Enterprise Platform and
Integration Concepts &
Information Systems

Master Project
Winter Term 2015/2016

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Learning to Note:
Intelligent Support for Document Annotation using Semi-Supervised Learning

Dr. Mariana Neves, Dr. Ralf Krestel

September 30th, 2015

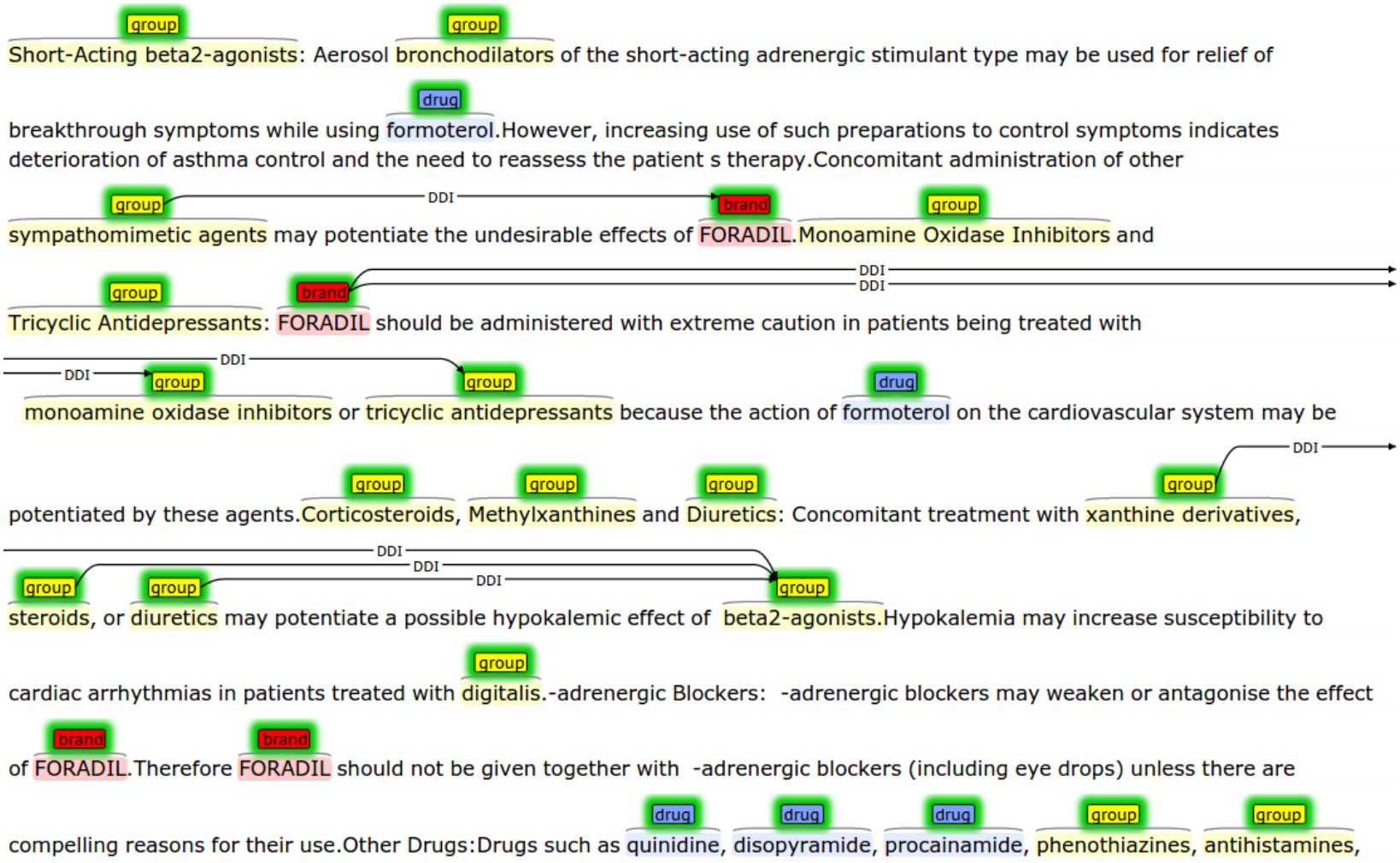
Overview

- Motivation
- NLP in HANA
- Tasks

Overview

- Motivation
- NLP in HANA
- Tasks

Motivation: annotation tool for corpus construction



(http://corpora.informatik.hu-berlin.de/index.xhtml#/ddi2013/DrugBank/DrugBank_d103)

Motivation: annotation tool as a curation tool

Arginase (L-arginine urea amidino hydrolase, EC 3.5.3.1) catalyses the hydrolysis of arginine to ornithine and urea and requires a bivalent metal ion, specially Mn^{2+}, for catalytic activity [1], [2], [3], [4], [5] and [6] and structural stabilization [4], [6] and [7]. Manganese ions are thought to activate a metal-bound water molecule, generating the hydroxide ion that nucleophilically attack the scissile guanidinium carbon of arginine [8] and [9]. An specially interesting aspect of the studies reported to date has been the detection of a Mn^{2+}–Mn^{2+} cluster in the active site of fully activated arginases from rat liver and *Bacillus caldovelox* [10] and [11]. One of the Mn^{2+}, designated Mn^{2+}_A in the case of rat liver arginase, is more weakly bound than the other, Mn^{2+}_B [12].

