Event Recognition Engine

Eine Analysis Engine im UIMA Framework

Hauptseminar Information Retrieval

Tobias Beck

10.01.2011

Übersicht:

- Einordnung UIMA
- Komponenten einer UIMA Pipeline
- Selbst erstellte Event Recognition Engine
- Bewertung
- Quellen und Materialien

Hintergrund

- Mehrheit der Daten liegen in unstrukturierter Form vor
- Ziel: Gewinnung von strukturierten Daten

Was ist UIMA? (1/2)

- **U**nstructured **I**nformation **M**anagement **A**rchitecture

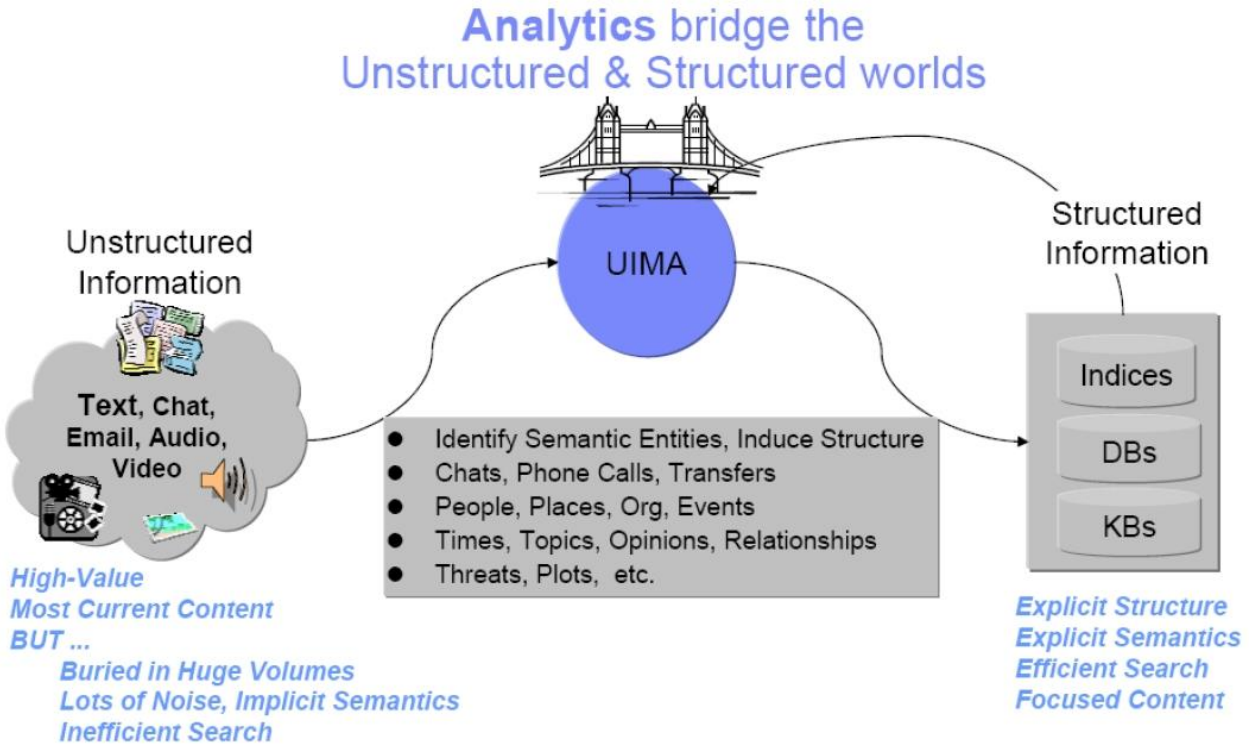
- OpenSource Framework und SDK
  - entwickelt bei IBM
  - jetzt Apache-Projekt

- Standardisiert von OASIS

Was ist UIMA? (2/2)

- unterstützt Komponenten in verschiedenen Programmiersprachen
- unterstützt Kombination von Komponenten zu einer Pipeline
- modularisierbar, skalierbar

Was macht UIMA?



