Interaktive Wissensextraktion und Wissenssuche

Michael Stoll[2], **Katja Hose**[1], **Steffen Metzger**[1], **Ralf Schenkel**[2]
[1]Max-Planck-Institut für Informatik, Saarbrücken, Germany
[2]Universität des Saarlandes, Saarbrücken, Germany

Abstract

Die hochwertige Annotation von Entitäten und ihren Beziehungen ist ein Schlüssel zur Erschließung großer Textmengen, erfordert aber eine Kombination von effizienten maschinellen Verfahren und manueller Überprüfung. Darüber hinaus ist auch eine ausdrucksstarke Suche, die über eine reine Schlüsselwortsuche hinausgeht, von entscheidender Bedeutung. Dieser Artikel stellt die Knowledge Workbench vor, die automatische Techniken zur Entitätenerkennung und Informationsextraktion zur Verfügung stellt. Durch Interaktion mit dem Benutzer werden die Verfahren dabei inkrementell für das jeweilige Anwendungsgebiet optimiert. Darüber hinaus erlaubt sie eine strukturierte Suche von Texten auf Basis extrahierter Informationen.

1 Einleitung

Die steigende Anzahl an verfügbaren Informationsquellen macht im heutigen Informationszeitalter eine effiziente Auswertung von Rohdokumenten zur Verarbeitung des eigentlich essentiellen Wissens, das darin textuell abgebildetet wurde, immer wichtiger. Im Bereich der Geisteswissenschaften beispielsweise werden Texte manuell mit Annotationen angereichert. So lässt sich z.B. der Ort *"Sarabriga"*, der in historischen Texten genannt wird, mit der heutigen Bezeichnung des Ortes *"Saarbrücken"* verknüpfen. Weiteres Wissen, wie z.B. dass es sich bei Saarbrücken um die Hauptstadt des Saarlandes handelt, lässt sich über eine Verknüpfung mit externen Ontologien einbinden; auch das Vorkommen solcher Fakten wird oft in Texten annotiert. Die existierenden Werkzeuge für Geisteswissenschaftler sind zumeist auf das Annotieren von Texten in aufwendiger und zeitintensiver Handarbeit beschränkt, so dass nur wenige Texte umfassend annotiert werden können. Werkzeuge zum Modellieren von Ontologien sind nicht transparent in den Annotationsprozess eingebunden. Um mit der enormen Flut an historischen und aktuellen Texten umgehen zu können, müssen daher computergestützte Methoden eingesetzt werden, die Domänenexperten Hinweise auf essentielle semantische Informationen geben.

Hierbei bietet es sich an, Methoden aus dem Gebiet der Wissensextraktion zu verwenden. Dabei wird eine eindeutige Erkennung im Text referenzierter Entitäten sowie eine einheitliche Abstraktion von im Text enthaltenen Beziehungen zwischen diesen Entitäten angestrebt. Betrachten wir z.B. die Textfragmente *"Universität des Saarlandes"*, *"UdS"* und *"Saarland University"* so stellen diese verschiedene Namen derselben eindeutigen Entität `UdS` dar. Auf dieser Basis kann Wissen über Beziehungen zwischen solchen Entitäten extrahiert werden. Formal kann dieses Wissen etwa in Form von RDF-Tripeln erfasst werden. Das Wissen, dass die Universität des Saarlandes in Saarbrücken liegt, entspricht dabei dem RDF-Tripel (`UdS`, `liegtIn`,`Saarbrücken`). Die Kombination mehrerer solcher Tripel ergibt eine Ontologie, die zusätzlich mit Typinformation und Regeln angereichert werden kann.

Auch wenn moderne Extraktionssysteme eine annehmbare Genauigkeit (Precision) erzielen können, ist die erzielte Qualität für eine wissenschaftliche Weiterverwendung der annotierten Entitäten und Fakten in der Regel nicht ausreichend. Eine manuelle Überprüfung der generierten Annotationen ist daher oft unerlässlich. Zudem leidet bei hochpräzisen Systemen oft die Ausbeute (Recall), so dass beispielsweise viele Beziehungen gar nicht gefunden werden. Desweiteren muss in der Regel Vorwissen, z.B. über die möglichen Namen einer Entität, in maschinenlesbarer Form angegeben werden, was insbesondere in großen Domänen vorab kaum vollständig zu leisten ist.

