Softwaretechnik-Praktikum
2016
Datum: 7. Januar 2016

Gruppe: wrd16
Betreuer: René Speck
Tutor: Marius Brunnert

Team:
Tim Niehoff
Felix Lange
Philip Fritzsche
Matti Wilhelmi
Simon Kaleschke
Johannes Leupold

Recherchebericht

Inhaltsverzeichnis

1 Begriffe

1.1 Korpus

Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst sowie möglicher aus Metadaten, die diese Daten beschreiben und aus linguistischen Annotationen, die diesen zugeordnet sind.

1.2 Framework

Ein Framework (englisch für Rahmenstruktur) ist ein Programmiergerüst, das in der Softwaretechnik, insbesondere im Rahmen der objektorientierten Softwareentwicklung sowie bei komponentenbasierten Entwicklungsansätzen, verwendet wird. Ein Framework ist selbst noch kein fertiges Programm, sondern stellt den Rahmen zur Verfügung, innerhalb dessen der Programmierer eine Anwendung erstellt. Die Entwickler bauen das Framework in ihre eigene Applikation ein, und erweitern es derart, dass es ihren spezifischen Anforderungen entspricht.

1.3 Entität (Entity)

Eine Entität ist ein eindeutig identifizierbares, konkretes oder abstrakes Objekt der realen Welt oder unserer Anschauung.

1.4 Relation (Relationship)

Eine Relation (Beziehung) besteht zwischen zwei Entitäten und beschreibt eine statische Beziehung dieser Entitäten zueinander oder eine Interaktion miteinander.

1.5 Unified Modeling Language (UML)

UML ist eine grafische Modellierungssprache. Diese kommt in Form von verschiedenen Diagrammen in der Planungs- und Definitionsphase der Softwareentwicklung zum Einsatz. Dazu gehören zum Beispiel Klassendiagramme und Sequenzdiagramme.
UML-Diagramme werden in Strukturdiagramme, welche die Struktur eines Systems beschreiben, und Verhaltensdiagramme, welche das Verhalten und die Interaktionen eines Systems beschreiben, eingeteilt.

1.5.1 Klassendiagramm

Ein Klassendiagramm ist ein Strukturdiagramm für die objektorientierten Modellierung. In Klassendiagrammen werden im Modell existierende und interagierende Objekte mit ihren Eigenschaften und ihre Beziehungen in Form von Assoziation, Aggregation, Komposition und Generalisierung dargestellt. Klassendiagramme sind durch den UML-Standard definiert.

1.5.2 Sequenzdiagramm

Ein Sequenzdiagramm ist ein Verhaltensdiagramm, welches Interaktionen zwischen verschiedenen Entitäten darstellt. Dazu wird auf Lebenslinien der Entitäten ein Austausch von Nachrichten in der zeitlichen Abfolge dargestellt.

1.6 Entity-Relationship-Modell (ERM)

Das Entity-Relationship-Modell dient in der Definitionsphase der Beschreibung von für die Software bedeutsamen Entitäten und deren Relationen. Dazu bietet es Konzepte wie Attribute, um Informationen über Entitäten zu speichern, sowie spezielle Beziehungen für Generalisierung bzw. Spezialisierung und Aggregation oder Komposition. Das Entity-Relationship-Modell besteht aus einem Entity-Relationship-Diagramm sowie einer textuellen Beschreibung der Semantik des Modells.

1.7 Entity-Relationship-Diagramm

Ein Entity-Relationship-Diagramm dient der grafischen Darstellung eines Entity-Relationship-Modells. Alle Konzepte des ERM werden grafisch umgesetzt. Für ERDs existieren verschiedene Notationen, die am häufigsten gebrauchten sind die Chen-Notation (*Peter Chen, 1976*) und die im UML-Standard definierte Notation.

1.8 Uniform Resource Identifier (URI)

URIs sind ein eindeutiger Bezeichner einer Resource im Internet, welcher als Zeichenfolge kodiert wird, die aus ASCII-Zeichen besteht.

1.9 Internationalized Resource Identifier (IRI)

IRIs sind eine Erweiterung der URIs, wobei hier nicht nur ASCII-Zeichen zugelassen sind, sondern fast alle Unicode Zeichen.

1.10 Uniform Resource Locator (URL)

URLs sind eine Form der URI, mit der Ressourcen in Netzwerken lokalisiert werden und meist die zu verwendende Zugriffsmethode festgelegt wird.

