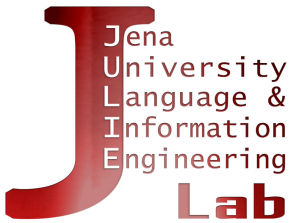
An UIMA-based Tool Suite for Semantic Text Processing

Katrin Tomanek, Ekaterina Buyko, Udo Hahn

Jena University Language & Information Engineering Lab



Bundesministerium
für Bildung
und Forschung

StemNet - Knowledge Management for Immunology

- in life-sciences: increasing amount of knowledge stored in (unstructured) textual documents
- semantic access to this knowledge necessary
- biomedical subdomain: hematopoetic *stem cell transplantation*

➔ semantic search engine for advanced document and information retrieval

- example user query:

  “get me relevant documents on **human IL2Ra** and **CTL**”

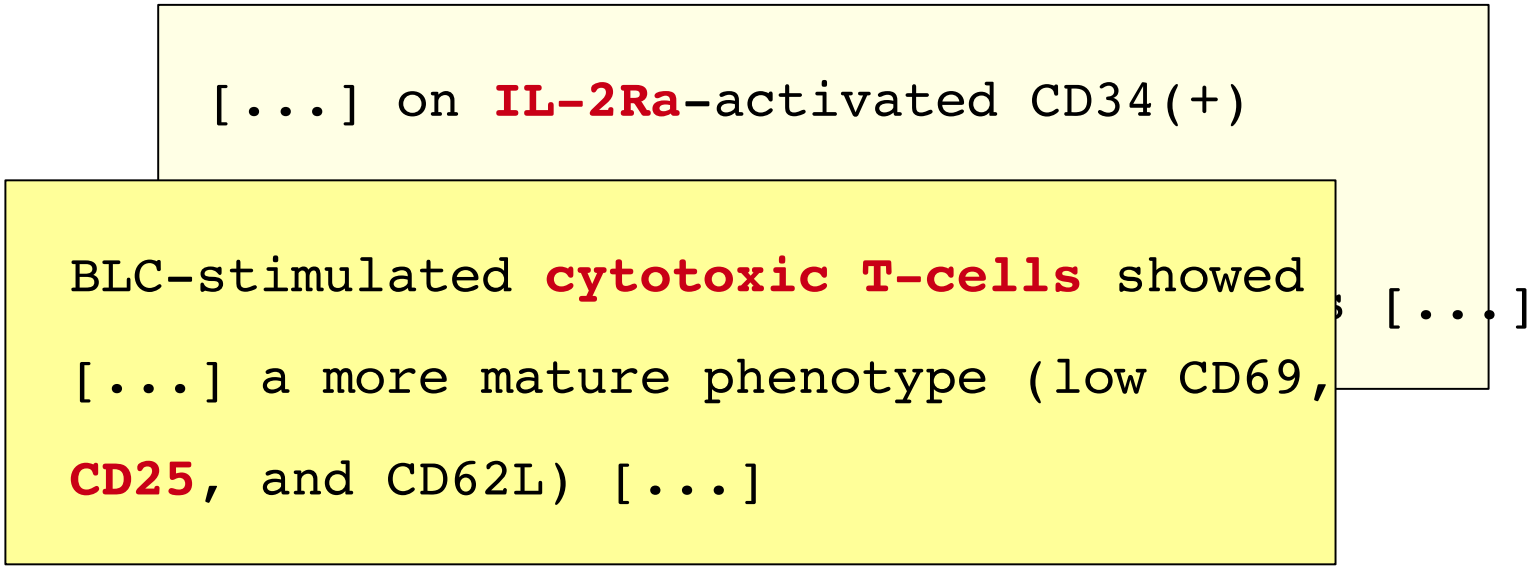
StemNet - Knowledge Management for Immunology

- user query: “human IL2Ra” AND “CTL”

> [...] on **IL-2Ra**-activated CD34(+)
> cytotoxic T-cells (**CTL**s). p3hr-1,
> the Burkit's lymphoma cell line, was [...]