General information	
Organism	Homo sapiens
Tissue	liver
EC Class	3.5.3.1
SABIO reaction id	574
Variant	mutant H101N activated
Recombinant	expressed in Escherichia coli
Experiment Type	in vitro
Pathways	Arginine and Proline metabolism Insulin signaling pathway Urea cycle
Event Description	-

Substrates		
name	**location**	**comment**
H2O	-	-
L-Arginine	-	-

Products		
name	**location**	**comment**
L-Ornithine	-	-
Urea	-	-

SABIO-RK Biochemical Reaction Kinetics Database

Intelligent annotation tool

→ Suggesting entities and relationships

Acetaminophen diminished the binding of theophylline to human serum by a net change of 5.7% (percentage increase in free drug fraction [FDF], 11.0%) at 662 micromol/L and by a net change of 7.1% (percentage increase in FDF, 13.7%) at 1324 micromol/L.

Theophylline decreased the binding of acetaminophen by a net change of 6.8% (percentage increase in FDF, 8.8%) at 277.5 micromol/L; phenobarbital reduced it by a net change of 6.6% (percentage increase in FDF, 8.5%) at 431 micromol/L.

Valproic acid diminished binding of phenobarbital by a net change of 9.9% (percentage increase in FDF, 21.2%) at 1732 micromol/L.

No significant effects were noted with other drug combinations or with the addition of ethanol.

Coingestion of acetaminophen with theophylline, phenobarbital with acetaminophen, and valproic acid with phenobarbital at high to toxic concentrations decreases the binding of the target drug.

The resulting increase in free drug concentration may lead to enhanced drug effect in vivo.

Intelligent annotation tool

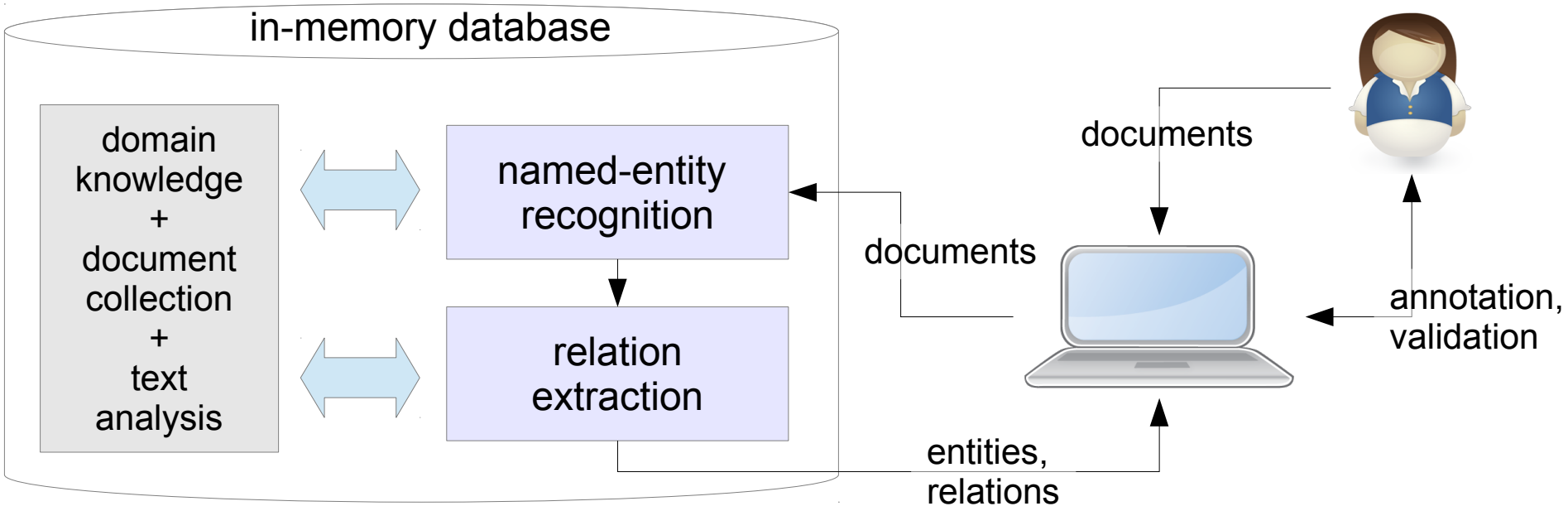
→ Suggesting documents

Effect of rofecoxib on the pharmacokinetics of digoxin in healthy volunteers. (PMID 11144988)
The authors examined the effect of the cyclooxygenase-2 (COX-2) inhibitor, rofecoxib, at steady state on the pharmacokinetics of digoxin following a single dose in healthy subjects. …

Glucose and insulin exert additive ocular and renal vasodilator effects on healthy humans. (PMID 11206417)
There is evidence that insulin and glucose cause renal and ocular vasodilation. There is, however, currently no data on the effect of combined hyperglycaemia and hyperinsulinaemia on the renal and ocular blood flow seen in diabetic patients on insulin therapy. …

Effect of diazepam and midazolam on the antinociceptive effect of morphine, metamizol and indomethacin in mice. (PMID 11210678)
The influence of midazolam and diazepam on antinociceptive effect of morphine (10 mg/kg), metamizol (500 mg/kg) and indomethacin (10 mg/kg) was investigated in a mouse model using the tail-flick and hot-plate tests. ...