Quelle:
http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/html/overview_and_setup/overview_and_setup.html
Abbildung 2.1, Letzter Zugriff: 02.02.2011

Komponenten einer UIMA Pipeline

- UIMA Framework Core

Komponenten einer UIMA Pipeline

- UIMA Framework Core
- Analysis Engine (AE)

Analysis Engine

- eigenständige Komponente für eine Teilaufgabe
- Bestandteile:
  - Component Descriptor
  - Annotator
- aggregierbar zu komplexen Analysis Engines

Komponenten einer UIMA-Pipeline

- UIMA Framework Core
- Analysis Engine (AE)
- Common Analysis Structure (CAS)

Common Analysis Structure

- Datenstruktur, die Austausch zwischen den Komponenten ermöglicht

- umfasst:
  - Dokument
  - Beschreibung des Typsystems
  - Annotationen
  - Index (referiert auf Annotationen)

Komponenten einer UIMA Pipeline

- UIMA Framework Core
- Analysis Engine (AE)
- Common Analysis Structure (CAS)
- Collection Processing Engine (CPE)

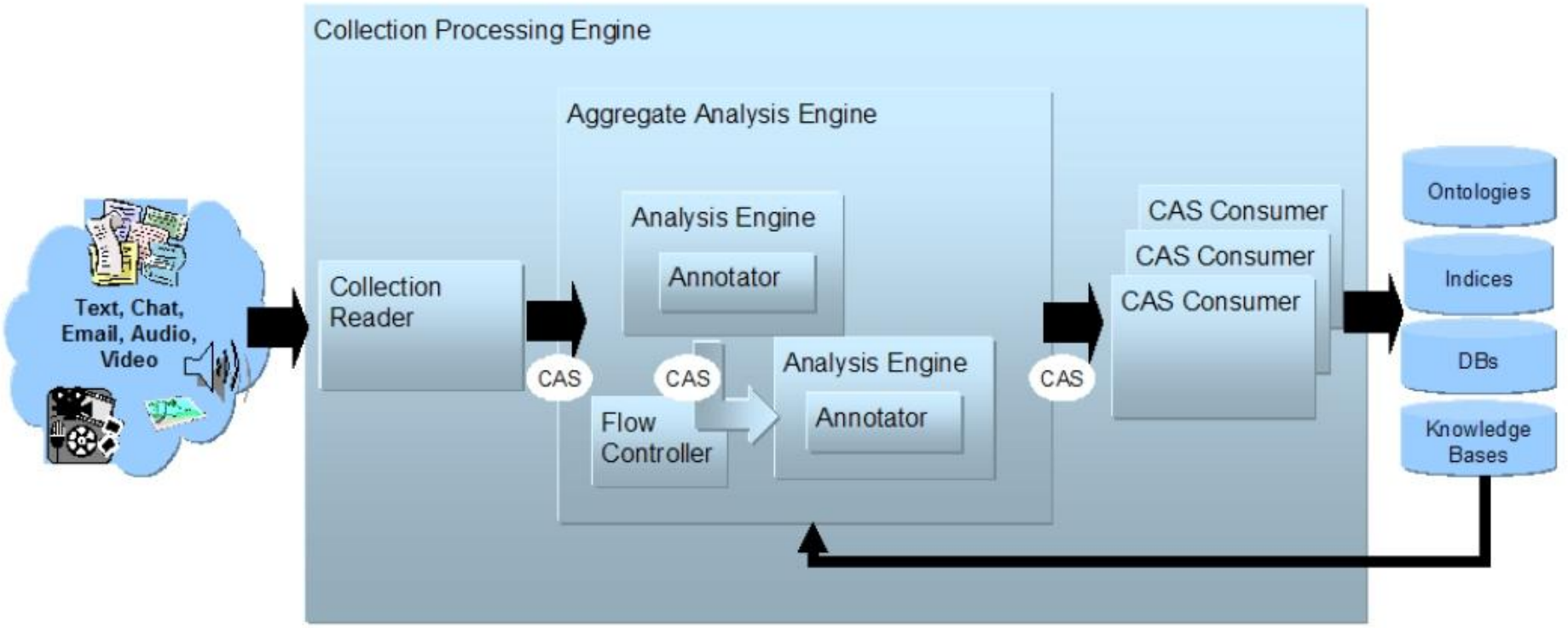
Collection Processing Engine

- liest Daten über Collection Reader
- präpariert die Daten zur Weiterverarbeitung
- steuert die Analyse
- gibt Daten über CAS Consumer aus

Komponenten einer UIMA Pipeline

- UIMA Framework Core
- Analysis Engine (AE)
- Common Analysis Structure (CAS)
- Collection Processing Engine (CPE)
- Collection Processing Manager (CPM)

Architektur der Pipeline



Quelle:
http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/html/overview_and_setup/overview_and_setup.html
Abbildung 2.5, Letzter Zugriff: 02.02.2011

Event Recognition Engine

- selbst erstellte Komponente
- dient als Analysis Engine in der vorgestellten Pipeline
  - zerlegt Eingabedokument in einzelne Tokens
  - sucht nach Events in den Satzsegmenten
  - annotiert Dokument mit zeitlicher Information