1.11 Ontologie

Eine Ontologie ist eine formale Definition von Typen, Eigenschaften und Beziehungen von Objekten eines bestimmten Gebietes.

1.12 Semantisches Datenmodell

Ein Semantisches Datenmodell (SDM, englisch auch conceptual schema) ist im Rahmen der Datenmodellierung eine abstrakte, formale Beschreibung und Darstellung eines Ausschnittes der in einem bestimmten Zusammenhang (z. B. eines Projekts) 'wahrgenommenen Welt'. Zur Formulierung semantischer Datenmodelle existieren verschiedene Modellierungssprachen, von denen das Entity-Relationship-Modell das bekannteste ist.

1.13 DBPedia

DBpedia German stellt die strukturierten Informationen der deutschsprachigen Wikipedia frei zur Verfügung. Dabei ist dies ein Teil des internationalen DBpedia Projektes, wobei wir uns auf die Extraktion und Bereitstellung der Informationen der deutschsprachigen Wikipedia beschränken. Damit wird es ermöglicht, optimale strukturierte Informationen zu nutzen die auf Anwendungen für deutsche Benutzer zugeschnitten sind. Das Ziel ist die Integration der deutschen Informationen in eine DBpedia Linked Data Wolke, welche die nationalen Datensätze miteinander verknüpft.

2 Konzepte

2.1 Apache Lucene

Apache Lucene ist eine freie Programmbibliothek zur Volltextsuche und ein Projekt der Apache Software Foundation.

2.2 Apache Solr, Apache Lucene

Apache Solr (ausgesprochen: s̈olar") ist eine Opensource Enterprise Suchplattform, die in Java geschrieben wurde und zum Apache Lucene Projekt gehört. Die hauptsächlichen Features beinhalten Volltextsuche, Trefferhervorhebung, facettenreiches Suchen, Echtzeitindexierung, dynamisches Clustern, Datenbankintegration, NoSQL Features und Handhabung mit Rich Documents (wie z.B. Word oder PDF). Durch das Anbieten verteilter Suche und Indexnachbau ist Solr für Skalierbarkeit und Fehlertoleranz konzipiert. Es läuft als standalone Volltext Suchserver. Zudem benutzt es die Lucene Java Suchbibliothek für das Volltextindexieren und -suchen und umfasst HTTP/XML und JSON APIs, sodass es für die beliebtesten Programmiersprachen anwendbar ist.

2.3 BOA

BOA (BOotstrapping linked datA) ist eine iterative Bootstrapping Strategie, um RDF aus unstrukturierten Daten zu extrahieren. Während die meisten Wissensquellen im Data Web aus strukturierten oder semi-strukturierten Daten extrahiert werden, enthalten diese nur einen Bruchteil der Informationen, der im dokumentenorientierten Web zur Verfügung steht. Hinter BOA steht die Idee, das Data Web als Hintergrundwissen für die Extraktion natürlicher Sprachmustern zu benutzen, die im Data Web gefundene Prädikaten repräsentieren. Diese Muster werden benutzt, um Instanzwissen aus natürlichen Sprachtexten zu extrahieren. Dieses Wissen wird dem Data Web zurückgegeben, um den Kreis zu schließen. Die Herangehensweise wird durch zwei Datenmengen evaluiert, indem DBpedia als Hintergrundwissen benutzt wird. Die Resultate zeigen, dass mehrere tausend neue Fakten in einer Iteration mit hoher Genauigkeit extrahiert werden. Zudem wurde so das erste Repositorium natürlicher Sprachrepräsentationen aus im Data Web gefundenen Prädikaten erzeugt.

2.4 Named Entity Recognition

(kurz NER) ist ein Teilgebiet der Wissenextraktion. Hierbei werden meist unannotierte Textblöcke genommen und in annotierte Textblöcke umgewandelt. Dabei können im Text Personen, Unternehmen, Zeitangaben etc. aufgefunden werden. Die Ausgabe kann dann weiterverwendet werden.

2.5 Natural Language Processing

(kurz NLP) ist ein Teilgebiet der Informatik. Man beschäftigt sich mit der Interaktion zwischen Menschen und Computern. Durch Informationen über das Verhalten verschiedener menschlicher Sprachen (bspw. Groß- und Kleinschreibung) kann man Informationen für NER erhalten.

2.6 NER Tools

sind Programme, die an Daten NER durchführen, bspw. FOX.

2.7 NLP Interchange Format

(kurz NIF) ist ein Format, in dem NLP Daten gespeichert werden. Es kann z. B. in RDF umgewandelt werden und dann bei der NER benutzt werden.