Architecture of the system



Master project „Learning to Note“

- Goals
  - Develop an annotation tool which relies on text mining and machine learning
  - Integrate the named-entity recognition and relation extraction components
  - Evaluate the tool on benchmarks and for curation of real data
  - Submit a paper describing the system and/or the methods

Overview

- Motivation
- NLP in HANA
- Tasks

NLP in SAP HANA

- Built-in features
  - Support to many languages

Language modules that support linguistic analysis:

- Catalan
- Chinese (traditional)
- Croatian
- Czech
- Danish
- Norwegian Bokmål
- Norwegian Nynorsk
- Serbian
- Slovak
- Slovenian
- Swedish

Language modules that support linguistic analysis and predefined entity extraction:

- Arabic
- Chinese (simplified)
- Dutch
- English
- Farsi
- French
- German
- Italian
- Japanese
- Korean
- Portuguese
- Russian
- Spanish

Language modules that support basic linguistic analysis:

- Greek
- Hebrew
- Hungarian
- Polish
- Romanian
- Thai
- Turkish

(http://help.sap.com/hana/sap_hana_text_analysis_language_reference_guide_en.pdf)

NLP in SAP HANA

- Built-in features
  - Document indexing
  - Sentence splitting, tokenization, part-of-speech tagging

	ID	SEQ_SNIPPET	TA_RULE	TA_	TA_TOKEN	TA_L	TA_TYPE	TA_NORMALIZED	TA_STEM	TA_PARAGRAPH	TA_SENTENCE	TA_CRE	TA_OFFSET
1	5133b9455	0	LXP	1	the	en	determiner	the	?	1	1	Mar 11,	0
2	5133b9455	2	LXP	1	The	en	determiner	the	?	1	1	Mar 11,	0
3	5133b9455	3	LXP	1	the	en	determiner	the	?	1	1	Mar 11,	0
4	5133b9455	5	LXP	1	the	en	determiner	the	?	1	1	Mar 11,	0
5	5133b9455	4	LXP	1	Functional	en	adjective	functional	?	1	1	Mar 11,	0
6	5133b9455	1	LXP	1	Two	en	number	two	?	1	1	Mar 11,	0
7	5133b9455	6	LXP	1	Two	en	number	two	?	1	1	Mar 11,	1
8	5133b9455	3	LXP	2	early	en	adverb	early	?	1	1	Mar 11,	4
9	5133b9455	0	LXP	2	cannonical	en	unknown	cannonical	?	1	1	Mar 11,	4
10	5133b9455	1	LXP	2	AATAAA	en	unknown	aataaa	?	1	1	Mar 11,	4
11	5133b9455	5	LXP	2	appropriate	en	adjective	appropriate	?	1	1	Mar 11,	4
12	5133b9455	2	LXP	2	results	en	noun	results	result	1	1	Mar 11,	4
13	5133b9455	6	LXP	2	AATAAA	en	unknown	aataaa	?	1	1	Mar 11,	5
14	5133b9455	3	LXP	3	poly	en	unknown	poly	?	1	1	Mar 11,	10
15	5133b9455	4	LXP	2	polyadenylation	en	unknown	polyadenylation	?	1	1	Mar 11,	11
16	5133b9455	1	LXP	3	motifs	en	noun	motifs	motif	1	1	Mar 11,	11

NLP in SAP HANA

- Built-in features
  - Named entity recognition (dictionaries and/or rules)

	ID	TA_RULE	TA_COUNTER	TA_TOKEN	TA_LANGUAGE	TA_TYPE	TA_NORMALIZED
1	5319a6e9b166e2b806000023	Entity Extraction	1	genetic lesion	en	NOUN_GROUP	?
2	5319a6e9b166e2b806000023	Entity Extraction	2	Huntington	en	PERSON	?
3	5319a6e9b166e2b806000023	Entity Extraction	3	Huntington	en	LOCALITY	?
4	5319a6e9b166e2b806000023	Entity Extraction	82	disease	en	MESH	D004194
5	5319a6e9b166e2b806000023	Entity Extraction	83	disease	en	DO	DOID:4

	ID	TA_RULE	TA_COUNTER	TA_TOKEN	TA_LANGUAGE	TA_TYPE	TA_NORMALIZED
1	532f03e6d6d3ac6a34000022	Entity Extraction	2	GATA	en	SwissProt	GATA_GEOSE
2	532f03e6d6d3ac6a34000022	Entity Extraction	3	GATA	en	SwissProt	GATA_GEOSE
3	532f03e6d6d3ac6a34000022	Entity Extraction	4	GATA	en	SwissProt	GATA_GEOSE
4	532f03e6d6d3ac6a34000022	Entity Extraction	5	GATA	en	SwissProt	GATA_GEOSE
5	532f03e6d6d3ac6a34000022	Entity Extraction	209	heart	en	MESH	D006321
6	532f03e6d6d3ac6a34000022	Entity Extraction	210	myocardial	en	DO	DOID:5844
7	532f03e6d6d3ac6a34000022	Entity Extraction	211	myocardial	en	MESH	D009203
8	532f03e6d6d3ac6a34000022	Entity Extraction	208	regeneratic	en	MESH	D012038
9	532f03e6d6d3ac6a34000022	Entity Extraction	207	regeneratic	en	GO	GO:0031099
10	532f03e6d6d3ac6a34000022	Entity Extraction	1	role	en	MESH	D012380

NLP in SAP HANA

- Built-in features
  - Fuzzy search

```
SELECT DISTINCT SCORE() AS SCORE, "TA_TOKEN"  FROM "SEARCH"."$TA_BIOASQ_PUBMED_INDEX"
WHERE  (CONTAINS ("TA_TOKEN",'infer',FUZZY(0.9)) OR CONTAINS ("TA_TOKEN",'functional',FUZZY(0.9)) OR
CONTAINS ("TA_TOKEN",'associations',FUZZY(0.9)) OR CONTAINS ("TA_TOKEN",'event',FUZZY(0.9)))
ORDER BY SCORE DESC
```