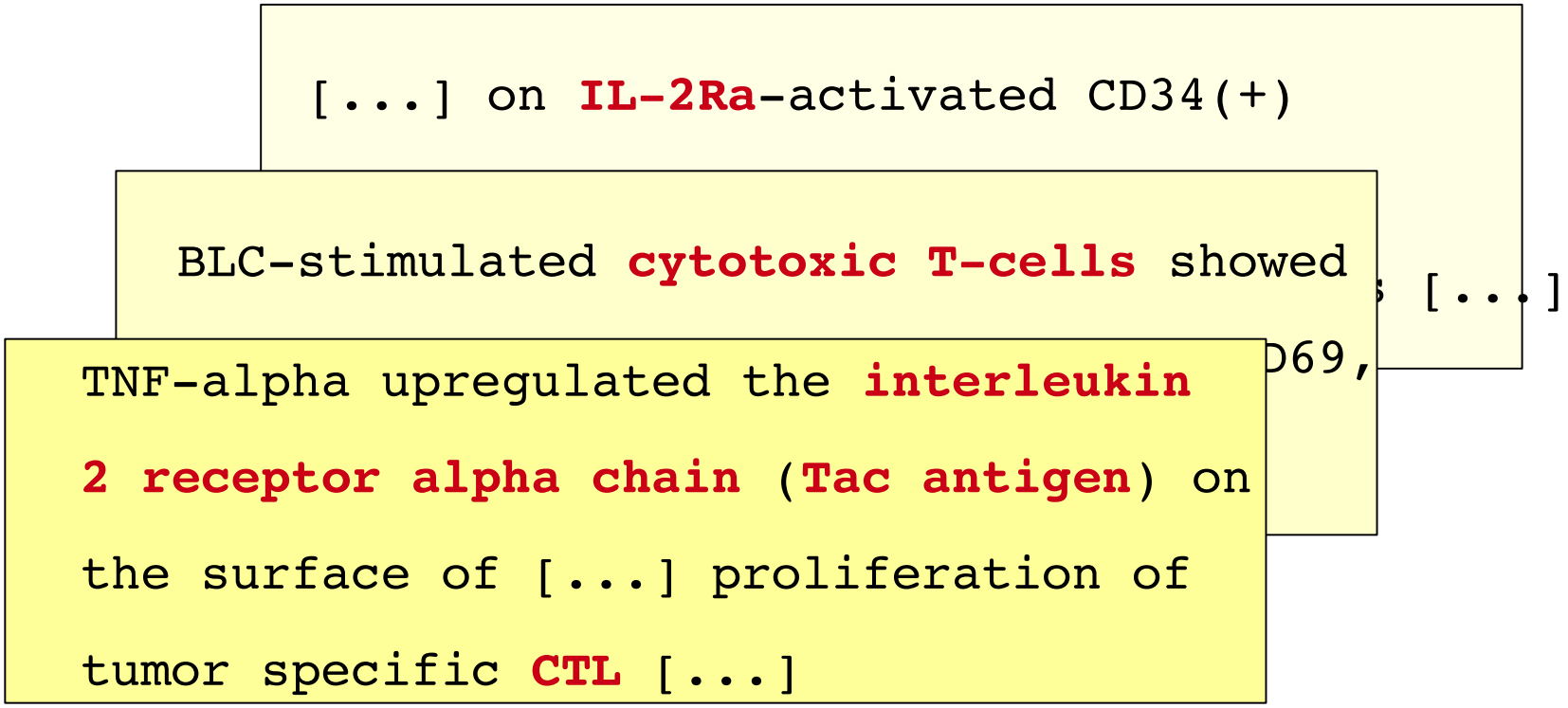
StemNet - Knowledge Management for Immunology

- user query: “human IL2Ra” AND “CTL”

