A Collaborative, Indeterministic and Partly Automatized Approach to Text Annotation

Thomas Bögel\textsuperscript{1}, Evelyn Gius\textsuperscript{2}, Marco Petris\textsuperscript{2}, Jannik Strötgen\textsuperscript{1}

The heureCLÉA Project
\textsuperscript{1}Heidelberg University, \textsuperscript{2}University of Hamburg
http://www.heureclea.de/

July 8, 2014
A Collaborative, Indeterministic and Partly Automatized Approach to Text Annotation

Thomas Bögel\textsuperscript{1}, Evelyn Gius\textsuperscript{2}, Marco Petris\textsuperscript{2}, Jannik Strötgen\textsuperscript{1}

The heureCLÉA Project
\textsuperscript{1}Heidelberg University, \textsuperscript{2}University of Hamburg
http://www.heureclea.de/

July 8, 2014
Motivation

Different types of annotations:

- some annotation tasks are complex, e.g.,
  - flashbacks, prolepsis, analepsis, ...

Manual creation of simple annotations is slow and time-consuming. Spend time on complex tasks, add simple annotations automatically.
Motivation

Different types of annotations:

- some annotation tasks are complex, e.g.,
  - flashbacks, prolepsis, analepsis, ...
- some annotation tasks are simple, e.g.,
  - sentences, part-of-speech information, tenses, ...
Motivation

Different types of annotations:

- some annotation tasks are **complex**, e.g., flashbacks, prolepsis, analepsis, ...
- some annotation tasks are **simple**, e.g., sentences, part-of-speech information, tenses, ...

**Manual creation of simple annotations**

- slow
- time-consuming
- boring
Motivation

Different types of annotations:
- some annotation tasks are **complex**, e.g.,
  - flashbacks, prolepsis, analepsis, ...
- some annotation tasks are **simple**, e.g.,
  - sentences, part-of-speech information, tenses, ...

Manual creation of simple annotations
- slow
- time-consuming
- boring

Spend time on complex tasks, add simple annotations automatically
Natural Language Processing

Automatic processing of text data
- goals (among others):
  - information extraction
  - automatic annotations
Natural Language Processing

Automatic processing of text data
  - goals (among others):
    - information extraction
    - automatic annotations

UIMA

UIMA: Unstructured Information Management Architecture
  - component framework for unstructured data (e.g., text)
  - helps to connect tools not developed to be used together
    → all components rely on the same data structure
      (Common Analysis Structure, CAS)
Natural Language Processing

Automatic processing of text data

- goals (among others):
  - information extraction
  - automatic annotations

UIMA

UIMA: Unstructured Information Management Architecture

- component framework for unstructured data (e.g., text)
- helps to connect tools not developed to be used together
  \[ \rightarrow \text{all components rely on the same data structure} \]
  \[ \text{(Common Analysis Structure, CAS)} \]

UIMA programs are processing pipelines
Components of a UIMA Pipeline

UIMA pipelines contain three components
Components of a UIMA Pipeline

- Docs
- Collection Reader
- Analysis Engines
- CAS Consumer
- Results
Components of a UIMA Pipeline
Components of a UIMA Pipeline

- **Collection Reader**
  - reads documents from source (e.g., file system, database)
Components of a UIMA Pipeline

Collection Reader
- reads documents from source (e.g., file system, database)
- instantiates a CAS for each document

Analysis Engines
CAS Consumer
Results
Components of a UIMA Pipeline

Collection Reader
- reads documents from source (e.g., file system, database)
- instantiates a CAS for each document
- initializes CAS with doc text (metadata, etc.)

Collection Reader
Docs
Collection Reader
Analysis Engines
CAS Consumer
Results

CAS
- doc text
- metadata
Motivation

NLP & UIMA

UIMA-Catma

Components of a UIMA Pipeline

Collection
Reader
CAS
Consumer
Docs
Results
Analysis
Engines
Analysis
Engines
Analysis
Engines
Analysis
Engines
CAS
- doc text
- metadata

Analysis Engines

- usually several Analysis Engines
- analyze the document
Components of a UIMA Pipeline

Analysis Engines
- usually several Analysis Engines
- analyze the document
- read content of the CAS
Components of a UIMA Pipeline

- Docs
- Collection Reader
- Analysis Engines
  - CAS
    - doc text
    - metadata
    - annotations
- CAS Consumer
- Results

Analysis Engines
- usually several Analysis Engines
- analyze the document
- read content of the CAS
- add annotations to the CAS
Components of a UIMA Pipeline

- **Docs**
- **Collection Reader**
- **Analysis Engines**
- **CAS Consumer**
- **Results**

**CAS Consumer**
- reads content of the CAS

CAS - doc text
- metadata
- annotations
Components of a UIMA Pipeline

CAS Consumer
- reads content of the CAS
- does final processing
  - evaluation, visualization, indexing
Components of a UIMA Pipeline

Docs -> Collection Reader -> Analysis Engines -> CAS Consumer -> Results

CAS

UIMA - What’s the clue?
Components of a UIMA Pipeline

UIMA - What’s the clue?
- single components are not directly connected to each other
- instead: CAS object
Components of a UIMA Pipeline

UIMA - What’s the clue?