	SCORE	TA_TOKEN
1	0.5	associations
2	0.5	Associations
3	0.5	FUNCTIONAL
4	0.5	Functional
5	0.5	functional
6	0.5	Event
7	0.5	event
8	0.5	infer
9	0.48500001430511475	association
10	0.48500001430511475	Association
11	0.47999998927116394	functionals
12	0.47999998927116394	functionaly
13	0.47999998927116394	functionnal
14	0.4650000035762787	infero
15	0.4650000035762787	eventS
16	0.4650000035762787	afunctional

NLP in SAP HANA

- Built-in features
  - Sentiments

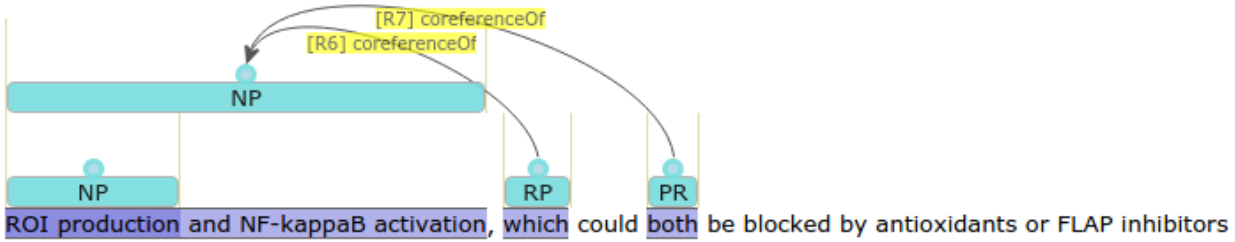
	ID	SEQ_	TA_RULE	TA_CC	TA_TOKEN	TA_L	TA_TYPE
1	52f5	24	Entity Extraction	1	fmr1 knockout	en	StrongPositiveSentiment
2	52f5	24	Entity Extraction	2	fmr1 knockout mouse model	en	Sentiment
3	52f7	0	Entity Extraction	1	Tuberous sclerosis complex (TSC) is a genetic disease in the group knov	en	Sentiment
4	52f5	33	Entity Extraction	1	In the MIA model of ASD, adult females are exposed to a simulated vira	en	Sentiment
5	52d	2	Entity Extraction	1	In patients with type 2 DM, the presence of SH serves as an additional ri	en	Sentiment
6	52d	2	Entity Extraction	2	patients	en	WeakPositiveSentiment
7	52d	6	Entity Extraction	1	In CHF patients TSH levels even slightly above normal range are indeper	en	Sentiment
8	52d	6	Entity Extraction	2	greater	en	StrongPositiveSentiment
9	52e	0	Entity Extraction	1	Our work describes, for the first time, distinct GATA-1 interactions with	en	Sentiment
10	52d	0	Entity Extraction	1	The decision to treat elderly people is still an unresolved clinical challer	en	Sentiment
11	52d	0	Entity Extraction	3	appropriately	en	WeakPositiveSentiment
12	52d	4	Entity Extraction	2	Subclinical hyperthyroidism seems to be a risk factor of developing ma	en	Sentiment
13	52d	5	Entity Extraction	1	SCH appears to influence the postoperative outcome for patients by in	en	Sentiment
14	52d	5	Entity Extraction	3	patients	en	WeakPositiveSentiment
15	530	2	Entity Extraction	1	Like	en	WeakPositiveSentiment
16	530	2	Entity Extraction	2	Like IκBs, Sef sequesters NF-κB in the cytoplasm of resting cells.	en	Sentiment

Overview

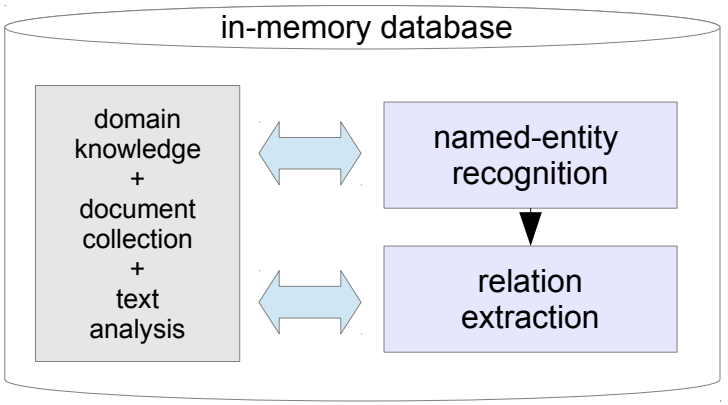
- Motivation
- NLP in HANA
- Tasks

Teams

- Team 1 (two students):
  - development of the annotation tool



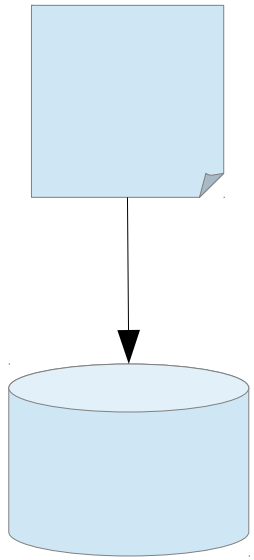
- Team 2 (two students):
  - development of text mining



(http://www.pubannotation.org/)

