Anwendungen Linguistische Informatik – Seminar –

Sommersemester 2018
Übersicht

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Organisatorisches

Seminartermin: Donnerstag, 15:15 Uhr im Raum 3-14

Teil 1:

- Vorträge durch Mitarbeiter der Abteilung zu Forschungsthemen und Projektarbeiten
- Vorbereitung der Vorträge durch die Teilnehmer, Gruppengröße 1-2.

Teil 2:

- Vorträge durch Teilnehmer (je 45 Minuten)
- Erstellen der Ausarbeitung (10-20 Seiten je nach Anteil von Programmierarbeit und Gruppengröße)

Info: *http://asv.informatik.uni-leipzig.de/de/courses/243*

Termine für Teil 1

12.04.2018 Einführung und Themenvergabe

26.04.2018 U. Quasthoff: Computerlexikographie

xx.xx.2018 xxx

Themenvorschläge für Teil 2 (1)

Thema 1: Erstellung eines Textkorpus als Canonical Text Service. (Ansprechpartner: Jochen Tiepmar)

Thema 2: Analyse moderner Speech-APIs kommerzieller Anbieter (Thomas Eckart)

Thema 3: Automatisierter Journalismus: Textgenerierung (Uwe Quasthoff)

Thema 4: Unterstützung der Namenberatungsstelle: Statistik von Babynamen (Fabian Schmidt)

Thema 5: Zwei-Ebenen-Morphologie mit Verwendung des HFST-Toolkits (Maciej Sumalvico)

Thema 6: Generische Wikipedia-Biografie (Uwe Quasthoff)

Thema 7: Qualität der Satzvereinfachung bewerten (Lena Schiffer)

Themenvorschläge für Teil 2 (2)

Thema 8: Lexikographie: Multilinguale Verknüpfung der Wortschatz-Datenbanken mit Wörterbüchern (Uwe Quasthoff)

Thema 9: REALonline: Arbeiten mit Bildbeschreibungen (Thomas Efer)

Thema 10: Erstellen eines spaCy-Modells für POS-Tagging (Christian Kahmann)

Thema 11: Satzsegmentierung – Evaluierung verschiedener NLPTools Verfahren (Christian Kahmann)

Eigene Themenvorschläge sind bis 26.4.2018 willkommen.

Thema 1: Erstellung eines Textkorpus als Canonical Text Service

Recherchieren Sie verfügbare Datensätze und erstellen Sie einen per TEI/XML kodierten thematisch zusammenhängenden Textkorpus. Der Datensatz soll am Ende als ein Canonical Text Service publiziert werden. Als Domäne werden Videospiel-Komplettlösungen vorgeschlagen, es kann aber auch eine eigene - möglichst noch nicht abgedeckte - Domäne gewählt werden.

Betreuer: Jochen Tiepmar <jtiepmar@informatik.uni-leipzig.de>

Thema 2: Analyse moderner Speech-APIs kommerzieller Anbieter

Im zunehmenden Maße wird Spracherkennungs-Technologie von Endkunden genutzt. Kommerzielle Anbieter stellen dabei Schnittstellen zur Verfügung, die für die Kommunikation mit mobilen und stationären Endgeräten geeignet sind und auf deren Basis auch von externen Anbietern Anwendungen entwickelt werden können. Bekannte Beispiele für solche APIs sind Alexa Voice Service, Google Assistant API oder SiriKit.

Zur Bewertung dieser Systeme aus Entwicklersicht sind u.a. folgende Charakteristiken relevant und sollen im Rahmen des Seminars systematisch erhoben werden.

- Funktionsumfang der API: unterstützte Funktionalität, unterstützte (natürl.) Sprachen
- Entwicklerunterstützung: Entwicklungsumgebungen vorhanden? Unterstützung welcher Programmiersprachen? Veröffentlichungsprozesses einfach? API stabil?
- Auf welchen Plattformen/Geräten ist die Anwendung lauffähig?
- Einschränkungen durch Anbieter (inklusive Kosten)

Bei Interesse kann das Thema (mit einem stärkeren praktischen Bezug) auch im Rahmen einer Bachelor-/Masterarbeit weiter bearbeitet werden.

