Semantische Suche in Zeitungsartikeln

Ina Baumgarten, Niklas Meinzer

6. Oktober 2011

Motivation

- Im Internet finden sich viele Nachrichtenseiten
- Für die Suche auf diesen Seiten wird meißt Volltextsuche verwendet
- $\rightarrow$ Suche nach Vorkommen der Suchbegriffe in den Texten.

Angenommen, man möchte etwas abstrakter suchen:

"Artikel über Sportler, die verletzt sind und US-Bürger sind."

Suchbegriffe: *Sportler, US-Bürger, verletzt*

Ziel

Broccoli

Eine semantische Suchmaschine entwickelt an der Uni Freiburg.
Derzeit existiert eine Instanz, die die englische Wikipedia durchsucht.

Ziel der Arbeit

Eine semantische Suchmaschine für deutsche Zeitungsartikel aufsetzend auf *Broccoli*

Teilaufgaben

Es wurden insbesondere folgende Teilaufgaben bearbeitet:

- Vorverarbeitung
- (mehrstufige) Entitätenerkennung
- Evaluation

Datenbeschaffung

Um die Artikeltexte für die Suche vorverarbeiten zu können, werden lokale Kopien der Daten benötigt.

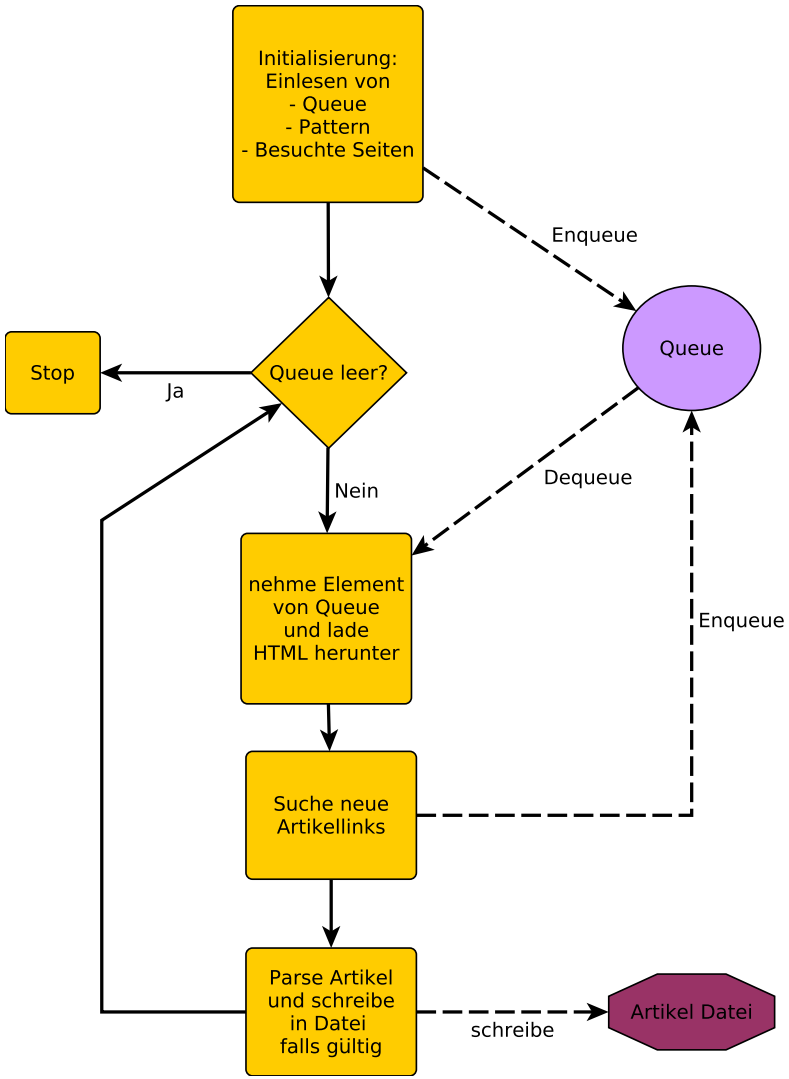
Da diese als HTML Dateien im Web verfügbar sind, können sie mit Hilfe eines Crawlers gesammelt werden.