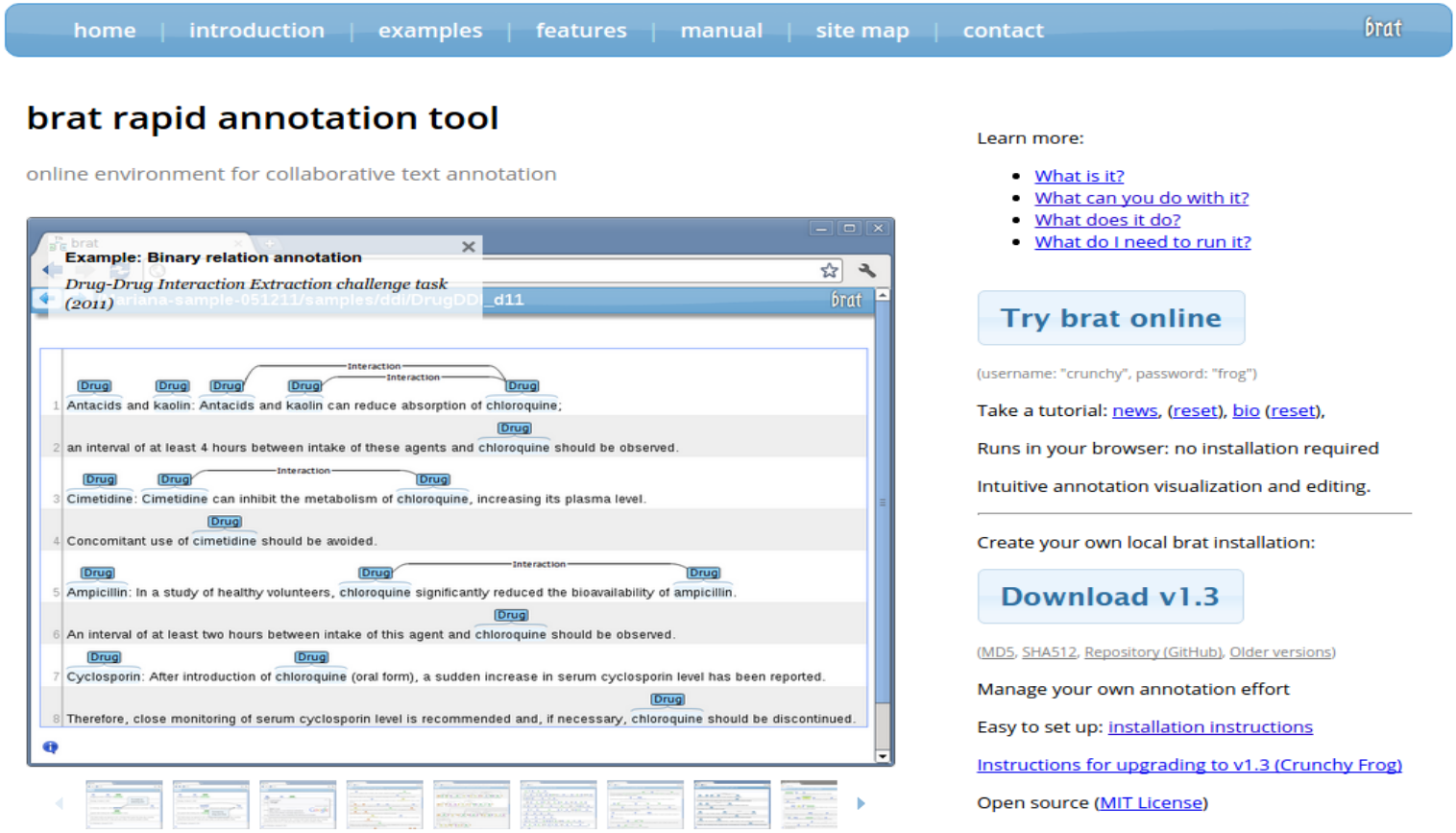
Development of the annotation tool

- Step 1: Selection of documents
  - Upload documents by the user
    - Indexing of the document
  - Search local documents
    - Keywords
    - Entity types



Development of the annotation tool

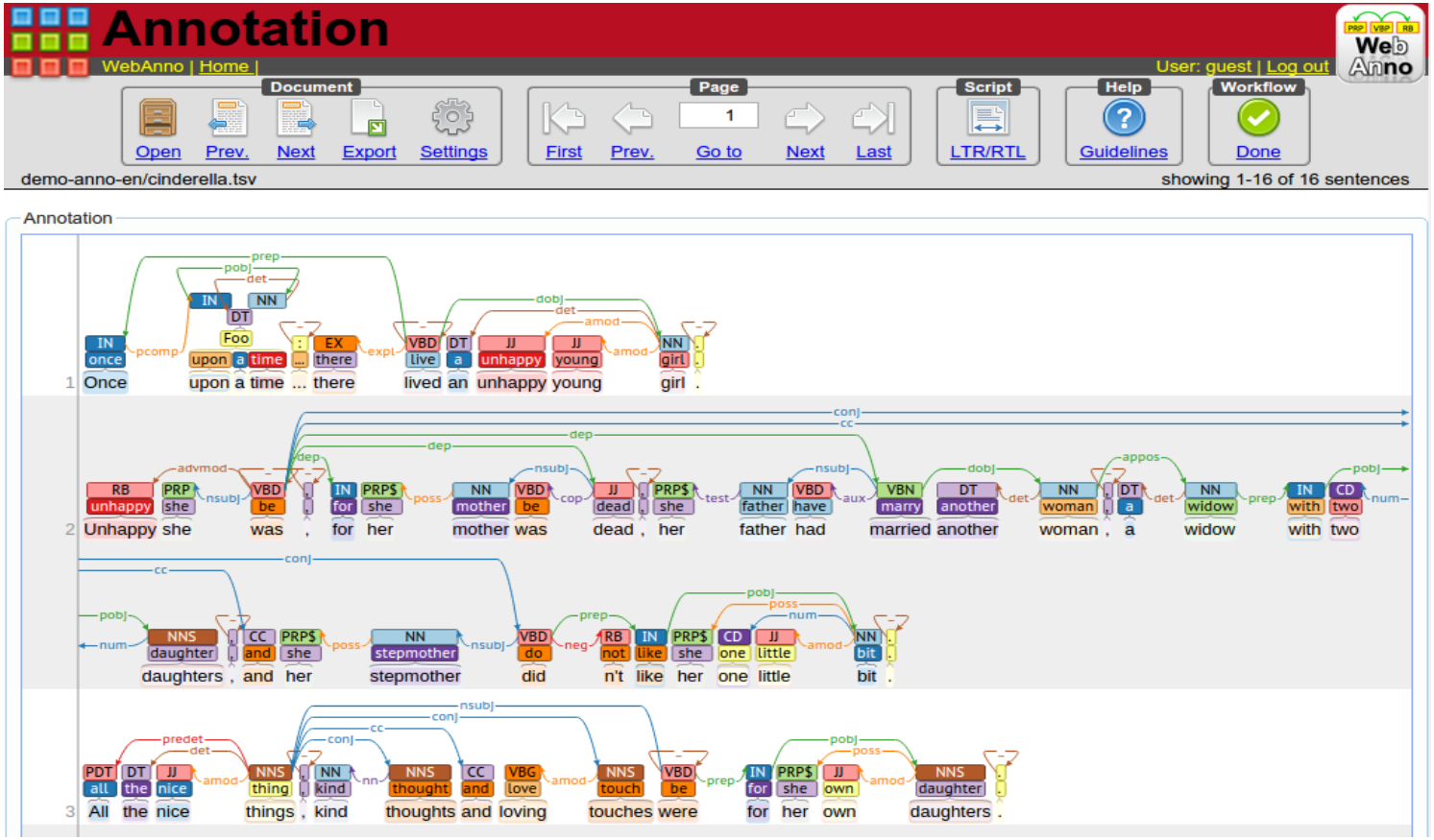
- Step 2: manual annotation of entities and relationships