In diesem Artikel stellen wir die *Knowledge-Workbench* vor, die die Lücke zwischen automatischer Extraktion und manueller Annotation schließt, indem sie manuelle Überprüfung und Korrektur sowie die Eingabe von Vorwissen in einem iterativen Prozess vereint. Die Knowlegde-Workbench bindet Verfahren zur automatischen Entitätenerkennung und Beziehungsextraktion ein und präsentiert einem menschlichen Experten das automatisch erkannte Wissen zur Überprüfung und Erweiterung. Es bleibt dabei nicht bei einem einzelnen Durchlauf durch diesen Prozess, im Gegenteil ergibt sich ein Kreislauf von automatischer Erkennung und Korrektur der Ergebnisse. Diese Interaktion mit dem System verbessert die Güte des Extraktionsprozesses. Das Erlernen von nötigem Vorwissen wird direkt in die Nutzung des Tools integriert.

Durch die zur Verfügung gestellte Suchfunktionalität werden Texte nicht nur auf Basis einer Schlüsselwortsuche suchbar, sondern auch auf Basis der darin enthaltenen Fakten oder Entitäten, da die Dokumente mit diesem in Ontologien abgelegten Wissen verknüpft sind.

2 Arbeitsoberfläche

Aus Anwendersicht bietet die Knowledge Workbench insbesondere zwei Funktionalitäten zur Interaktion: Wissensextraktion und Wissenssuche.

2.1 Wissensextraktion

Der Benutzer kann in einem ersten Schritt auswählen, welche Ontologie als Vorwissen verwendet werden soll (Abb. 1 links oben: „Vorwissen wählen...“). Notwendig ist zudem

die Angabe des Verzeichnisses, in dem sich die zu untersuchenden Dokumente befinden. Alle Dateien des angegebenen Verzeichnises werden in einen Dateibrowser geladen, in dem der Benutzer diejenigen Dokumente auswählt, in denen Entitäten und/oder Fakten erkannt werden sollen.

Neues Wissen wird in zwei Schritten erzeugt. Im ersten Schritt wird wahlweise (Abb. 1 links unten: „Aktion“) eine automatische Entitäten-/Fakten-Erkennung durchgeführt, deren Ergebnisse dann in einem zweiten Schritt manuell inspiziert und bearbeitet werden. Durch die Unterscheidung in bestätigte und unbestätigte Entitäten/Fakten kann die Bearbeitung jederzeit unterbrochen und zu einem späteren Zeitpunkt fortgesetzt werden.

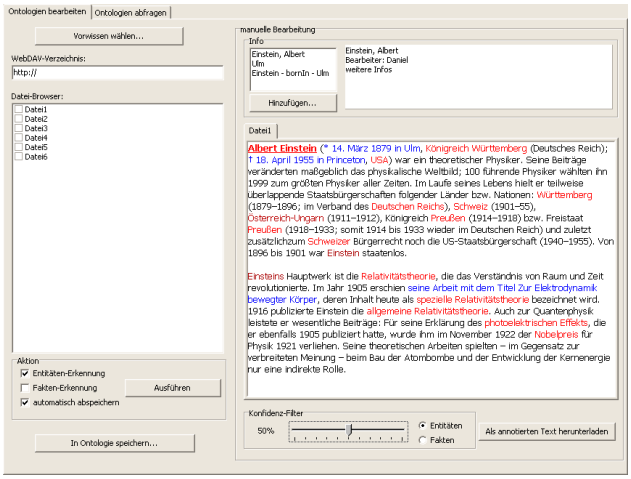

Abbildung 1: Prototyp der Knowledge Workbench

Das Ergebnis der Faktenerkennung wird dem Benutzer in Form einer Liste, die alle erkannten Fakten mitsamt der jeweils zugrundeliegenden Dokumente enthält, präsentiert. Durch entsprechende Sortierung ist eine schnelle Übersicht über mehrfach erkannte oder inkonsistente Fakten in verschiedenen Dokumenten gegeben. Alternativ können alle erkannten Fakten pro Dokument aufgelistet werden.

Die manuelle Inspektion (Abb. 1 rechts Mitte: „manuelle Bearbeitung“) einzelner Dokumente bietet zum einen die Möglichkeit, Entitäten bzw. Fakten nach einem zuvor automatisch zugewiesenen Konfidenzwert („Konfidenz-Filter“) zu filtern – dieser Konfidenzwert gibt die Sicherheit an, mit der der Fakt bzw. die Entität vom System korrekt erkannt wurde. Zum anderen werden gefundene Entitäten und Fakten farblich hervorgehoben und können vom Benutzer bearbeitet werden. Dazu gibt es ein eigenes Info-Fenster (Abb. 1 rechts oben: „Info“), das Fakten, Entitäten sowie weitere Metainformationen wie etwa Konfidenz, Bearbeiter etc. in einer Übersicht enthält. Ebenso können weitere, neue Textstellen markiert und Entitäten bzw. Fakten dazu angegeben werden. Das gewonnene Wissen kann entweder in einer neuen Ontologie oder in einer annotierten Textdatei gespeichert werden.