Funktionsweise des Crawlers

- Nimm Link von Queue
- Lade Seite herunter
- Suche nach neuen Links
- Verarbeite Seite

Ausgabe

Eine CSV-Datei in der sämtliche gesammelte Artikel mit Metainformationen enthalten sind.



Deutsch-Englisch Mapping

Problem

Broccoli arbeitet mit einer englischsprachigen Ontologie, es sollen aber deutsche Artikel durchsucht werden.

$\rightarrow$ Nutze Wikipedia, um deutsche Entitäten auf englische zu mappen.

Disambiguierungsseiten

Wikipedia besitzt Disambiguierungsseiten, die für einen Begriff je die möglichen Bedeutungen auflisten

- Meist in Reihenfolge ihrer Bedeutungshäufigkeit

$\Rightarrow$ Nutze diese, um für mehrdeutige Begriffe die richtige Bedeutung finden zu können

Zählen von Entitäten

Entitäten kommen meist vollständig vor oder sie wurden zuvor vollständig genannt, z.B. Angela Merkel

- Zähle diese Entitäten in einer Menge von Artikeln, um die Entitäten nach ihrer Wahrscheinlichkeit ordnen zu können

$\Rightarrow$ Merkel als Angela Merkel wahrscheinlicher als Merkel als Stadt von Texas

unwichtige Wörter

Viele unwichtige Wörter, insbesondere Füllwörter, sind (Teil von) Entitäten, wie z.B. „und“

- Finde diese durch Durchsuchen nach Wörtern, die sowohl groß als auch klein geschrieben werden
- Anschließend manuelle Anpassung

PosTagger

Nicht alle unwichtigen Wörter werden als unwichtig erkannt

- nutze Wortarten (Verb, Adjektiv, Nomen, etc.)
- wird ermöglicht durch PosTagger
- neues Format der Artikel

Die#Artikel Kommune#Nomen von#Präposition
Michael#Eigenname ist#Verb arm#Adjektiv.

PosTagger

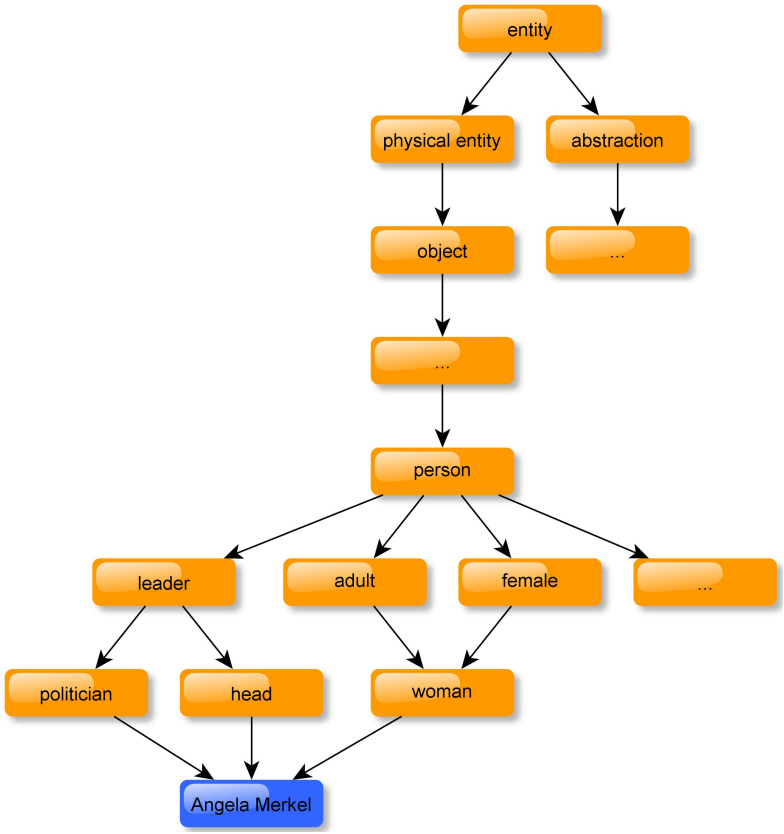
Nicht alle unwichtigen Wörter werden als unwichtig erkannt

