Citation Segmentation from Sparse & Noisy Data:

An Unsupervised Joint Inference Approach with Markov Logic Networks

Dustin Heckmann, Anette Frank
Institut für Computerlinguistik
Universität Heidelberg
{heckmann,frank}@cl.uni-heidelberg.de

Matthias Arnold, Peter Gietz, Christian Roth
Exzellenzcluster "Asien and Europa in a Global Context"
Universität Heidelberg
{arnold,croth}@asia-europe.uni-heidelberg.de, gietz@daasi.de

Bibliographische Daten sind das Rückgrat der wissenschaftlichen Forschung. Liegen sie nur in Druckform vor, können sie ihr Potential nicht voll entfalten. Erst ihre Speicherung in bibliographischen (Online-) Datenbanken öffnet effiziente Suchmöglichkeiten für den zeitgemäßen und breitgefächerten Einsatz in internationalen Forschungsgemeinschaften. Zu diesem Zweck ist es erforderlich, die Materialien zu digitalisieren und die den bibliographischen Referenzen innewohnende Struktur automatisch zu erkennen, indem einzelne Felder (z. B. Autor, Titel, Erscheinungsort) extrahiert werden.

In diesem Paper präsentieren wir ein Verfahren für die Zitationsanalyse auf spärlichen und verrauschten OCR-Daten[1]. Als Datenbasis nutzen wir den *Turkologischen Anzeiger* (TA)[2], ein wichtiges Referenzwerk für die Turkologie und die Osmanistik. Der *Turkologische Anzeiger* ist eine systematische Bibliographie in 28 Bänden, die bisher nur in gedruckter Form verfügbar war. Er umfasst Einträge in vielen verschiedenen Sprachen, einschließlich Transkriptionen aus dem

[1] Das hier vorgestellte Projekt ist eine Zusammenarbeit mehrerer Institutionen der Universität Heidelberg, dem Seminar für Sprachen und Kulturen des Vorderen Orients (Islamwissenschaft), dem Institut für Computerlinguistik und der Heidelberg Research Architecture, der Abteilung Digitale Geisteswissenschaften am Exzellenzcluster "Asia and Europe in a Global Context". Gestützt auf eine Vereinbarung mit dem Redaktionsausschuss des TA, die Datenbank als Open Access Ressource der Öffentlichkeit zur Verfügung zu stellen, formierte sich eine Arbeitsgruppe, die beim Exzellenzcluster erfolgreich Mittel zur Umsetzung einwerben konnte. Die Ergebnisse des Projektes sowie die Online Datenbank selbst können unter dieser Adresse aufgerufen werden: http://kjc-fs2.kjc.uni-heidelberg.de:8000/.

[2] Siehe Hazai & Kellner-Heinkele (1975) sowie die Online Datenbank.

Arabischen und aus Sprachen mit kyrillischem Alphabet, wobei selbst einzelne Einträge aus Abschnitten in verschiedenen Sprachen bestehen können.

Bestehende Ansätze für die Zitationsanalyse stützen sich auf sprachspezifische lexikalische Daten und Mehrfachvorkommen von Zitationen in Online-Publikationsverzeichnissen. Bei der Verarbeitung mehrsprachiger Daten wird die Nutzung sprachspezifischen Wissens jedoch erschwert. Darüber hinaus enthalten in sich abgeschlossene Datenquellen, wie gedruckte Bibliographien, naturgemäß wenige wiederkehrende Referenzen, was die Berufung auf Datenredundanz verhindert.

Mit den in hoher Auflösung digitalisierten und mit der OCR Software Abbyy FineReader Professional in Volltext umgewandelten Zitationen des *Turkologischen Anzeigers* verhielt es sich genauso. Der Mangel an Redundanz in den Zitationen, Erkennungsfehler in der OCR-Volltextumwandlung sowie Inkonsistenzen in der Zitationsstruktur der Druckausgabe erschwerten die Anwendung bestehender statistischer Ansätze für Zitationsanalyse.

Nach Beispiel von Poon & Domingos (2007) stützt sich unser Verfahren auf Markov Logic Networks (MLN), einem Framework für Statistical Relational Learning, das Prädikatenlogik mit probabilistischer Modellierung verbindet (Richardson & Domingos, 2006). Die Formulierung in Prädikatenlogik bietet hohe Ausdrucksstärke und Flexibilität. Dadurch kann die Zitationsanalyse auf die besonderen Konventionen einer Bibliographie -- in unserem Fall dem *Turkologischen Anzeiger* -- zugeschnitten werden. MLNs können auf der Basis annotierter Daten in einem überwachten Lernverfahren trainiert werden. Sie können aber mit Hilfe manuell gewichteter Regeln auch in einem nicht-überwachten Verfahren angewandt werden. Bei fehlenden Trainingsdaten und Mangel an Datenredundanz in bibliographischen Quellen bieten MLNs ein attraktives Framework für die Zitationsanalyse durch Verwendung unüberwachter Verfahren.

In unserer Arbeit präsentieren wir ein Verfahren für Zitationsanalyse mittels Markov Logic Networks und Joint Inference. Wir wenden dieses auf eine umfangreiche mehrsprachige Bibliographie an, die aus verrauschtem OCR-Output gewonnen wurde. Die zu bewältigenden Probleme beinhalten insbesondere Rauschen durch OCR-Fehler, die Mehrsprachigkeit der einzelnen Einträge, komplexe Zitationsstrukturen, Inkonsistenzen in den Zitationen, sowie Mangel an Redundanz. Unser Joint Inference Verfahren erweitert den Ansatz von Poon & Domingos (2007), indem Redundanz auf Feldebene ausgenutzt wird. Dadurch sind wir in der Lage, dem Fehlen redundanter Zitationen beizukommen.