(http://brat.nlplab.org/)

Development of the annotation tool

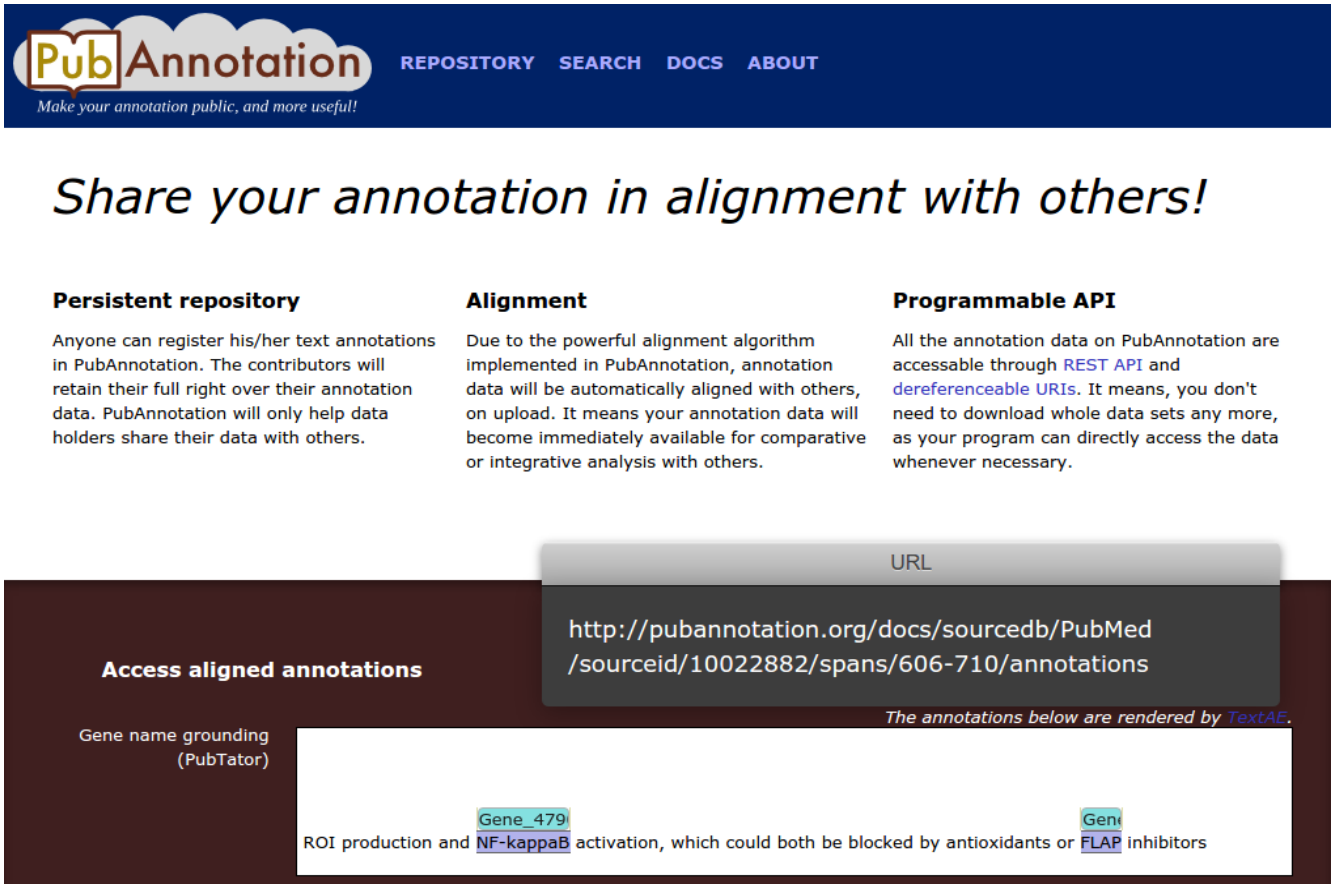
- Step 2: manual annotation of entities and relationships