- nutze Wortarten (Verb, Adjektiv, Nomen, etc.)
- wird ermöglicht durch PosTagger
- neues Format der Artikel

Die#Artikel Kommune#Nomen von#Präposition
Michael#Eigenname ist#Verb arm#Adjektiv.

YAGO-Facts

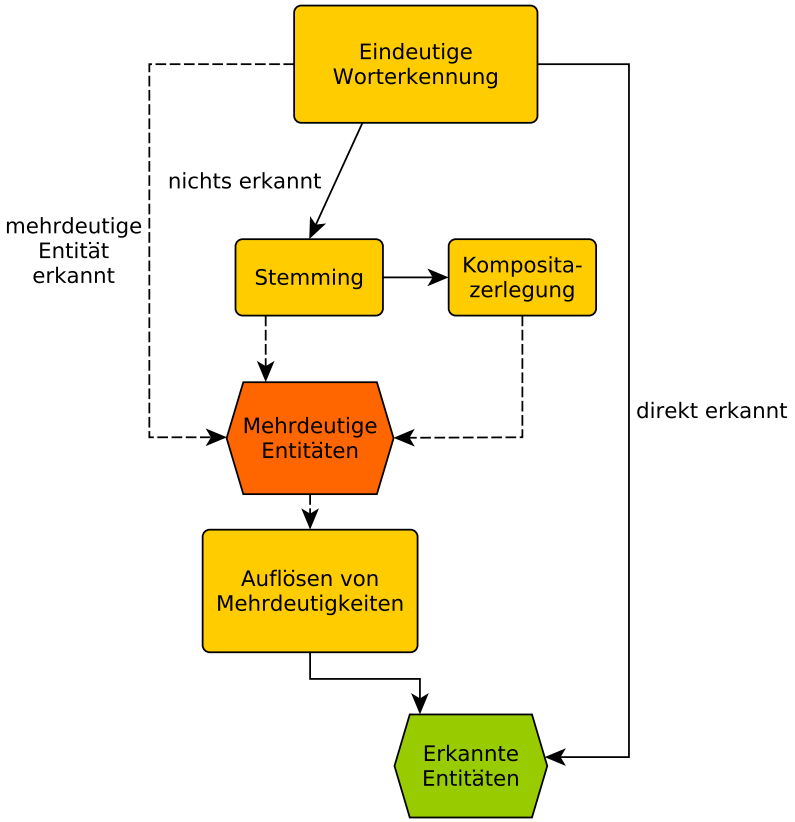
Die Pfade der YAGO Ontologie werden für die Disambiguierung verwendet.



Entitätserkennung

Teilmodule

- Eindeutige Entitätserkennung
- Uneindeutige Entitätserkennung
- Kompositazerlegung
- Stemming



Datenstruktur

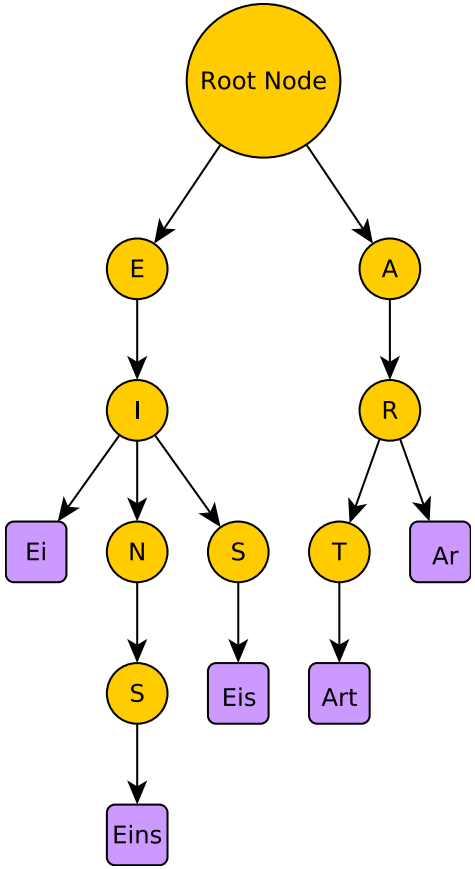
Für die Repräsentation der Entitätenliste bedarf es einer Datenstruktur die folgende Operationen effizient unterstützt:

