Technischer Bericht der AG Texttechnologie

Number: TTLab-TR-2016-12

CEDIFOR-Forschungs-Workflow

AG Texttechnologie
Goethe-Universität Frankfurt am Main
Fachbereich für Informatik und Mathematik
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
Prof. Dr. Alexander Mehler

Workflow zur Arbeit mit Bildern und Text-Bild-Aggregaten

Ansprechpartner: Prof. Dr. Alexander Mehler, Giuseppe Abrami

Zusammenfassung
Workflow zur Annotation und Analyse von Bilddateien und Text-Bild-Beziehungen mit dem Ressourcen-Management und dem OWLnotator. Der Workflow und die involvierten Ressourcen werden beispielsweise im assoziierten CEDIFOR-Projekt Computational Historical Image Semantics eingesetzt.

Inhaltsverzeichnis

1. Hypothesenbildung

Formulierung einer Forschungsfrage, welche mittels einer Bild-Datenbank bzw. einer Datenbank von Bild-Text-Aggregaten untersucht werden soll.

2. Datengewinnung

Identifikation oder Erstellung einer Sammlung von Bildern bzw. multimedialen Dokumenten, die im Sinne der Frage geeignet ist. Uni- und multimediale Korpora werden mittels des Ressourcen-Management verwaltet. Das Ressourcen-Management erlaubt ein ausführliches Rechte-Management. Über diese Grundfunktionen hinaus bietet das Ressourcen-Management für Bilddaten spezialisierte Methoden an: beispielsweise können Bilder segmentiert, mittels einer generischen Benutzerschnittstelle oder einer REST-API skaliert und basierend auf Benutzer- und Gruppenrechten gezielt freigegeben werden. Im Falle der Bild-Segmentierung werden die Segmente als eigenständige Bild-Objekte dem System automatisch hinzugefügt, bleiben aber auf ihr Ausgangsbild (innerhalb der Datenbank wie auch innerhalb der Webschnittstelle) bezogen. Das Original wird also nicht verändert.

3. Datenmodellierung

Modellierung einer Ontologie im Kontext der Forschungsfrage bzw. des Erkenntnisinteresses. Grundlage für die spätere Datenannotation ist ein Datenmodell, welches zuvor definiert werden muss. Da die Annotationen mit dem OWLnotator erfolgen, ist ein Datenmodell im Format OWL/RDF erforderlich. Konzepte des Forschungsinteresses werden in *Klassen*, die Beziehungen zwischen den Klassen als `ObjectProperties` und die Attribuierung der Klassen als `DataProperties` modelliert. Wenn bereits bestehende Ontologien für die gewünschten Forschungsfragen vorliegen, können diese nachgenutzt oder erweitert werden (siehe Sektion 4). Eigene Ontologien können vom Anwender mit Unterstützung der freien Software Protegé erstellt und anschließend im OWLnotator hochgeladen und instantiiert werden. Durch die Datenmodellierung mittels OWL/RDF wird die Wissenschaftlerin/der Wissenschaftler im nachfolgenden Schritt der Datenannotation maßgeblich unterstützt.

4. Datenannotation

Annotation der Phänomene/Forschungsgegenstände, welche untersucht werden sollen bzw. Instanziierung der gemäß Schritt 3 erstellten Ontologie. Die Annotation soll nach klaren Regeln erfolgen, die in einer

Annotationsanleitung festgelegt sind. Die Annotationsanleitung folgt einem flexiblen Schema, das zu Beginn des Annotationsprozesses festzulegen und auszuwählen ist (siehe Schritt 3). Alle Bilder und Text-Bilddaten können im Rahmen des Ressourcen-Managements durch beliebige Annotationen angereichert werden. Hierzu dient der OWLnotator, der die flexible Annotation ontologiebasiert unterstützt. Hierzu muss zunächst eine Ontologie ausgewählt werden, wie sie zuvor in Schritt 3 ausgewählt oder erstellt wurde. Der OWLnotator kann beliebige Ontologien im OWL/RDF-Format aufnehmen, interpretieren und (nach-)nutzen und erlaubt Annotatoren die Annotation ihrer Daten basierend auf den in der Ontologie festgelegten Restriktionen. Annotationen werden manuell erstellt oder automatisiert über eine CSV-Datei eingelesen. Die Spalten der CSV-Datei beziehen sich auf die *Relationen* bzw. *Klassen* der jeweiligen Ontologie. Zur Annotation von textuellen Daten kann beispielsweise WebAnno oder der TextImager verwendet werden.

5. Experimentelle Nachnutzung der Daten

Für textuelle Daten stellt der TextImager eine Vielzahl an automatischen Vorverarbeitungs- und Visualisierungsmethoden zur Verfügung, die zur Aufbereitung der Textbestandteile von multimedialen Dokumenten verwendet werden können. Gleichzeitig kann auch die ImageDB als Komponente des Ressourcen-Management zur graphischen Visualisierung der Bildbestände und ihrer Relationen (basierend auf den Annotationen mittels des OWLnotator) verwendet werden. Auf diese Weise können Text-Bild-Bestände schließlich webbasiert traversiert werden. Ferner sind alle annotierten Daten und Modelle mittels der zugrundeliegenden Graphdatenbank webbasiert für Suchanfragen verfügbar.

6. Datenanalyse

Die Datenanalyse besteht in der Regel aus deskriptiver Statistik, analytischer Statistik oder *Machine Learning*. Das Ressourcen-Management bzw. der OWLnotator halten bereits grundlegende deskriptive Angaben bereit und können – je nach vorliegenden Zugriffsrechten – jederzeit durch die zugrundeliegende REST-Schnittstelle abgefragt werden. Darüber hinaus wird ein Datenexport für externe Programme unterstützt, mit denen die exportierten Daten weiter analysiert werden können. Je nach zugrundeliegender Ontologie kann der Datenexport unter Umständen nicht vollständig auf eine zweidimensionale Struktur abgebildet werden (Serialisierung). Hierzu müssen unter Rückgriff auf entsprechende Bezugsgrößen möglicherweise besser geeignete Exporte durchgeführt werden.

7. Metadaten und Verlinkung

Die Spezifikation von Metadaten und die Verlinkung von Datensätzen können mittels des OWLnotator vorgenommen werden. Inhalte, die über das Ressourcen-Management verwendet werden, sind überdies über das Clarin Virtual Language Observatore auffindbar und stehen somit einer größeren Forschungsgemeinschaft zur Verfügung. Die Sichtbarmachung über CLARIN ist über das Rechtemanagement des Ressourcen-Management einstellbar.

8. Evaluation

Bestimmung der *Interrater-Agreement* und/oder Evaluation des Modells auf Testdaten. Unabhängig von der technologischen Absicherung des Workflows stehen CEDIFOR-Mitarbeiter ggf. bei der Durchführung der einzelnen Arbeitsschritte beratend zur Seite.

Rückfragen und Kontakt

Bei Rückfragen wenden Sie sich bitte an Prof. Dr. Alexander Mehler oder an Giuseppe Abrami.