(https://webanno.github.io/webanno/)

Development of the annotation tool

- Step 2: manual annotation of entities and relationships



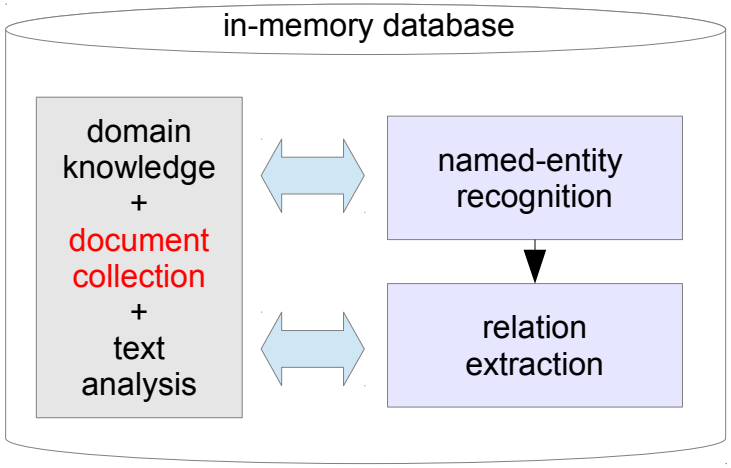
(http://textae.pubannotation.org/)

Integration of text mining

- Document collection

- SAP HANA instance of 1 Tb of memory (Future SOC Lab)

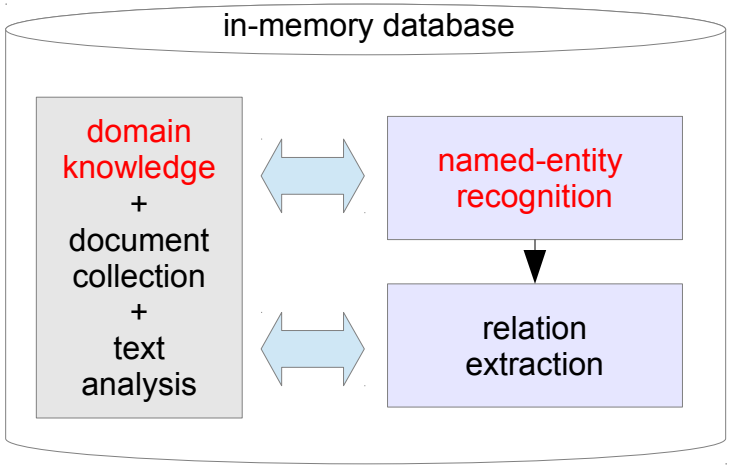
- Local copy of PubMed
  - Currently occupies 300 Gb
    - 7,2 millions of 24 millions citations
    - 38 millions sentences
    - 1,7 billions tokens (words)



(http://www.ncbi.nlm.nih.gov/pubmed)

Integration of text mining

- Named-entity recognition
  - Based on dictionaries

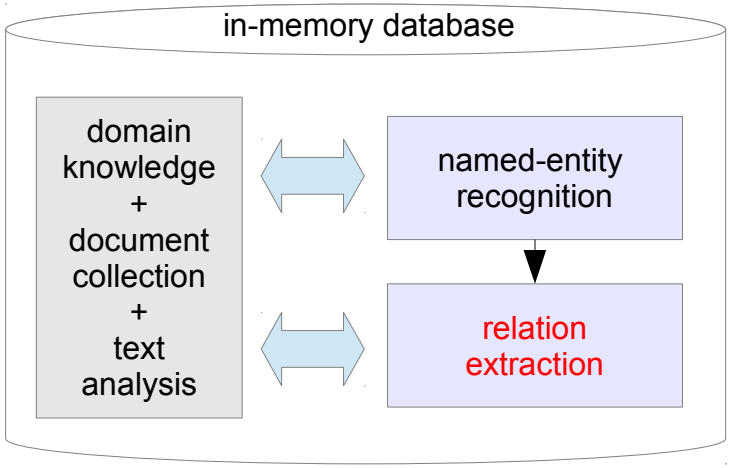


(http://www.uniprot.org/ http://www.drugbank.ca/ http://pubchem.ncbi.nlm.nih.gov/ http://www.nlm.nih.gov/research/umls/ http://www.ncbi.nlm.nih.gov/mesh)

Integration of text mining

- Relation extraction
  - Co-ocurrence
  - Machine learning (training data)

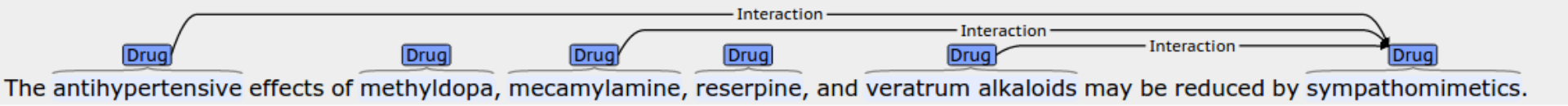
Valproic acid diminished binding of phenobarbital by a net change of 9.9% (percentage increase in FDF, 21.2%) at 1732 micromol/L.