- Einfügen
- Suchen
- Finde Suffixe

In dieser Arbeit kommt eine *Trie* Datenstruktur zum Einsatz.
(*Trie* von *retrieval Tree*)

Trie

- Speichert Daten mit String-Schlüssel
- Knoten sind die Buchstaben der Schlüssel



Eindeutige Entitätserkennung

Während der Entitätserkennung wird der Text Wort für Wort nach Entitäten durchsucht.

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Eindeutige Entitätserkennung

Während der Entitätserkennung wird der Text Wort für Wort nach Entitäten durchsucht.

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland (Stemming!)
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

Beispiel

Beispielsatz

Staatschef Deutschlands ist Angela Merkel aus Hamburg.

Entitätenliste

- Staatschef
- Deutschland
- Angela Merkel
- Angela Müller
- Hamburg
- Hamburg Harburg

→ *Hamburg* wird an Disambiguierung übergeben.

Stemming

Um konjugierte und deklinierte Wörter erkennen zu können wird ein Stemmingverfahren angewandt.

Caumanns, Jörg: A Fast and Simple Stemming Algorithm for German Words / Free University of Berlin, CeDiS

Entferne Schrittweise bestimmte Buchstabenkombinationen vom Ende der Wörter

- e
- s
- n
- t
- em
- er
- nd

Kompositazerlegung

Auch in Komposita können wichtige Informationen stecken

- Zerlege die Worte in alle möglichen Varianten

Beispiele

- Unfallopfer

Kompositazerlegung

Auch in Komposita können wichtige Informationen stecken

- Zerlege die Worte in alle möglichen Varianten

Beispiele

- Unfall opfer
- Landtagswahl

Kompositazerlegung

Auch in Komposita können wichtige Informationen stecken

- Zerlege die Worte in alle möglichen Varianten

Beispiele

- Unfall opfer
- Landtagswahl

Kompositazerlegung

Auch in Komposita können wichtige Informationen stecken

- Zerlege die Worte in alle möglichen Varianten

Beispiele

- Unfall opfer
- Landtag wahl

$\Rightarrow$ Fugenlaute können enthalten sein

Kompositazerlegung

Auch in Komposita können wichtige Informationen stecken

- Zerlege die Worte in alle möglichen Varianten

Beispiele

- Unfall opfer
- Landtag wahl

$\Rightarrow$ Fugenlaute können enthalten sein

Beispiel - falsch

- Datenstrukturen

Kompositazerlegung

Auch in Komposita können wichtige Informationen stecken

- Zerlege die Worte in alle möglichen Varianten

Beispiele

- Unfall opfer
- Landtag wahl

$\Rightarrow$ Fugenlaute können enthalten sein

Beispiel - falsch

- Date nstrukturen

$\Rightarrow$ Beide Wörter müssen Entitäten sein

Uneindeutige Entitätserkennung

Mehrdeutige Entitäten werden während der eindeutigen Entitätenerkennung in einer Liste gespeichert.

Beispiele

Merkel	Angela Merkel; Merkel (Stadt); ...
Angela von	Angela von Merzhausen; ...
und der	Und der Haifisch, der hat Zähne; ...

Minimierungsregeln

- Entitäten enthalten keine unwichtigen Wörter
- Entitäten sind Substantive
- Beispiel: „Angela von“ → „Angela“

Uneindeutige Entitätserkennung

Mehrdeutige Entitäten werden während der eindeutigen Entitätenerkennung in einer Liste gespeichert.