Thomas Eckart <teckart@informatik.uni-leipzig.de>

Thema 3: Robo-Journalismus: Automatisierte Textgenerierung

Moderne kommerzielle Systeme erstellen vollautomatisch ganze Artikel, die völlig unverändert in Newsmedien veröffentlicht werden. Momentan betrifft dies noch ausgewählte Bereiche wie Sportnachrichten, Wetterberichte oder Wirtschaftsmeldungen. Siehe z.B. *http://www.sueddeutsche.de/kultur/kuenstliche-intelligenz-robo-journalismus-1.3921660* mit Links zu Anbietern.

Finden Sie heraus,

- welche Anforderungen an die Eingaben gestellt werden;
- welche Methoden verwendet werden und wie Variabel die Ausgabe ist (siehe z.B. www.retresco.de/textgenerierung/
- welche Handarbeit für die Textgenerierung investiert wurde.

Welche Bereiche bieten sich als nächstes an?

Uwe Quasthoff <quasthoff@informatik.uni-leipzig.de>

Thema 4: Statistik von Babynamen

Die Namenberatungsstelle der Uni Leipzig verfügt über Babynamenlisten vieler Jahre, darunter die in Deutschland vergebenen Babynamen seit ca. 2010 in einer MySQL-Datenbank. Zusätzlich dazu geographische Informationen.

Daraus lassen sich interessante Fragestellungen bearbeiten, z.B.

- Trendvorhersagen für einzelne häufige Vornamen
- Untersuchungen zur Länge der Vornamen
- Regionale Besonderheiten bei Vornamen (Sepp, Peer, ..)
- Vergleich mit den Häufigkeitslisten der GfdS (regional anderer Schwerpunkt)

In Absprache mit der Namenberatungsstelle sind weitere Aufgabenstellungen denkbar.

Ansprechpartner am Institut: Fabian Schmidt fschmidt@informatik.uni-leipzig.de

Thema 5: Zwei-Ebenen-Morphologie mit Verwendung des HFST-Toolkits

Die Seminararbeit stellt das Thema "Zwei-Ebenen-Morphologie" aus anwendungsorientierter Sicht. Das Hauptziel ist der Aufbau eines einfachen morphologischen Analysierers fürs Deutsche. Dabei sollten folgende Themen ausgearbeitet und vorgestellt werden:

- Hauptidee und Motivation der Zwei-Ebenen-Morphologie
- Transduktoren als Werkzeug für Stringabbildungen; einfache Operationen auf Transduktoren (Disjunktion, Schnitt, Komposition etc.)
- Realisierung der Zwei-Ebenen-Morphologie auf Transduktoren und die Compose-Intersect-Operation
- Aufbau eines Lexikons fürs Deutsche inkl. Flexion, Derivation, Komposition (nicht vollständig; Abdeckung nur ausreichend zur Illustration des Vorgehens)
- Behandlung von morphophonologischen Phenomänen im Deutschen (z.B. "behandeln" vs. "*behandelen")
- Behandlung von Stammalternationen (Umlaut, Ablaut)

Maciej Sumalvico <sumalvico@informatik.uni-leipzig.de>

Thema 6: Generische Wikipedia-Biografie

Aus möglichst vielen Wikipedia-Biographien (Abschnitt: „Leben“) soll deren wiederkehrende Struktur extrahiert und beschrieben werden. Ziel ist nicht eine Überführung in eine dbpedia-ähnliche Struktur, sondern die Identifikation von Teilen sowie deren Beschreibung, z.B. durch wiederkehrende Zwischenüberschriften sowie in den Absätzen enthaltene typische Wörter. Evtl. Strukturierung nach Zeiträumen.

Angestrebt (z.B. in einer nachfolgenden Bachelor- / Masterarbeit) wird ein automatisches Verfahren, welches konkrete Biographien auf diese generische Biographie abbildet.

Betreuer: Uwe Quasthoff <quasthoff@informatik.uni-leipzig.de>

Thema 7: Qualität der Satzvereinfachung bewerten

Ein vorliegendes Verfahren vereinfacht längere Sätze, indem unwichtige Teile entfernt werden und bei Bedarf der verbleibende Satz umgestellt wird.

Beispiele:

Der zweite Angeklagte Silvio G. (27)

wurde freigesprochen. => *Der Angeklagte wurde freigesprochen.*

Derzeit schreibt sie eine Doktorarbeit. => *Sie schreibt eine Doktorarbeit.*

Dies ist momentan nicht fehlerfrei möglich (*Die Polizei hatte [den Täter] festgenommen.*, *Die Edelsteine leuchten in [vielen] Farben.*)

Die Aufgabe besteht darin, die Ergebnisse zu evaluieren, typische Fehler festzustellen und Verbesserungsvorschläge zu machen.

