UIMA-based Annotation Type System for a Text Mining Architecture

Udo Hahn, Ekaterina Buyko, Katrin Tomanek, Scott Piao, Yoshimasa Tsuruoka, John McNaught, Sophia Ananiadou

Jena University Language and Information Engineering Lab & School of Computer Science, University of Manchester
BOOTStrep NLP Infrastructure

Bootstrapping Of Ontologies and Terminologies STStrategic REsearch Project

NLP Components Repository

Team 1
Team 2
Team n

Tool 1 -> Tool 2 -> Tool n

Annotated Facts
Annotation in Natural Language Processing (NLP)

NLP System
Tokenizer
POS Tagger
Entity Tagger
Relation Tagger

Fred is CEO of IBM

<document source ../>
<sentence begin ... end../>
<token begin .. end ../>
<token begin .. end ..>

<entity person begin .. end>
<entity organization begin .. end/>
<relation is_ceo_of begin .. end/>
Annotation in NLP Systems
Annotation in NLP Systems

NLP Tool Suite 1

Token POS1

Data Conversion

Token POS2

NLP Tool Suite 2

Token POSn

NLP Tool Suite 3

NLP Tool Suite n

..
Advantages of the UIMA Framework

Interoperability between NLP systems

- Portability of components
- Flexible exchange of components
Annotation in NLP Systems

NLP Tool Suite 1

Token POS1

NLP Tool Suite 2

Token POSn

NLP Tool Suite 3

Token POS2

NLP Tool Suite n

..
Annotation in NLP Systems

NLP Tool Suite 1

NLP Tool Suite 2

NLP Tool Suite 3

NLP Tool Suite n

Token POS1

Token POS2

Data Conversion

Token POSn

..
Annotation in NLP Systems

NLP Tool Suite 1

Token POS1

NLP Tool Suite 2

Token POSn

Data Conversion

NLP Tool Suite 3

Token POS2

NLP Tool Suite n

..
Advantages of the UIMA Framework

Interoperability between NLP systems

✔ Portability of components

✗ Flexible exchange of components
Exchange of components in UIMA

• Adaptation Efforts
  • Over-write Wrappers
  • Create Matching Files

• Define a Common Annotation Type System in advance
Annotation in NLP Systems
Advantages of the UIMA Framework

Interoperability between NLP systems

✔ Portability of components

✔ Flexible exchange of components
Design of an Annotation Type System

• Requirements from various NLP teams

• Annotation guidelines and schemata
Requirements for an Annotation Type System

• Broad coverage for the information extraction

• Compatible to “standard” NLP annotation schemata

• Definition of the core type system which is extensible

• Using UIMA specific features
  - Multiple annotation of the same type
  - Annotation control through the restriction of values
Annotation Guidelines & Schemata
Corpus Annotation

• Annotation languages (e.g. XML (in-line, stand-off))

• Annotation levels:
  - Document Meta (e.g. Dublin Core Metadata Initiative)
  - Linguistic Analysis (e.g. TEI, XCES (EAGLES), Penn Treebank)
  - Semantic Analysis (e.g. MUC, ACE, GENIA)

• NLP system annotation guidelines?
1. **Document Meta**: author, publication data, source

2. **Document Structure & Style**: title, sections, text bold

3. **Morpho-Syntax**: token, part-of speech, lemma

4. **Syntax**: chunks, constituents, dependency relations

5. **Semantics**: entities, relations, events

6. **Discourse**: anaphora
Basic Annotation Type

```
-uima.tcas.Annotation
 +begin: uima.cas.Integer
 +end: uima.cas.Integer

Annotation
 +componentID: uima.cas.String
 +confidence: uima.cas.Double

DiscontinuousAnnotation
 +value: uima.cas.FSArray = Annotation
```
Morpho-Syntax

POSTag

Token

Lemma

Abbreviation

GrammaticalFeats

StemmedForm

Acronym
Morpho-Syntax I

Token

+posTag: uima.cas.FSArray = POSTag
+feats: GrammaticalFeats
+stemmedForm: StemmedForm
+lemma: Lemma
+orthogr: uima.cas.FSArray = String
+...
Morpho-Syntax IV

GrammaticalFeats
  +language: Language

  NounFeats
    +gender: uima.cas.String
    +case: uima.cas.String
    +number: uima.cas.String

  VerbFeats

  Token
    +posTag: uima.cas.FSArray = POSTag
    +feats: GrammaticalFeats
    +stemmedForm: StemmedForm
    +lemma: Lemma
    +orthogr: uima.cas.FSArray = String
    +...
Syntax

Chunk  Constituent  DependencyRelation
Shallow Parsing

- Chunk
  - PhraseChunk
    - NP
    - VP
    - PP
    - ADJP
    - ...

Full Parsing (constituent-based)
Full Parsing (dependency-based)
Resource Connection

ResourceEntry
+source: uima.cas.String
+entryID: uima.cas.String
+version: uima.cas.String

OntologyEntry

DBEntry

LexiconEntry
To wrap up ..

- Multi-layered annotation
- **Core** annotation type system
- Extended for the *biomedical* domain
- Can easily be extended for other domains
- **Restriction of values** for the annotation control
- Sub-Types for **multiple** annotation (e.g. POS, Chunk)
- **Connection** to external resources
Open Issues

- **Performance** measure of the type system

- **Definitions:**
  - Semantics (Relation, Event)
  - Discourse (Anaphora)
UIMA Annotation Type System Working Group?

Download: http://www.julielab.de/

Contact: buyko@coling-uni-jena.de

Sponsored by Boot Strep Information Society Technologies