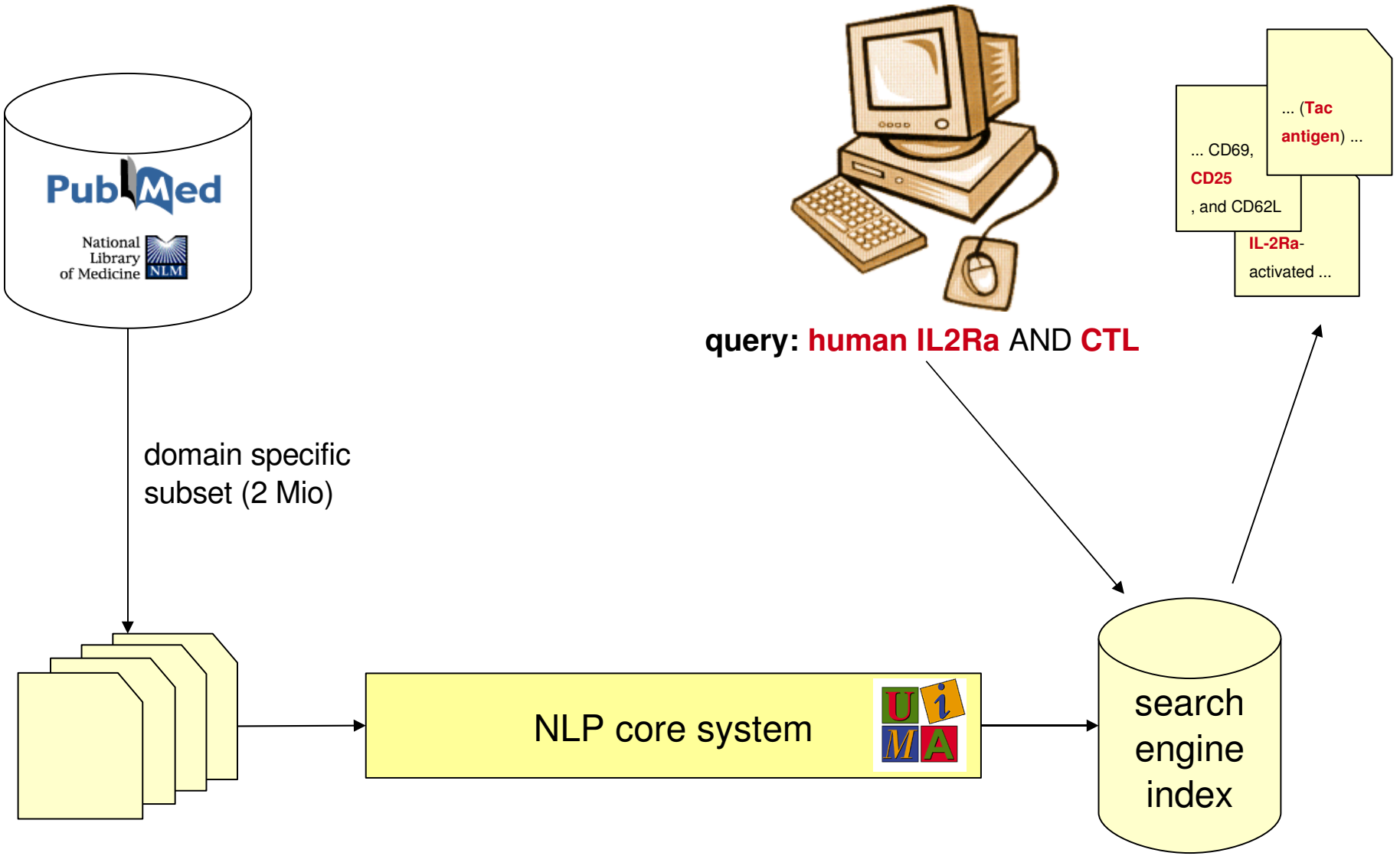


StemNet - Knowledge Management for Immunology

- user query: “human IL2Ra” AND “CTL”



UIMA in the StemNet Project



JULIE NLP Tool Suite based on UIMA (1/2)

1) comprehensive UIMA type system

- covers the full NLP pipeline
- five layers:
  - document meta information (bibliographic and content information)
  - document structure and style information (sentences, rhetorical zones, ...)
  - morpho-syntax (tokenisation, POS, acronyms, lemmatisation, ...)
  - syntax (shallow and full parsing information)
  - semantics (named entities, relationships, events...)