2.2 Wissenssuche

Die Knowledge-Workbench erlaubt strukturierte Anfragen (z.B. SPARQL) an die erzeugten Ontologien. Die Ergebnisse werden in Textform angezeigt und dienen zudem als Startpunkt(e) für eine grafische Navigation durch das Wissen der gewählten Ontologie. Dabei wird Ontologiebrowsing mit Document Retrieval kombiniert, so dass einerseits die Ontologie durchsucht werden kann und andererseits Dateien, die bestimmte Entitäten oder Beziehungen beinhalten, gefunden werden können. Hierfür verwenden wir das in [Elbassuoni *et al.*, 2010] vorgestellte Framework.

3 Architektur

Die *Knowledge Workbench* besteht aus den drei in Abbildung 2 dargestellten Komponenten *Extraktionsworkflowkontrolle*, *Wissenssuche* und *Wissensexport*, die jeweils einzelnen Nutzeranwendungen entsprechen. Zudem ist die Workbench auf zwei externe Komponenten angewiesen: ein Extraktionssystem wie z.B. SOFIE [Suchanek *et al.*, 2009] und ein Ontologiestore wie z.B. RDF-3X [Neumann and Weikum, 2010]. Sollen Dokumente bearbeitet werden, übergibt die *Extraktionsworkflowkontrolle* die Daten dem Extraktionssystem, lässt die Ergebnisse durch den Nutzer verifizieren, korrigieren und erweitern, und startet gegebenenfalls eine erneute Analyse unter Beachtung des neuen Wissens. Das für die Extraktion nötige und vom Benutzer ergänzte Hintergrundwissen kann je nach Domäne sehr groß werden und wird daher in einem performanten Backend, dem Ontologiestore, verwaltet. Das extrahierte Wissen kann über die *Wissensexport*-Komponente in unterschiedlicher Form exportiert werden, nämlich als eigenständige Ontologie oder als annotierte Textdatei. Beide Varianten erlauben die Weiterverarbeitung mit verschiedenen externen Hilfmitteln, um z.B. über Texteditoren den Text mit weiteren Annotationen anzureichern oder den durch die Extraktion aufgespannten semantischen Raum mit Visualisierungstools zu erforschen. Die Workbench bietet jedoch mit der *Wissenssuche* auch eine eigene Suche auf dem im Ontologiestore gespeicherten Wissen.

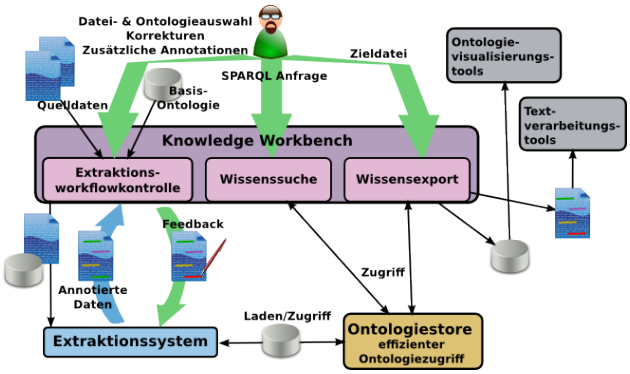

Abbildung 2: Architektur

4 Ausblick

Die Knowledge-Workbench ist ein erster Schritt zur Unterstützung hochwertiger manueller Textannotation durch automatische Verfahren. Sie wird derzeit in TextGrid integriert, einem geisteswissenschaftlichen Gridprojekt. Funktionalitäten und Usability des Systems werden durch ständige Interaktion mit den Anwendern weiter verbessert.

Literatur

[Elbassuoni *et al.*, 2010] Shady Elbassuoni, Katja Hose, Steffen Metzger, and Ralf Schenkel. ROXXI: Reviving witness dOcuments to eXplore eXtracted Information. *Proceedings of the VLDB Endowment*, 3(1-2):1589–1592, 2010.

[Neumann and Weikum, 2010] Thomas Neumann and Gerhard Weikum. The RDF-3X engine for scalable management of RDF data. *The VLDB Journal*, 19(1):91–113, 2010.

[Suchanek *et al.*, 2009] Fabian Suchanek, Mauro Sozio, and Gerhard Weikum. SOFIE: A self-organizing framework for information extraction. In *WWW*, pages 631–640, 2009.