Die Ergebnisse einer einfachen MLN Formalisierung für einzelne Datensätze übertreffen sowohl die guten Referenzdaten der traditionellen Herangehensweise mit auf regulären Ausdrücken basierendem Parsing, als auch die des überwachten statistischen Ansatzes unter Nutzung von Conditional Random Fields (CRF). Werden auch joint references auf *Entitäts*- und *Feldebene* einbezogen, steigt die

Perfomanz bei recall und precision, wobei Joint Inference auf Feldebene die besten Ergebnisse produziert. Dabei nutzt unser Verfahren manuell gewichtete Regeln und ist komplett unüberwacht.

Unsere Evaluationsergebnisse zeigen, dass unsere MLN Formalisierung angesichts besonderer Herausforderungen bei der Analyse von Zitationen aus einer digitalisierten Bibliographie sowohl regelbasierte als auch moderne statistische Verfahren übertreffen. Auf unseren Testdaten erreichen wir einen F_1-Wert von 88% für exakte Übereinstimmung von Feldern, was einen Zuwachs von 21.9% gegenüber einem CRF-basierten Vergleichssystem darstellt.

Zusammenfassend kann festgestellt werden, dass wir im Gegensatz zu früheren Arbeiten Datensätze aus dem Bereich der Digitalen Geisteswissenschaften adressieren, die verrauscht und nur gering strukturiert sind. Dabei erweitert unsere Methode den Ansatz von Poon & Domingos (2007), indem wir *Joint Inference* auf *Feldebene* einsetzen. Dadurch sind wir in der Lage, ohne wiederkehrende Referenzen und mit verrauschten Daten umzugehen. Unser Verfahren ist komplett unüberwacht und benötigt keine annotierten Trainingsdaten. Das von uns erarbeitete Regelwerk kann auch auf die Verarbeitung anderer Bibliographien oder digitale Quellen, wie historische Wörterbücher oder Enzyklopädien, angewandt werden.

Die Ergebnisse unseres Projektes stellen wir auf der Website *Turkology Annual Online*[3] der Öffentlichkeit zur Verfügung. Das Interface bietet sowohl Such- und Browse-Funktionalitäten für den TA an. Bibliographische Subfelder (wie Titel oder Autor) sind explizit gemacht und können als Suchkriterien oder zum Sortieren von Ergebnissen genutzt werden, Querverweise sind als Hyperlinks ausgegeben. Referenzen können selektiert und in verschiedene bibliographische Formate wie beispielsweise BibTeX exportiert werden.

Abbildung 1 zeigt einen Beispieldatensatz, wie er im Webinterface dargestellt wird.
Abbildung 2 zeigt ein Beispiel für einen Datensatz im Originalscan.

[3] http://kjc-fs2.kjc.uni-heidelberg.de:8000/

Hazai, G. and Kellner-Heinkele, B. eds. (1975ff). *Turkologischer Anzeiger*, Universität Wien. Institut für Orientalistik and Universität Wien. Orientalisches Institut. Available at: http://orientalistik.univie.ac.at/forschung/publikationen/turkologischer-anzeiger, [Accessed: July 19, 2013].

Poon, H. and Domingos, P. (2007). Joint Inference in Information Extraction. In *Proceedings of the Twenty-Second National Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press.

Richardson, M. and Domingos, P. (2006). Markov Logic Networks. *Machine Learning*, 62(1), pp.107–136.

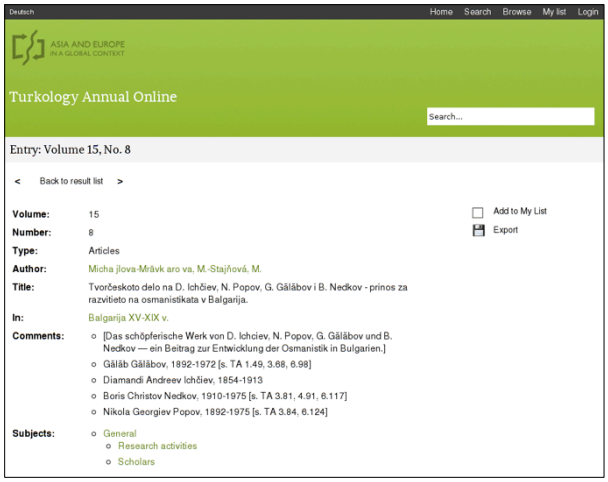

Abbildung 1: Turkologischer Anzeiger Online: Darstellung eines einzelnen Datensatzes.

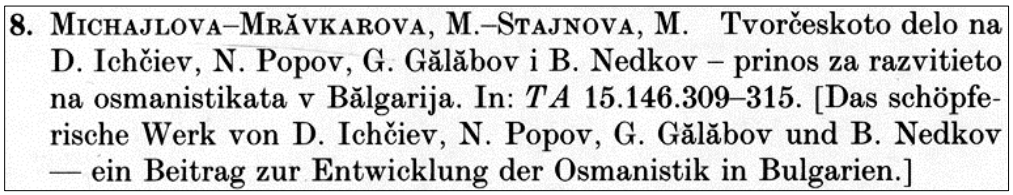
8. MICHAJLOVA–MRĂVKAROVA, M.–STAJNOVA, M. Tvorčeskoto delo na D. Ichčiev, N. Popov, G. Gălăbov i B. Nedkov – prinos za razvitieto na osmanistikata v Bălgarija. In: *TA* 15.146.309–315. [Das schöpferische Werk von D. Ichčiev, N. Popov, G. Gălăbov und B. Nedkov — ein Beitrag zur Entwicklung der Osmanistik in Bulgarien.]

Abbildung 2: Turkologischer Anzeiger: Beispiel für einen Datensatz im Originalscan.