Angewandte Linguistische Informatik
– Seminar –

Sommersemester 2015
Übersicht

Uwe Quasthoff

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

Organisatorisches

Seminartermin: Montag, 9:15 Uhr im Raum 3-12

Teil 1:

- Vorträge durch Mitarbeiter der Abteilung zu Forschungsthemen und Projektarbeiten
- Vorbereitung der Vorträge durch die Teilnehmer, Gruppengröße 1-3.

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/courses/176*

Termine für Teil 1

13.04.2015 Seminar

20.04.2015 U. Quasthoff: Computerlexikographie

27.04.2015 Abschlußberichte von Praktikumsgruppen aus
Wissens- und Contentmanagement

Themenvorschläge für Teil 2

Themenbereich 1: Wörter des Tages. (Ansprechpartner: Maciej Janicki)
Themenbereich 2: Crawling (Dirk Goldhahn)
Themenbereich 3: Wortschatz (Thomas Eckart, Uwe Quasthoff, Christoph Kuras)
Themenbereich 4: Multilingualität (Christoph Kuras)
Themenbereich 5: Metadaten (Thomas Eckart)
Themenbereich 6: Maschinelles Lernen für linguistische Probleme (Uwe Quasthoff)
Themenbereich 7: Theater: Dramen-Markup (Thomas Efer)
Themenbereich 8: Daten des Universitätsarchivs (Thomas Efer)
Themenbereich 9: Unterstützung der Namenberatungsstelle (Fabian Schmidt)

Eigene Themenvorschläge sind bis 20.4.2015 willkommen.

Bereich 1: Wörter des Tages

1.1. Signifikanzmaße und Auswahlkriterien für die *Wörter des Tages*

- Literaturrecherche über Signifikanzmaße für Wörter (frequency ratio, Poisson, tf-idf, velocity usw.)
- (Vorschlag:) Implementierung am besten in SQL oder R
- Auswahlkriterien: Wikipedia-Filterung? Musterbasierte Filtrierung (z.B. Datum)?

1.2. Visualisierung der *Wörter des Tages*

- einfache Variante: Word Cloud (Größe eines Wortes ~ Signifikanz, Farbe ~ Topic, verschiedene Sortierungen ausprobieren)
- zusätzliche Informationen: Timelines? Tageskookkurenzen? Topic-Entwicklung? Beispielsätze? Links zu relevanten Artikeln? Schlagzeilen?
- Literaturrecherche über Topic Modelle und ihre Visualisierung

1.3. NER auf den *WdT*

- ein Standard-Tool (z.B. Stanford) benutzen
- folgende Tabelle NER erzeugen: Satz_ID, start, end, type
- Literaturrecherche über NER + ein paar Anwendungsbeispiele auf WdT

Themenbereich 2: Crawling

Thema 2. CommonCrawl

Commoncrawl stellt Ergebnisse des eigenen umfangreichen Webcrawls auf seiner Seite zum Download bereit: http://commoncrawl.org/
Ziel wäre eine Analyse der Daten:
- Zusammensetzung des Crawls bezüglich TLDs, Sprachen
- Anzahl URLs pro Datei
- Durchschnittliche Dokumentenlänge, Satzanzahl....
- Vergleich der früher Ergebnisse eines Crawls mit den späteren desselben Crawls (bezüglich der zuvor genannten Faktoren)
- Download als warc-File oder als extrahierter Text möglich: Vergleich unseres HTML-Extractors mit dem dort verwendeten, um zu bestimmen, ob die dort verwendete Rohtextextraktion von guter Qualität ist oder unsere eigenen Tools zum Einsatz kommen müssten

Themenbereich 3: Wortschatz

Thema 3.1: Logfileanalyse Wortschatzportal

Das Wortschatzportal mit seinen hunderten Korpora ist eine wichtige Ressource für Sprachinteressierte. Im Rahmen einer Logfileanalyse soll typisches Nutzerverhalten und -interesse analysiert werden. Dazu gehören Analysen betreffend: -häufige Anfragen (Korpus/Wörter) -häufige Anfragemuster (Wort1 -> Wort2 -> Wort3) - Herkunft von Anfragen (geographische Verteilungen via IP-Adressen) - Identifizierung häufiger Nutzer etc.

Thema 3.2: Qualitätsbewertung der Beispielsätze

Allgemeine Qualitätsbewertung der Satzsegmentierung. Wie groß ist der Anteil der wohlgeformten Sätze? Was sind die häufigsten Fehlerklassen?

Wie sinnvoll ist das Ranking der Sätze nach GDEX? Welche Folgen hat das Verwerfen der am schlechtesten gerankten Sätze?