2.8 FOX

ist ein von René Speck und Axel-Cyrille Ngonga Ngomo entwickeltes NER Tool (das u.a. drei andere NER Tools zusammenfasst) mit hoher Effizienz. Es unterstützt verschiedene Eingabeformate (z.B. normalen Text oder eine URL) und kann die gefundenen Entitäten in verschiedenen Formaten ausgeben (u. a. Turtle, Json, XML) und mit Einträgen in bspw. dbpedia verknüpfen. Im Projekt kann FOX zum Finden von Entitäten benutzt werden, soll aber auch durch andere NER Tools ersetzbar sein.

2.9 Resource Description Framework (RDF)

ist ein Framework mit dem Informationen über Resourcen dargestellt werden können. Die Grundstruktur bilden dabei Tripel aus Subjekt, Prädikat und Objekt, wobei diese Resourcen darstellen und das Prädikat die Beziehung zwischen Subjekt und Objekt beschreibt. Die Menge dieser Tripel bildet einen (gerichteten) Graph, in dem Subjekte und Objekte die Knoten und Prädikate die Kanten sind. Resourcen werden durch IRIs identifiziert, welche an allen Positionen des Tripels auftreten können. Alternativ können Subjekte und Objekte auch blank nodes sein, Prädikate auch Literale.

2.9.1 Blank node

Blank nodes können in RDF als Platzhalter für unbekannte Resourcen verwendet werden oder um indirekte Aussagen über Resourcen ohne eigene IRI machen zu können. Sie sind vergleichbar mit Variablen, da sie keinen eigenen (bekannten) Wert haben.

2.9.2 Literal

Literale sind generische Werte, die keine IRIs sind. Sie können beispielsweise Strings oder Zahlen sein, wobei der Datentyp stets festgelegt sein muss. (Im RDF Standard sind erlaubte Datentypen definiert.) Zusätzlich kann einem String mittels Tag eine Sprache zugeordnet werden.

2.10 Terse RDF Triple Language (Turtle)

Turtle ist ein Format mit dem RDF Daten repräsentiert werden können. Dabei werden unter anderem Tripel direkt über die URIs angegeben, wobei URIs in spitzen Klammern geschrieben werden. Ebenso können Präfixe definiert werden, um URIs abzukürzen.

2.11 RDF/XML

RDF/XML ist wie Turtle ein Format für RDF Daten, in dem der RDF Graph als XML Dokument dargestellt wird.

2.12 SPARQL Protocol And RDF Query Language (SPARQL)

SPARQL ist eine Abfragesprache für RDF, die dazu dient Daten im RDF-Graph auszulesen oder zu verändern. Resourcen können dabei unter anderem über ihre Beziehungen zu anderen Resourcen oder über Muster gefunden werden.

2.13 OWL 2 Web Ontology Language

OWL 2 ist eine Beschreibungssprache für Ontologien, die RDF erweitert. Mittels OWL 2 können Klassen, Attribute und Beziehungen zwischen diesen definiert werden. Eine solche Beschreibung ist selbst wieder ein RDF Graph, wobei RDF/XML das hauptsächlich genutzte Format ist.

2.14 PROV Ontologie

Die PROV Ontologie beschreibt die kodierung des PROV Datenmodells in OWL 2, also die dabei genutzten Klassen, Objekte und Attribute.

2.14.1 PROV Datenmodell

Provenance sind Informationen über Personen, Objekte und Vorgänge, die bei der Produktion von Daten beteiligt waren. Dadurch können Aussagen über die Qualität der Daten abgeleitet werden. Das PROV Datenmodell ist ein generisches Modell, was die Darstellung solcher Daten ermöglicht. So beinhaltet es Informationen darüber, zu welcher Zeit welche Objekte oder Vorgänge genutzt wurden, welche Daten von welchen abgeleitet wurden, wer verantwortlich für die Daten ist, sowie Verfahren um Provenance über Provenance zu ermöglichen und mehr.

2.15 Semantic Web

Das Semantic Web dient dem einfacheren Datenaustausch zwischen verschiedenen Anwendungen und der besseren Datenverwertung, in dem Inhalte mit weiteren Informationen verknüpft werden. Dadurch entsteht ein Graph mit Knoten und Kanten. Dies wird vor allem durch URIs und RDF realisiert (siehe RDF/URI). Außerdem ermöglicht die Semantic Web Technologie das automatische generieren von Webseiten, sodass nicht mehr jede Seite exakt mit HTML geschrieben werden muss, sondern sich erzeugt sobald ein User auf den Link klickt.