JULIE NLP Tool Suite based on UIMA (2/2)

2) collection of NLP components `(Analysis Engines)`:

- for morpho-syntactic analysis
- for syntactic analysis
- for named entity recognition and normalisation/mapping

3) data import and export `(Collection Reader/CAS Consumer)`:

- PubMed Reader
- Search Engine Indexer

- included tools:
  - mostly based on machine learning
  - external tools for which we have written UIMA wrappers
  - JULIE tools; have stand-alone and UIMA mode

PubMed Reader

- processes PubMed articles (XML)
- reads the following document meta-data:
  - bibliographic information: title, authors, publication date, journal name
  - content information (manually added): keywords (MeSH), list of chemicals
- writes data to CAS
  - ➔ our type system contains respective types for this kind of information

Sentence/Token Splitting, POS Tagging, Chunking

- configurable UIMA wrappers for OpenNLP tools
  - sentence splitter
  - tokeniser
  - POS tagger
  - chunker
- JULIE tools
  - sentence splitter
  - tokeniser
- available models for life-sciences:
  - trained on JULIE corpus (covers special cases and subtleties of bio-medical domain)
  - trained on well-known biomedical corpora (e.g. PennBioIE)

Parsing

- UIMA wrappers for external parser implementations:
  - OpenNLP Parser (Ratnaparkhi, 1998)
    - → consituency parser
  - MST Parser (McDonald, 2006)
    - → dependency parser

- different linguistic paradigms supported
- → type system supports both constituency and dependency parse information

Acronym Detection

- detection and resolution of local acronyms
- implementation of M. Hearst's algorithm (Hearst 2003)
- with extension: DB lookup for unresolved acronyms
- Acronym DB generator (CAS Consumer):
  - tuples (acronym, full form), associated with spelling variants, first year of occurrence, keywords (MeSH)

```
[...] on IL-2Ra-activated CD34(+)
cytotoxic T-cells (CTLs). p3hr-1,
the Burkit's lymphoma cell line, was [...]
```

Named Entity Recognition

- generic named entity recognizer
- ML-based
- flexibly configurable wrt:
  - mapping: predicted labels –> UIMA types
  - feature parametrization
    - user defined feature set (turn on/off, configure features)
    - CAS-specified feature information (e.g. POS tags)
- consistency preservation:
  - assures that same entity mentions within one abstract (document zone) are consistently annotated

Named Entity Mapping (1/2)

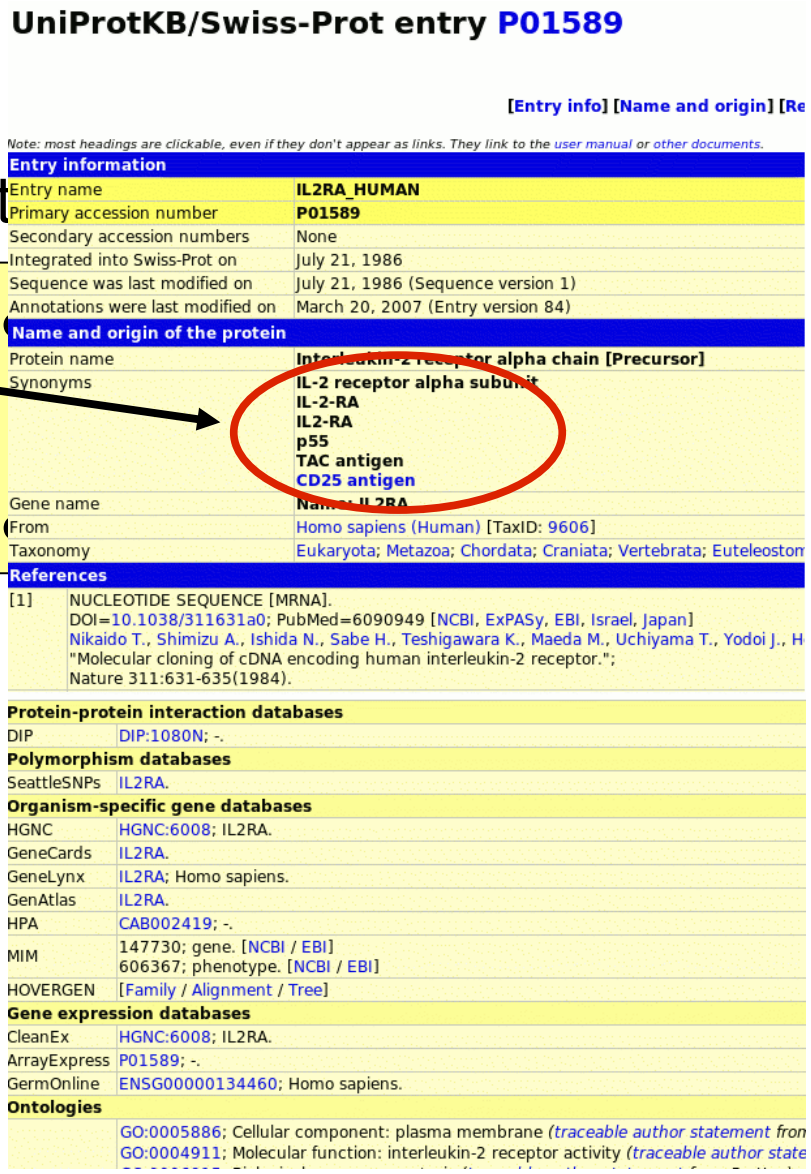
- associates identified NEs with DB entries
- in life-sciences: e.g. SwissProt

```
[...] on IL2Ra-activated CD34(+)
cytotoxic T-cells (CTLs). p3hr-1,
the Burkit's lymphoma cell line, was [...]
```