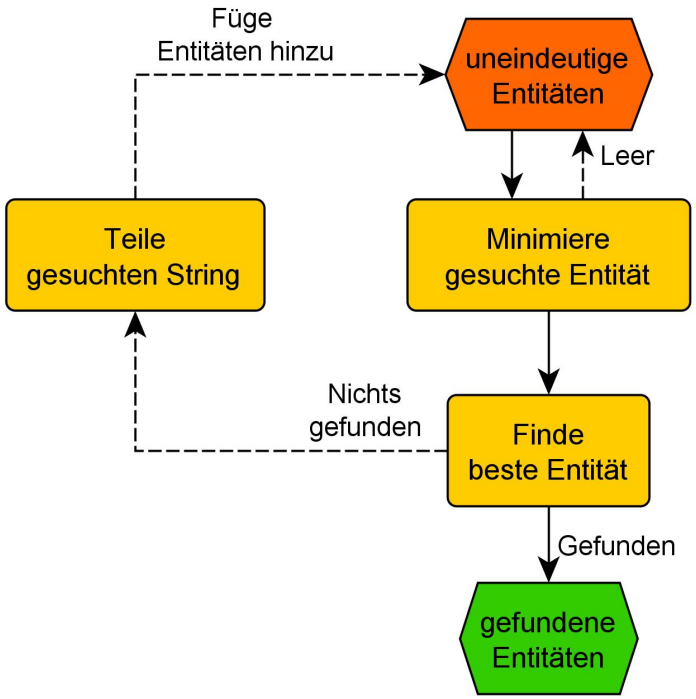
Beispiele

Merkel	Angela Merkel; Merkel (Stadt); ...
Angela von	Angela von Merzhausen; ...
und der	Und der Haifisch, der hat Zähne; ...

Minimierungsregeln

- Entitäten enthalten keine unwichtigen Wörter
- Entitäten sind Substantive
- Beispiel: „Angela von“ → „Angela“

Teilentitätenerkennung



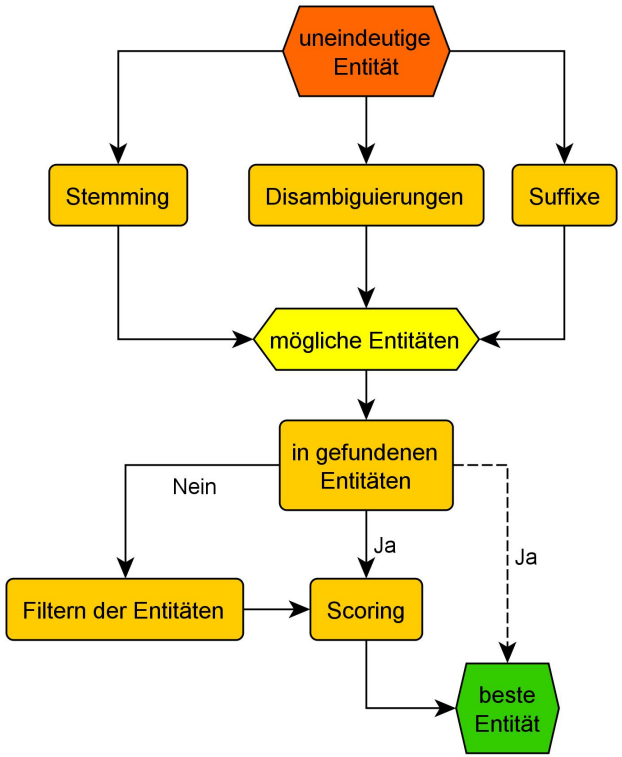
Teilen der Phrase

- evtl. keine passende Entität mehr durch Entfernung der Wörter
- falls mindestens zwei Worte: Aufteilung in neue, ambige Entitäten

Minimierungsregeln

- Entitäten enthalten keine unwichtigen Wörter
- Entitäten sind Substantive
- Beispiel: „Angela von“ → „Angela“