2.16 Linked Open Data

Linked Open Data (LOD) bezeichnet im World Wide Web frei verfügbare Daten, die per URI identifiziert sind und darüber direkt per HTTP abgerufen werden können und ebenfalls per URI auf andere Daten verweisen. Idealerweise werden zur Kodierung und Verlinkung der Daten das Resource Description Framework (RDF) und darauf aufbauende Standards wie SPARQL und die Web Ontology Language (OWL) verwendet, so dass Linked Open Data gleichzeitig Teil des Semantic Web ist.

3 Aspekte

3.1 Ziel

Ziel unseres Projektes ist es, aus unstrukturierten elektronischen Dokumenten automatisiert strukturierte RDF-Daten zu generieren. Dabei soll es möglich sein, sowohl Daten aus dem Web als auch bereits annotierte Daten aufzubereiten. Diese generierten RDF-Daten sollen anschließend automatisiert unter Verwendung unterschiedlicher Tools ausgewertet werden können.

3.2 Herausforderungen

Herausfordernd für unser Team wird es sein, clevere und effiziente Möglichkeiten (Pattern und Filter) zu finden, um die gewünschten Informationen aus den Dokumenten „herauszuhören“. Ein anderes Ziel ist Erstellung parallel verarbeitender Methoden, um in relativ kurzer Zeit die

Datenmengen des gegebenen Korpus zu analysieren. Des Weiteren müssen wir darauf achten, dass wir sowohl deutsche als auch englische Dokumente verarbeiten können und unsere erzeugten RDF-Daten ebenfalls in deutsch und englisch auswertbar sind. Außerdem wollen wir die Modularität aller verwendeten Tools mithilfe von Interfaces sicherstellen, um später eventuell Tools austauschen zu können.

3.3 ähnliche Projekte

AlchemyAPI ist ein Unternehmen, welches kostenpflichtige Tools anbietet, die natürliche Sprachen mittels maschinellem Lernen und semantischer Textanalysen verarbeiten. Wir wollen uns durch die Modularität unserer eingebunden Tools von AlchemyAPI und andere Projekten abheben.

4 Quellen

- *Lemnitzer, L. & Zinsmeister, H.*, 2010: Korpuslinguistik. Eine Einführung. Tübingen: Narr
- *Gerber, D. & Ngomo, A.-C.N.*, 2011: Bootstrapping the Linked Data Web.
  In: 1st Workshop on Web Scale Knowledge Extraction ISWC (2011).
  Unter: http://aksw.org/Projects/BOA (abgerufen am 06.01.2016)
- http://de.wikipedia.org/wiki/Framework (abgerufen am 06.01.2016)
- http://lucene.apache.org (abgerufen am 06.01.2016)
- http://lucene.apache.org/solr (abgerufen am 06.01.2016)
- http://lucidworks.com/blog/2012/05/21/solr-4-preview/ (abgerufen am 06.01.2016)
- http://de.wikipedia.org/wiki/Unified_Modeling_Language (abgerufen am 06.01.2016)
- http://de.wikipedia.org/wiki/Entity-Relationship-Modell (abgerufen am 06.01.2016)
- https://de.wikipedia.org/wiki/Semantisches_Datenmodell (abgerufen am 06.01.2016)
- https://de.wikipedia.org/wiki/Linked_Open_Data (abgerufen am 06.01.2016)
- http://www.w3.org/TR/prov-o/ (abgerufen am 06.01.2016)
- http://ceur-ws.org/Vol-1272/paper_70.pdf (abgerufen am 06.01.2016)
- http://persistence.uni-leipzig.org/nlp2rdf/ (abgerufen am 06.01.2016)
- http://www.w3.org/TR/owl2-overview/ (abgerufen am 06.01.2016)
- http://www.w3.org/TR/rdf11-primer/ (abgerufen am 06.01.2016)
- https://de.wikipedia.org/wiki/Uniform_Resource_Identifier (abgerufen am 06.01.2016)
- https://en.wikipedia.org/wiki/SPARQL (abgerufen am 06.01.2016)
- https://en.wikipedia.org/wiki/Ontology_(information_science) (abgerufen am 06.01.2016)
- http://www.w3.org/TR/2012/REC-owl2-overview-20121211/ (abgerufen am 06.01.2016)
- http://www.w3.org/TR/2013/REC-prov-dm-20130430/ (angerufen am 06.01.2016)