die entwickelte Prozesskette zur automatischen Referenzextraktion mit größeren Volltextkorpora zu testen und Parscit entsprechend anzupassen.

Literaturverzeichnis

Councill, Isaac G., Giles, C. Lee, & Kann, Min-Yen (2008). ParsCit: An open-source CRF reference string parsing package. In Proceedings of the Language Resources and Evaluation Conference.

Giles, C. Lee, Bollacker, Kurt D., & Lawrence, Steve (1998). CiteSeer: An Automatic Citation Indexing System. In Proceedings of ACM DL'1998.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.