Im Folgenden die wichtigsten Schritte
zur Erstellung meiner Anwendung

Event Dictionary

- Textdatei, generiert aus Wikipedia
- liefert Daten für Annotationen der Event Recognition Engine

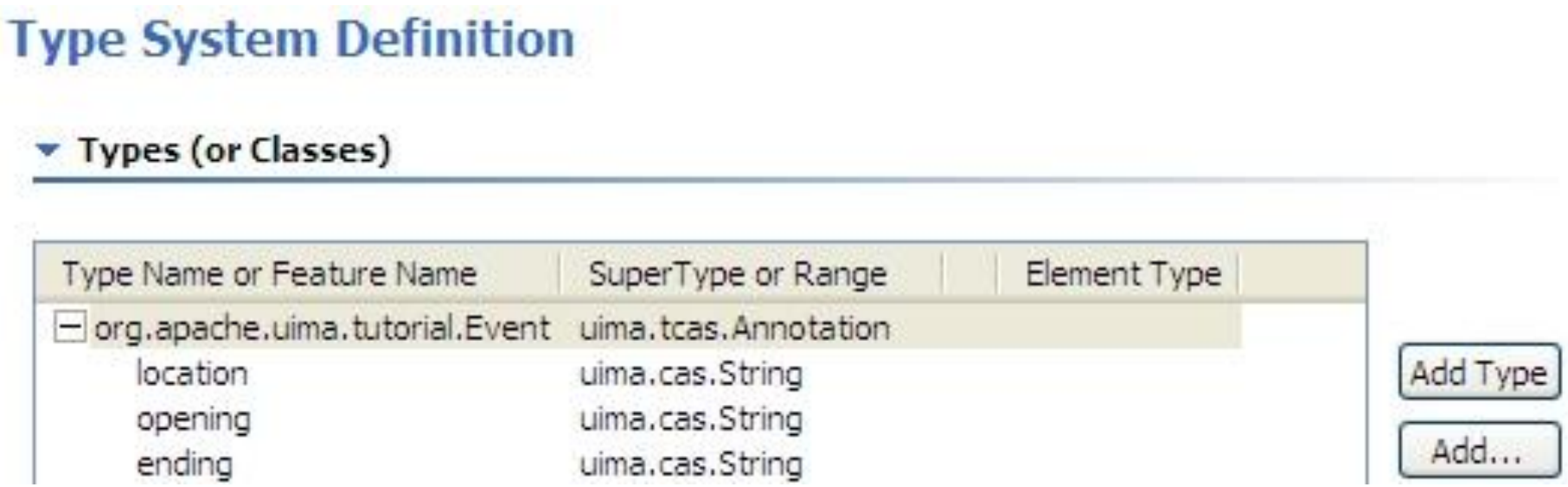
Event	Location	Begin	End
…			
FIFA-Fußball-WM 1990	Italien	8. Juni 1990	8. Juli 1990
FIFA-Fußball-WM 1994	USA	17. Juni 1994	17. Juli 1994
FIFA-Fußball-WM 1998	Frankreich	10. Juni 1998	12. Juli 1998
FIFA-Fußball-WM 2002	Japan und Südkorea	31. Mai 2002	30. Juni 2002
FIFA-Fußball-WM 2006	Deutschland	9. Juni 2006	9. Juli 2006
FIFA-Fußball-WM 2010	Südafrika	11. Juni 2010	11. Juli 2010
…			

Eingabedokument

Die DFB-Auswahl zählt zu den erfolgreichsten Fußballnationalmannschaften der Welt.
Dreimal ("FIFA-Fußball-WM 1954", "FIFA-Fußball-WM 1974" und "FIFA-Fußball-WM 1990") konnte die deutsche Mannschaft den Weltmeistertitel gewinnen,
viermal ("FIFA-Fußball-WM 1966", "FIFA-Fußball-WM 1982", "FIFA-Fußball-WM 1986", FIFA-Fußball-WM 2002) ging sie als Vize-Weltmeister vom Platz.
Dazu kamen vier dritte Plätze bei der "FIFA-Fußball-WM 1934", "FIFA-Fußball-WM 1970", "FIFA-Fußball-WM 2006" und "FIFA-Fußball-WM 2010"
sowie ein vierter Platz bei der "FIFA-Fußball-WM 1958".