Named Entity Mapping (1/2)

- associates identified NEs with
- in life-sciences: e.g. SwissProt

```
[...] on IL2Ra-activate
cytotoxic T-cells (CTLs
the Burkit's lymphoma c
```



UniProtKB/Swiss-Prot entry P01589

[Entry info] [Name and origin] [Re

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**IL2RA_HUMAN**
Primary accession number	**P01589**
Secondary accession numbers	None
Integrated into Swiss-Prot on	July 21, 1986
Sequence was last modified on	July 21, 1986 (Sequence version 1)
Annotations were last modified on	March 20, 2007 (Entry version 84)
Name and origin of the protein	
Protein name	**Interleukin-2 receptor alpha chain [Precursor]**
Synonyms	**IL-2 receptor alpha subunit** **IL-2-RA** **IL2-RA** **p55** **TAC antigen** **CD25 antigen**
Gene name	**Name: IL2RA**
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostom

References	
[1]	NUCLEOTIDE SEQUENCE [MRNA]. DOI=10.1038/311631a0; PubMed=6090949 [NCBI, ExPASy, EBI, Israel, Japan] Nikaido T., Shimizu A., Ishida N., Sabe H., Teshigawara K., Maeda M., Uchiyama T., Yodoi J., H "Molecular cloning of cDNA encoding human interleukin-2 receptor."; Nature 311:631-635(1984).

Protein-protein interaction databases	
DIP	DIP:1080N; -.
Polymorphism databases	
SeattleSNPs	IL2RA.
Organism-specific gene databases	
HGNC	HGNC:6008; IL2RA.
GeneCards	IL2RA.
GeneLynx	IL2RA; Homo sapiens.
GenAtlas	IL2RA.
HPA	CAB002419; -.
MIM	147730; gene. [NCBI / EBI] 606367; phenotype. [NCBI / EBI]
HOVERGEN	[Family / Alignment / Tree]
Gene expression databases	
CleanEx	HGNC:6008; IL2RA.
ArrayExpress	P01589; -.
GermOnline	ENSG00000134460; Homo sapiens.
Ontologies	
	GO:0005886; Cellular component: plasma membrane (*traceable author statement from*
	GO:0004911; Molecular function: interleukin-2 receptor activity (*traceable author state*

Named Entity Mapping (2/2)

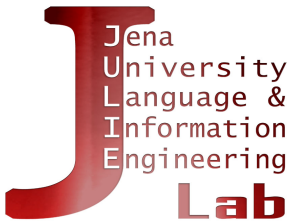
- for gene/protein entity mentions
- principles:
  - normalization rules for bio-medical entities
    - a -> alpha
    - R -> receptor, L -> ligand
    - numbers split away
    - word order ignored
    - `“IL2RA” -> “IL 2 receptor alpha”`
    - `“receptor of IL-4” -> “IL 4 receptor”`
  - requires well-curated synonym list

JULIE Lucene Indexer

- goal: directly build search engine index from processed documents
- Lucene
  - high-performance search engine
  - fielded search and special query types (e.g. range searches)
  - open source, freely available, provides Java API
- Lucene Indexer
  - directly consumes CAS
  - tokenization as in CAS
  - currently indexed fields:
    - document meta-data (as in PubMed)
    - entity mentions + synonyms (with same offset)
- work in progress: flexible configurability
  - external mapping file (UIMA type -> Lucene field)

for further information/download of tools:

http://www.julielab.de



Bundesministerium
für Bildung
und Forschung