Evaluation on bechmarks

- Benchmarks

Extraction of Drug-Drug Interactions from BioMedical Texts



(http://corpora.informatik.hu-berlin.de/index.xhtml#/ddi/train/DrugDDI_d105
https://www.cs.york.ac.uk/semeval-2013/task9/)

Evaluation on bechmarks

- Use cases

SABIO-RK
Biochemical Reaction Kinetics Database

General information	
Organism	Homo sapiens
Tissue	liver
EC Class	3.5.3.1
SABIO reaction id	574
Variant	mutant H101N activated
Recombinant	expressed in Escherichia coli
Experiment Type	in vitro
Pathways	Arginine and Proline metabolism Insulin signaling pathway Urea cycle
Event Description	-

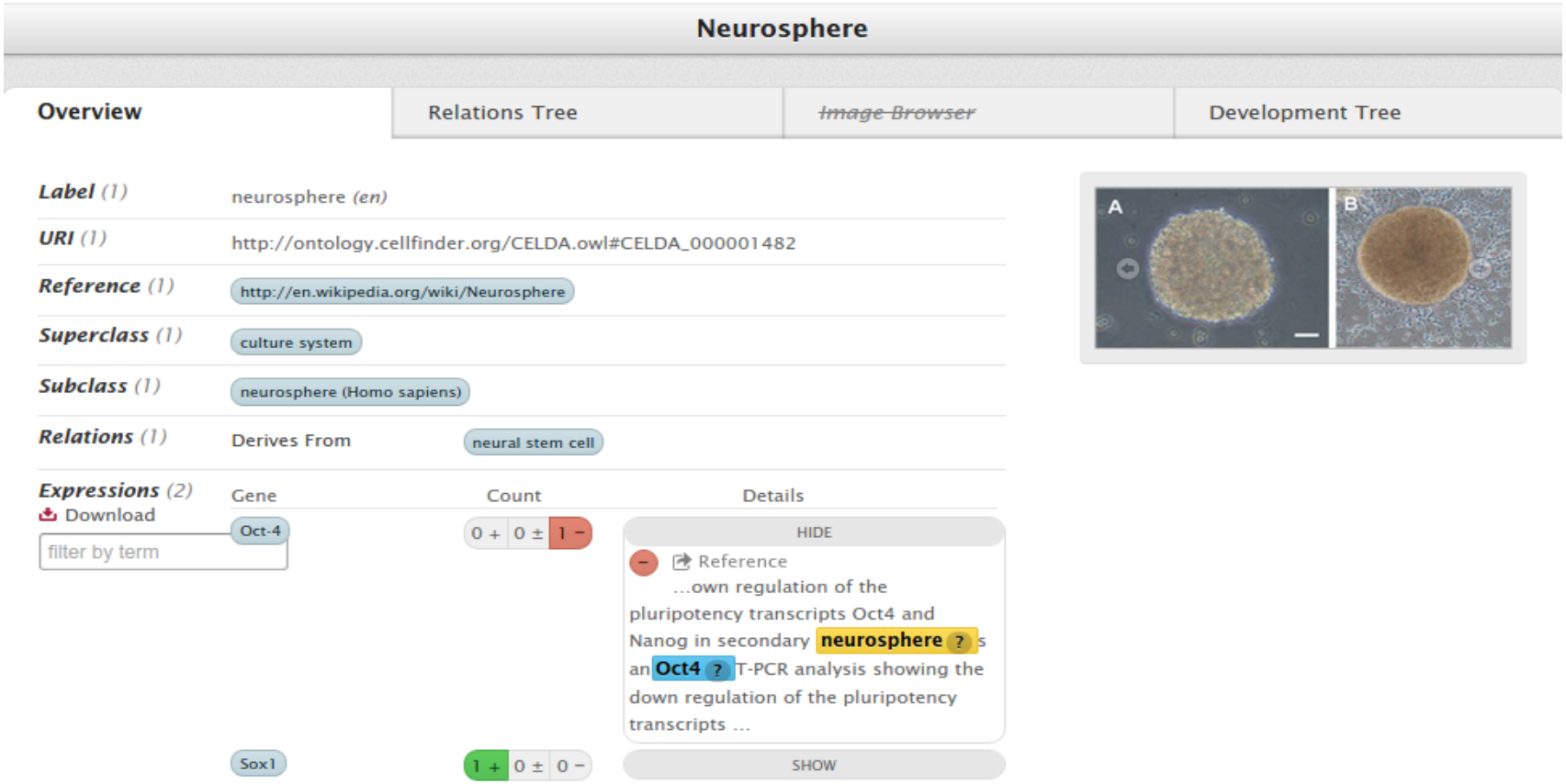
Substrates		
name	**location**	**comment**
H2O	-	-
L-Arginine	-	-

Products		
name	**location**	**comment**
L-Ornithine	-	-
Urea	-	-

(http://sabio.villa-bosch.de/)

Evaluation on bechmarks

- Use cases



(http://cellfinder.org/ http://hpscreg.eu/)

Submission of a paper

- Publication of the annotation tool
  - Journals
    - Database

  - Workshops and Conferences
    - Workshops in ACL'16
    - Poster in the BioCuration'16
    - Poster/Demo IUI'16

ACL 2016
AUGUST 7 – 12 | BERLIN, GERMANY

ISB International Society for Biocuration

iui 2016

(http://database.oxfordjournals.org/ http://acl2016.org/
http://www.isb-sib.ch/events/biocuration2016/ http://iui.acm.org/2016/)

Master project „Learning to Note“

- Grading
  - Commitment: 10%
  - Presentation: 10% (mid-term, final)
  - Implementation: 40%
  - Paper: 40%

Master project „Learning to Note“

- Contact: V0.01 (Villa) or 2-02.1 (Haus E), HPI Campus II

Dr. Mariana Neves

mariana.neves@hpi.de

Dr. Ralf Krestel

ralf.krestel@hpi.de