Lena Schiffer <schiffer@informatik.uni-leipzig.de>

Thema 8: Lexikographie: Multilinguale Verknüpfung der Wortschatz-Datenbanken mit Wörterbüchern

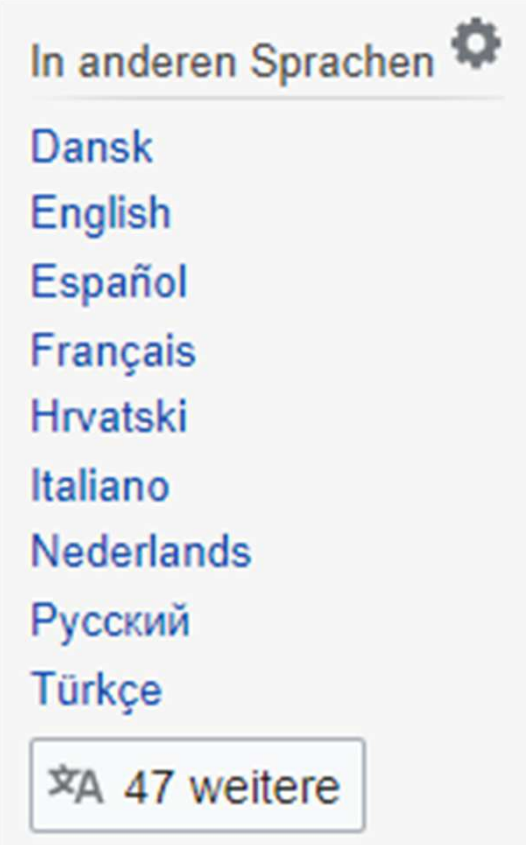
Die Stichwörter der Wortschatz-Datenbanken auf *http://corpora.uni-leipzig.de/* sollen multilingual verlinkt werden ähnlich wie bei Wikipedia in der linken Spalte. Dazu ist ein konzeptioneller Entwurf und ggf. eine prototypische Implementierung zu erstellen. Konzeptionelle Fragen sind:

Wie wird damit umgegangen, dass pro Sprache teilweise mehrere Übersetzungen vorhanden sind, die bei vollständiger Anzeige viel Platz benötigen?

In welcher Reihenfolge werden die Sprachen angezeigt? Alphabetisch oder nach Wichtigkeit?

Auf welche Datenbank einer Sprache zeigen die Links?

Exemplarische Daten für einige Sprachen sind vorhanden.



Uwe Quasthoff <quasthoff@informatik.uni-leipzig.de>

Thema 9: REALonline: Arbeiten mit Bildbeschreibungen

Am "Institut für Realienkunde des Mittelalters und der frühen Neuzeit" der Universität Salzburg wird seit 15 Jahren die Bilddatenbank "REALonline" aufgebaut. Darin befinden sich über 28.000 Aufnahmen der letzten 45 Jahre, u.A. von Gemälden, Zeichnungen und Buchillustrationen, den sogenannten "Realien" aus der Zeit des mitteleuropäischen Mittelalters (mit Fokus auf Österreich).

Das Portal wurde jüngst auf neue Technologien umgestellt und unterstützt eine feingliedrige facettierte Suche für alle erfassten Werke. Ein Datenexport des Portals liegt uns in Form einer 500MB großen JSON-Datei vor.

Die erfassten Bilder sind hinsichtlich der materiellen Beschaffenheit und künstlerischen Bearbeitung der Bildträger erschlossen:

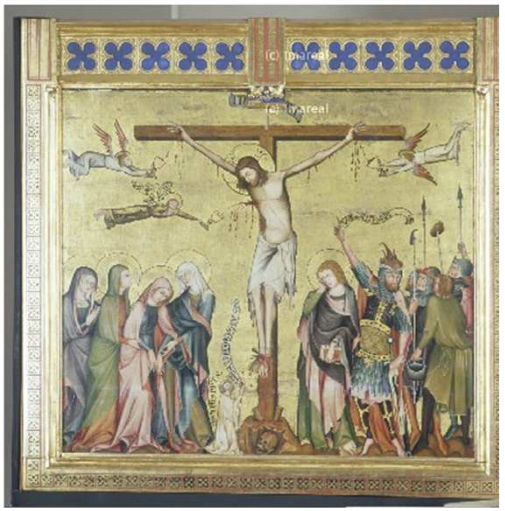
Daneben existieren auch Inhaltliche Beschreibungen, die z.T.