Algorithmus



Filtern der Entitäten

- „Michael“ in „Michael Klammhausen“ sollte nicht als „Michael Ballack“ erkannt werden

Regel - gefundene Entitäten

- Entitäten wiederholen sich
- Beispiel: „Angela Merkel“ gefunden → „Merkel“ als „Angela Merkel“

Scoring

3 gewichtete Faktoren:

- Länge der Entität
- Kontext durch YAGO-Facts
- Anzahl der Vorkommen

$\Rightarrow$ Normalisierung

Boni:

- Disambiguierungen
- Entität und gesuchte Entität haben (fast) den gleichen Wortlaut

Regel - simpelste Entität

- simple Entitäten sind wahrscheinlicher
- Beispiel: „Partei“ als „Partei“, nicht als „Partei Bibeltreuer Christen“

Grundlagen der Evaluation

Ground Truth zum Vergleichen mit den tatsächlichen Daten:

- <Wort> <PosTag> <gewünschte Entität>

⇒ Netz Nomen Stromnetz

⇒ gehen Verb

Klassifizierung der Entitäten:

- unmögliche Entitäten
- falsch negative Entitäten - nicht gefundene Entitäten
- falsch positive Entitäten - gefundene, aber falsche Entitäten
- richtig positive Entitäten - gefundene und korrekte Entitäten