- single components are not directly connected to each other
- instead: CAS object
- components are independent of each other
- components only have to be able to handle CAS
Example Tasks

heureCLÉA project:

- sentence splitting
- tokenization
- part of speech tagging
- temporal expressions
- temporal signals
- tense
Example Tasks

heureCLÉA project:
- sentence splitting
- tokenization
- part of speech tagging
- temporal expressions
- temporal signals
- tense

Further projects:
- sentence splitting
- tokenization
- part of speech tagging
- temporal expressions
- geographic expressions
- named entities (persons)
- event extraction
Example Tasks

heureCLÉA project:
- sentence splitting
- tokenization
- part of speech tagging
- temporal expressions
- temporal signals
- tense

further projects:
- sentence splitting
- tokenization
- part of speech tagging
- temporal expressions
- geographic expressions
- named entities (persons)
- event extraction
The Temporal Tagger HeidelTime

Extraction and normalization of temporal expressions

- July 8, 2014 → 2014-07-08
- today → 2014-07-08
The Temporal Tagger HeidelTime

Extraction and normalization of temporal expressions

- July 8, 2014 → 2014-07-08
- today → 2014-07-08

Most of the work so far: English news documents
The Temporal Tagger HeidelTime

Extraction and normalization of temporal expressions

- July 8, 2014 → 2014-07-08
- today → 2014-07-08

Most of the work so far: English news documents

HeidelTime: Multilingual, Cross-domain Temporal Tagger

Current languages:

- English, Spanish, German, French, Italian
- Dutch, Arabic, Vietnamese, Chinese, Russian

Publicly available:

- UIMA & standalone versions, online demo
- © Google Code
The UIMA-Catma Pipeline

- Collection Reader: reads documents from a source
- Analysis Engines: add annotations
- CAS Consumer: does final processing
The UIMA-Catma Pipeline

- Collection Reader: *reads document from CATMA*
- Analysis Engines: *add annotations*
- CAS Consumer: *sends annotations to CATMA*
Live Demo

So far: manual annotations of adverbs and superlatives

→ nice annotation task to get familiar with CATMA

→ boring if you want to annotate a whole book
Live Demo

So far: manual annotations of adverbs and superlatives
→ nice annotation task to get familiar with CATMA
→ boring if you want to annotate a whole book

Our CATMA-UIMA workflow:
- CATMA Collection Reader → get documents from CATMA
- TreeTagger wrapper (part of UIMA HeidelTime kit) HeidelTime
  → sentence splitting, tokenization, part-of-speech tagging
  → temporal expressions
- CATMA CAS Consumer: → send annotations to CATMA
Processing completed successfully.
Documents Processed: 1
Total Time: 75.251 seconds

100% (75251ms) – Collection Processing Engine
0.14% (104ms) – File System Collection Reader (Process)
5.35% (4028ms) – TreeTaggerWrapper (Analysis)
0% (0ms) – TreeTaggerWrapper (End of Batch)
38.2% (28744ms) – HeidelTime (Analysis)
0% (0ms) – HeidelTime (End of Batch)
56.31% (42375ms) – Catma CAS Consumer (Analysis)
0% (0ms) – Catma CAS Consumer (End of Batch)
being very fond of using them, his habit of stammering was not thereby improved. In fact, there were periods in his discourse when he would finally give up and swallow his discomfort—in a glass of water.

As I said, my uncle, Professor Hardwigg, was a very learned man; and I now add a most kind relative. I was bound to him by the double ties of affection and interest. I took deep interest in all his doings, and hoped some day to be almost as learned myself. It was a rare thing for me to be absent from his lectures. Like him, I preferred mineralogy to all the other sciences. My anxiety was to gain real knowledge of the earth. Geology and mineralogy were to us the sole objects of life, and in connection with these studies many a fair specimen of stone, chalk, or metal did we break with our hammers.

Steel rods, loadstones, glass pipes, and bottles of various acids were oftener before us than our meals. My uncle Hardwigg was once known to classify six hundred different geological specimens by their weight, hardness, fusibility, sound, taste, and smell.

He corresponded with all the great, learned, and scientific men of the age. I was, therefore, in constant communication with, at all events the letters of, Sir Humphry Davy, Captain Franklin, and other great men.

But before I state the subject on which my uncle wished to confer with me, I must say a word about his personal appearance. Alas! my readers will see a very different portrait of him at a future time, after he has gone through the fearful adventures yet to be related.