mit einer Ontologie von Motiven verknüpft ist:

Die Besonderheit der Datensammlung ist eine detaillierte und

strukturiert erfasste Beschreibung der jeweils abgebildeten

Szene als Graph:



Auf Grundlage dieses Datenbestandes sind verschiedene Themen denkbar, die einzeln oder in Kombination bearbeitet werden können:

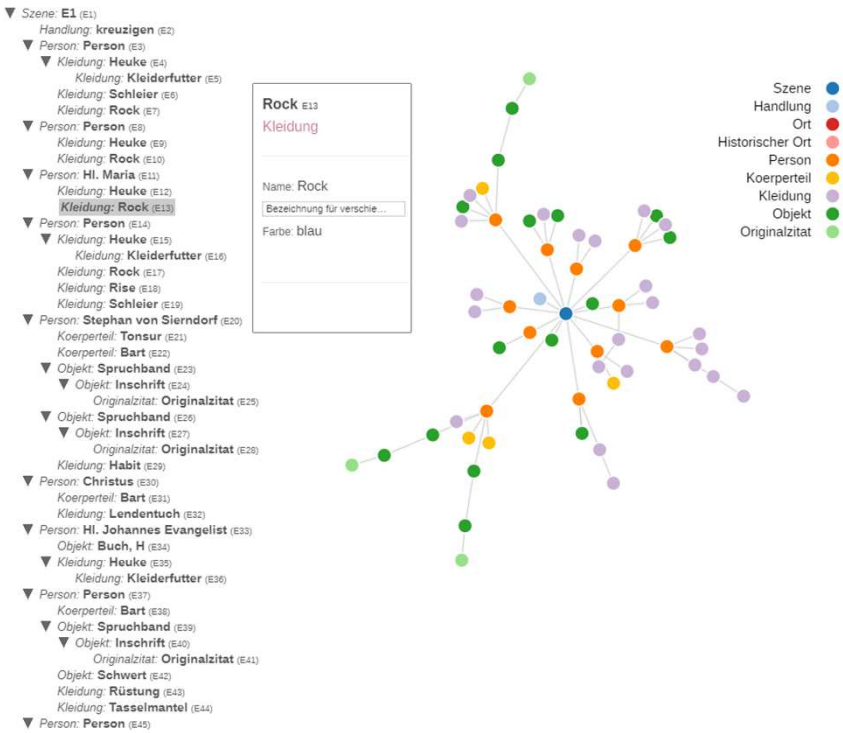
- Literaturrecherche und prototypische Erstellung eines Systems zur automatischen Bild-beschreibung (über NLG-Verfahren) aus dem Graphen

- Bilden von Ähnlichkeitsmaßen für Objekte und Szenen, die neben numerischen Attributen und Beschreibungen auch die Graphstruktur nutzen

- Recherche zu den Möglichkeiten automatischer Objekterkennung und -benennung (& ggf. -verknüpfung) aus Bilddaten auf der Basis des Datensatzes zum Training

- Erstellung eines protypischen Systems zur interaktiven Exploration der Graphen und Bilder der Sammlung

- Umsetzung eigener Ideen (nach Absprache)



Thomas Efer <efer@informatik.uni-leipzig.de>

Thema 10: Satzsegmentierung – Vergleich von Qualität, Performanz und typischen Problemen verschiedener NLPTools

Zu vergleichende NLP-Tools:

- spaCy: https://spacy.io/
- Quanteda: https://quanteda.io/
- OpenNLP: https://opennlp.apache.org/

Vergleich für die Sprachen Deutsch und Englisch, evtl. weitere.

Evaluierung mit Hilfe von Goldstandard aus Wortschatz-Daten.

Betreuer: Christian Kahmann <kahmann@informatik.uni-leipzig.de>

Thema 11: Erstellen eines spaCy-Modells für POS-Tags im Deutschen und Evaluierung dieses Modells

Die Qualität des gegebenen spaCy-Modells für die deutsche Sprache scheint unbefriedigend. Es besteht die Möglichkeit zum Erstellen von eigenen Modellen anhand von Trainingsdaten (https://spacy.io/usage/training).

Aufgaben:

- Erstellen von Trainingsdaten aus Wortschatz
- Erstellen eines spaCy Modells
- Vergleich von erstelltem Modell vs. Standard Modell

Betreuer: Christian Kahmann <kahmann@informatik.uni-leipzig.de>