Evaluation einer Ground Truth

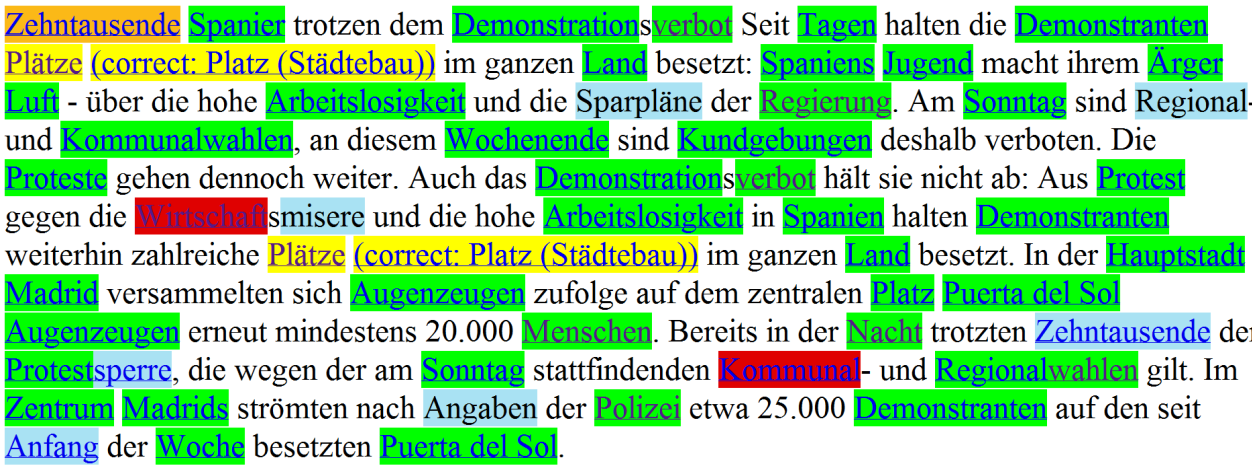

Zehntausende Spanier trotzen dem Demonstrationsverbot Seit Tagen halten die Demonstranten Plätze (correct: Platz (Städtebau)) im ganzen Land besetzt: Spaniens Jugend macht ihrem Ärger Luft - über die hohe Arbeitslosigkeit und die Sparpläne der Regierung. Am Sonntag sind Regional- und Kommunalwahlen, an diesem Wochenende sind Kundgebungen deshalb verboten. Die Proteste gehen dennoch weiter. Auch das Demonstrationsverbot hält sie nicht ab: Aus Protest gegen die Wirtschaftsmisere und die hohe Arbeitslosigkeit in Spanien halten Demonstranten weiterhin zahlreiche Plätze (correct: Platz (Städtebau)) im ganzen Land besetzt. In der Hauptstadt Madrid versammelten sich Augenzeugen zufolge auf dem zentralen Platz Puerta del Sol Augenzeugen erneut mindestens 20.000 Menschen. Bereits in der Nacht trotzten Zehntausende der Protestsperre, die wegen der am Sonntag stattfindenden Kommunal- und Regionalwahlen gilt. Im Zentrum Madrids strömten nach Angaben der Polizei etwa 25.000 Demonstranten auf den seit Anfang der Woche besetzten Puerta del Sol.

Abbildung: Graphische Evaluation eines Artikels

Berechnung von

- Trefferquote
- Genauigkeit
- unmögliche Entitäten

grün	-	korrekt
gelb	-	falsch
orange	-	fälschlicherweise gefunden
rot	-	nicht gefunden
blau	-	unmöglich

Ergebnisse

Ground Truth	Trefferquote in %	Genauigkeit in %	unmögliche Entitäten in %
1. Artikel	$90,1$	$97,1$	$14,2$
2. Artikel	$86,3$	$97,6$	$16,6$
3. Artikel	$75,8$	$97,0$	$27,3$

Abbildung: Resultate der Evaluation auf drei verschiedenen Ground Truths

- nur wenig repräsentativ, da zu wenige Ground Truths
- dennoch:
  Ergebnisse können sehr in Abhängigkeit des Artikels schwanken!

Diskussion - unmögliche Entitäten & Genauigkeit

Unmögliche Entitäten - ca. 19%:

- fehlende Entitäten (deutschsprachige und englischsprachige Wikipedia)

$\Rightarrow$ fehlerhaft gefundene Entitäten können Suchergebnis verfälschen

- „Anfang“ als der Film „Der Anfang“

Genauigkeit - ca. 97%:

- Kompositazerlegung:
  - „Finanzsenatorin“ $\rightarrow$ „Finanz“ existiert nicht
- Stemming:
  - „Händchen“ $\rightarrow$ „Hand“
  - „Internationalem Währungsfond“ $\rightarrow$ „Internationaler Währungsfond“

Diskussion - Trefferquote

Trefferquote - ca. 82%:

- unzureichende Kontexterkennung
  - unvollständig:
    YAGO-Fact für „Kernkraftwerk Lingen“ fehlt
  - ungenau:
    „Kernkraftwerk Krümmel“ und „Berliner U-Bahnstation Schönleinstraße“ haben denselben YAGO-Fact
- Kettenreaktionen
  - bedingt durch unzureichende Kontexterkennung
  - Annahme, dass gefundene Entitäten korrekt sind

  $\Rightarrow$ Folgefehler bei falsch erkannten Entitäten

Ausblick

Verbesserte Kontexterkennung

- Aktualisieren der YAGO-Facts
- Wahl einer genaueren Ontologie
  - muss auch mit den Entitäten der Wikipedia arbeiten

Minimierung von Kettenreaktionen

- Test auf gefundene Entitäten erst während des Scoring, dann aber Bonus

Minimierung unmöglicher Entitäten

- Einführung einer minimalen Punktzahl im Scoring

Literatur


Buchhold, Björn: SUSI: Wikipedia Search Using Semantic Index Annotations. Masterarbeit, Albert-Ludwigs-Universität Freiburg, Lehrstuhl für Algorithmen und Datenstrukturen, 2010


Meinzer, Niklas: Semantische Suche in Zeitungsartikeln. Bachelorarbeit, Albert-Ludwigs-Universität Freiburg, Lehrstuhl für Algorithmen und Datenstrukturen, 2011


Schmid, Helmut: TreeTagger - a language independent part-of-speech tag- ger. 1994, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ - Zugriff am 21.09.2011


Schmid, Helmut: Improvements In Part-of-Speech Tagging With an Application To German. Dublin, 1995, http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf - Zugriff am 21.09.2011