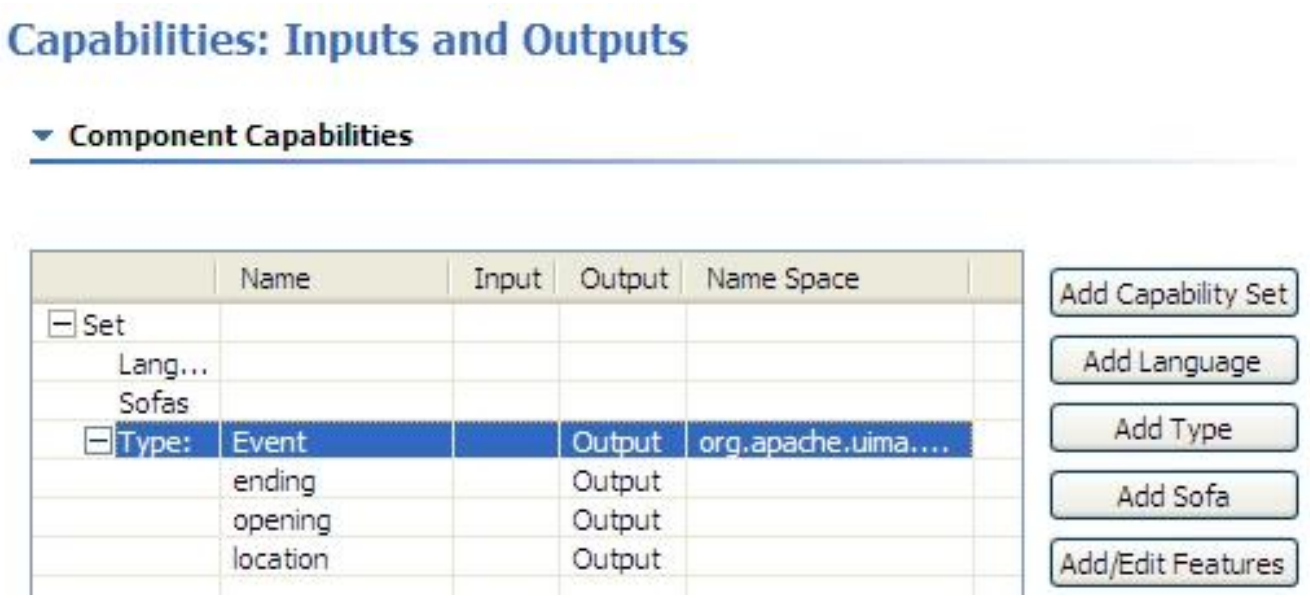
Type System Descriptor

- definiert Typsystem für die Event Recognition Engine



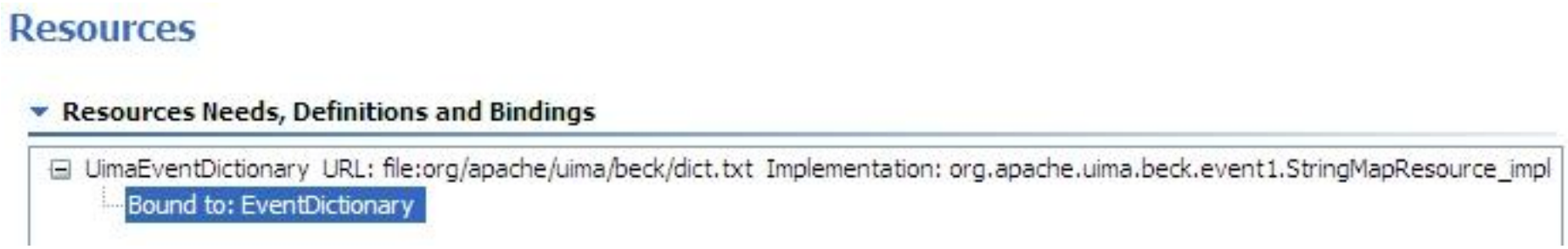
Component Descriptor

- definiert Ein- und Ausgabeparameter der Event Recognition Engine



Component Descriptor

- bindet externe Ressourcen wie das Event Dictionary ein



Annotator (1/2)