Themenbereich 4: Multilingualität

Thema 4: Integration von Wörterbuchdaten

Im Wortschatzportal sollen die monolingualen Ressourcen über Sprachgrenzen hinweg verknüpft werden. Dazu sollen freie Wörterbücher benutzt und aufbereitet werden.

Bisher aufbereitet: dbnary und JRC-Names

Themenbereich 5: Metadaten

Thema 5: Vokabularentwicklung zur Beschreibung von linguistischen Metadaten

Der Metadatenstandard CMDI sowie die Ontologie ISOcat wird von vielen Anbietern zur Beschreibung linguistischer Ressourcen (Korpora, Einzeltexte, Werkzeuge etc.) genutzt. Der modulare Aufbau dieses Standards erschwert allerdings die föderierte Nutzung der verteilt erstellten Dateien. Auf der Basis monatlicher Dumps der letzten drei Jahre soll die Entwicklung der verwendeten Vokabulare bei den verschiedenen Anbietern nachvollzogen werden. Der Fokus liegt dabei auf dem Vergleich des Vokabulars unterschiedlicher Ressourcenanbieter sowie der Frage wie sich Vokabulare über die Zeit verändern.

Bereich 6: Maschinelles Lernen

Thema 6: TiMBL für klassische Aufgabenstellungen der Linguistik

- Grundformreduktion
- Kompositazerlegung
- morphologische Zerlegung
- (und mehr)

Ziel ist die Schaffung einer Umgebung, die ein der jeweiligen Aufgabenstellung angepasstes einfacheres Datenformat für Trainings- und Testdaten ermöglicht und Aussagen über die Qualität der Ergebnisse zulässt.

Bereich 7: Theater: Dramen-Markup

Thema 7: Das XML-basierte Textencodingformat der TEI (http://www.tei-c.org/index.xml) bietet zahlreiche Tags und Strukturen um den Handlungsverklauf (Akt/- und Szenenstruktur, Liste handelnder Personen, Regieanweisungen, "Speeches", Prolog & Epilog, ...) zu Modellieren. Die letztendliche Umsetzung ist allerdings sehr frei geregelt und daher nicht immer einheitlich. Ziel der Seminararbeit ist es,

1. den technologischen und inhaltlichen Rahmen all dieser Dokumentenformate zu erkunden
2. die maximal vorhandene Erschließungstiefe der einzelen Stücke über mehrere Dramen-Sammlungen hinweg (unterschiedliche Sprachen, Zeitabschnitte & Formate) zu benennen und ggf. (falls der Aufwand dafür nicht zu hoch ist) Konvertierungsmöglichkeiten zu erfassen.
3. maschinenlesbare Regieanweisungen aus textuellen Originalstellen zu erzeugen, Bsp.: Wer ist momentan der Bühne, wenn es heißt "Enter Hamlet", "exeunt", "Heinrich (allein)", "bleeds heavily, sinks to the floor and dies", "rennt schreiend davon", "beide gehen nach rechts ab", "Re-enter Ghost", usw.

Bereich 8: Daten des Universitätsarchivs

Thema 8: Das Universitätsarchiv verfügt über eine Sammlung von dokument- und autorenbezogenen Metadaten, von denen ein Teil bisher nur im Konvolut digitalisiert wurde und noch nicht auf die einzelne Entität heruntergebrochen ist. Hieraus müssen zunächst strukturierte Daten erzeugt werden (über Parsing, reguläre Ausdrücke, kleine Scripts, ...)

Beispiel:

Abb, Edmund: geb. Trennfurt, Bez. Obernberg 8.6.1878-- Prüfungszeugnis 1902.- Abt, Otto: aus Saalfeld-- Reifezeugnis 1866, väterliche/vormundschaftliche Studiengenehmigung, Abgangszeugnis Universitätsgericht Leipzig 1873, Brief v. C. Brühus.- Achenbach, Rudolf: geb. Netra b. Eschwege-- Reifezeugnis 1873, …

Ausgabe:

Vorname: *Edmund*

Nachname: *Abb*

Geburtsort: *Trennfurt (Bezirk: Obernberg)*

Geburtsdatum: *1878-06-08*

Dazu erfasste Dokumente:

…

Themenbereich 9: Unterstützung der Namenberatungsstelle

Die Namenberatungsstelle verfügt über eine umfangreiche Vornamenstatistik. Die in Deutschland vergebenen Babynamen sind seit ca. 2010 in einer MySQL-Datenbank.

Viel mehr Daten liegen in anderen Formaten vor.

In Absprache mit der Namenberatungsstelle sind verschiedenste Aufgabenstellungen denkbar.

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