Darmstadt Knowledge Processing Repository
Based on UIMA

Iryna Gurevych, Max Mühlhäuser, Christof Müller,
Jürgen Steimle, Markus Weimer, Torsten Zesch

Ubiquitous Knowledge Processing Group
Telecooperation, Computer Science Department
Darmstadt University of Technology
Telecooperation

Ubiquitous Computing

Ubiquitous Computing Infrastructures

Ubiquitous Learning Infrastructures

Multimedia Software Engineering

Ubiquitous Security & Trust

Ubiquitous Knowledge Processing

Ubiquitous Peer-to-Peer Infrastructures
Ubiquitous Knowledge Processing

SIR
AQUA
THESEUS

Darmstadt Knowledge Processing Software Repository
Automatic Quality Assessment and Feedback in eLearning 2.0 (AQUA)
User Generated Discourse in Web 2.0
AQUA – Anoto pen

- Bluetooth Transceiver
  Bluetooth is a communications standard that enables all things that are now connected by cable to be connected without

- Battery
  A rechargeable battery enables a full day of use

- Camera

- Digital Paper

- Memory

- Processor

- Ink Cartridge / Force Sensor

- The predefined area

- Absolute coordinate addresses
  - 1.5mm
  - 0.3mm

- Basic pattern
  - Above
  - Left
  - Right
  - Below

- Anoto pattern

- Diameter 0.15mm

- 5mm
AQUA - System Architecture

Natural Language Processing

Import into UIMA -> Annotation of the Text -> Export to a Featurevector

Evaluation <-> Training

Machine Learning

Data split
SIR (in cooperation with Prof. Hinrichs)

- **Semantic Information Retrieval**

  Bridge the human – computer gap

  Semantic search (SIR) based on semantic relatedness

  Natural language expression of information need

  Low level communication interface
Information Retrieval (IR)

Boolean, Vector Space, ...

Keywords

- Document 1
- Document 2
- Document 3
- Document ...
- Document ...
- Document ...
SIR-Project

Semantic Relatedness

Essay

cake, computer, to read, ...

Semantic search (SIR) based on semantic relatedness

Natural language expression of information need

low level communication interface

baker, to program, quality assurance

Profession ...

Profession 2

Profession 1

Profession ...

Profession ...

Profession 3
SIR Example

find good index terms

compute semantic relatedness

Compound Splitting
Negation Detection
WSD
THESEUS - TEXO

- Large-scale BMBF-Project, industry (SAP, Siemens, etc.)

- Service Marketplaces in Web 2.0
  - Find services, both users and machines

- Problem:
  - Only keyword-based search
  - Lack of ontologies for semantic search

- Solution:
  - Use natural language descriptions of web services
  - Apply Semantic Information Retrieval
  - Community Mining for optimized service selection

  ➔ Darmstadt Knowledge Processing Repository
UIMA components

SIR  AQUA  THESEUS

Data import
Wikipedia reader, *Forum reader*, Plain text reader

Linguistic preprocessing
Tokenizer, Sentence splitter, Stopword tagger

Morphological analysis
Stemmer, Lemmatizer, Compound Splitter

Syntactic analysis
PoS-Tagger, *Parser*

Semantic analysis
*NE tagger*, *Sentiment detector*, *WSD component*

Project specific analysis
Swear word tagger (AQUA), *Negation detection (SIR)*

Data export
Indexer (Lucene, Terrier), *ARFF export*
Advantages of UIMA

• Components can be shared between projects

• Shared model of thinking
  ▪ “Reader + Annotators + Consumer”
  ▪ Configuration of components

• Descriptive component orchestration
Challenges

• Agree on a type system
  ▪ No automatic type mapping

• Some rough edges in UIMA
  ▪ No real plug’n’work with PEAR packages
  ▪ Using constraints to align annotations seems to be slow
Wish list

• Automatic type matching
• Better tool support
  ▪ Improving Eclipse plug-ins (robustness, features)
  ▪ Refactoring of UIMA components
  ▪ CPE runner ++ (automatic logging, performance monitor, etc.)
• Plug’n’work approach
• “Import by name” in CPEs
  ▪ Or make ${CPM_HOME}/path also work for readers/consumers
• Construct XML descriptors from Java annotations
• More intuitive API
Thank you very much!

• Acknowledgements:

  ▪ DFG for funding “Semantic Information Retrieval”

  ▪ DFG for funding “Automatic Quality Assessment and Feedback in eLearning 2.0”

http://www.ukp.tu-darmstadt.de/