- einlesen des Event Dictionary

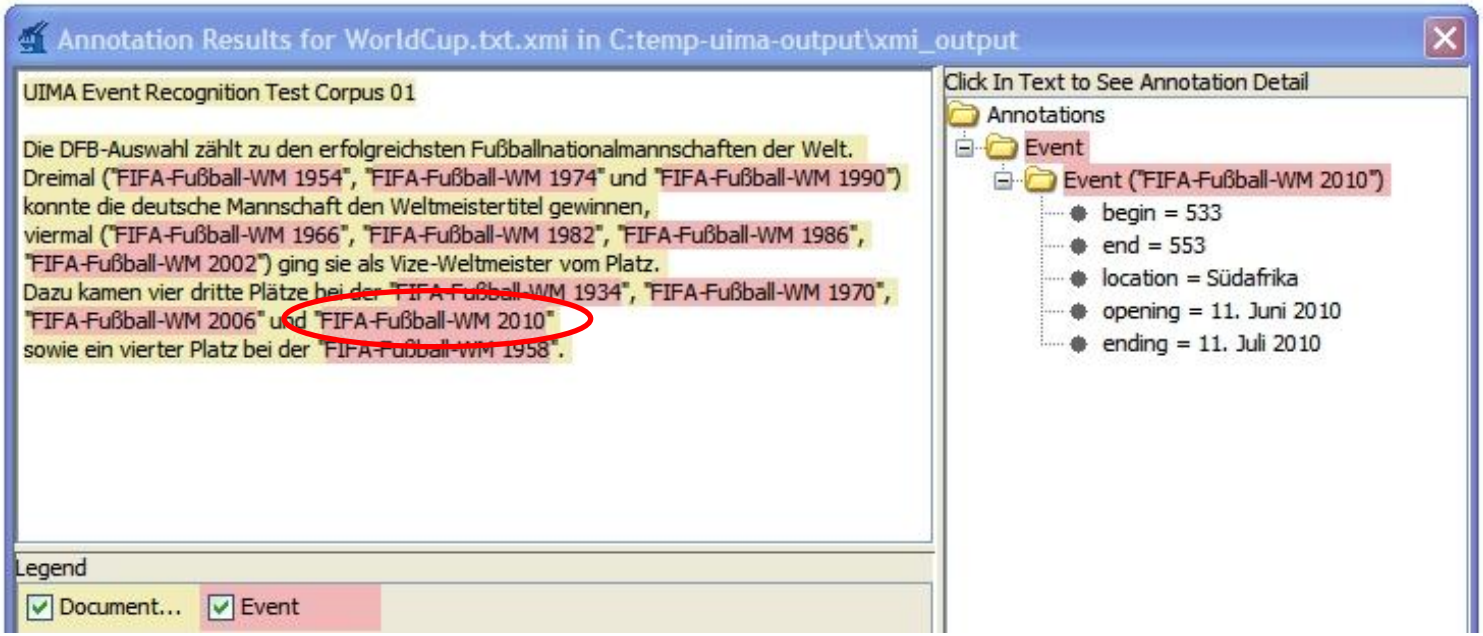
```
20 public class EventRecognitionAnnotator extends JCasAnnotator_ImplBase {
21
22     private StringMapResource mMap;
23
24     public void initialize(UimaContext aContext)
25             throws ResourceInitializationException {
26         super.initialize(aContext);
27         // get the ressourceObject
28         try {
29             mMap = (StringMapResource) getContext().getResourceObject(
30                     "EventDictionary");
31         } catch (ResourceAccessException e) {
32             throw new ResourceInitializationException(e);
33         }
34     }
```

Annotator (2/2)

```
70    public void process(JCas aJCas) {
71        String text = aJCas.getDocumentText();
72        int pos = 0;
73        StringTokenizer tokenizer = new StringTokenizer(text,
74                "\t\n\r.<.>/?\";:[{]}\\|=+()!", true);
75
76        while (tokenizer.hasMoreTokens()) {
77            String token = tokenizer.nextToken();
78
79            // look up token in map to see if it is an event
80            Value eventLine = mMap.getValue(token);
81            if (eventLine != null) {
82
83                // create annotation
84                Event event = new Event(aJCas, pos, pos + token.length());
85                event.setLocation(eventLine.getLocation());
86                event.setOpening(eventLine.getOpening());
87                event.setEnding(eventLine.getEnding());
88                event.addToIndexes();
89            }
90            // go to next token
91            pos += token.length();
92        }
93    }
```

UIMA Annotation Viewer

- visualisiert Annotationen (hier Weltmeisterschaft 2010) der Event Recognition Engine



Bewertung meiner Event Recognition Engine

- nur explizit angegebene Tokens aus dem Wörterbuch werden erkannt
  - FIFA Fußball-WM 2010

    vs.
  - Fußball WM im Jahre 2010 ausgerichtet von der FIFA

Quellen und Material

- Apache-Projekt :

  http://uima.apache.org/

- Tutorial:

  http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/html/index.html

- Verfügbare Analysis Engines:
  - http://www.julielab.de/Resources/Software/NLP+Tools/Download.html
  - http://incubator.apache.org/opennlp/
  - http://uima.lti.cs.cmu.edu
  - …