Annotating German Clinical Documents for De-Identification

Tobias Kolditz,[a] Christina Lohr,[a] Johannes Hellrich,[a] Luise Modersohn,[a]
Boris Betz,[b] Michael Kiehntopf,[b] Udo Hahn [a]

[a] Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena

[b] Department of Clinical Chemistry and Laboratory Diagnostics and Integrated Biobank Jena (IBBJ), Jena University Hospital

Aug 29, 2019 – T1-08 De-identification

Clinical text data

Discharge Summary

Provider: Ken Cure, MD

Patient: Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

Attachment Control Number: XA728302

personal data

HOSPITAL DISCHARGE DX

- 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- 163.8 Other specified sites of pleura.

HOSPITAL DISCHARGE PROCEDURES

1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that was originally diagnosed in the early 70's. At that time she had a radical mastectomy with postoperative radiotherapy. In the mid 70's she developed a chest wall recurrence and was treated with further radiation therapy. She then went without evidence of disease for many years until the late 80's when she developed bone metastases with involvement of her sacroiliac joint, right trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done well until recently when she developed shortness of breath and was found to have a larger pleural effusion. This has been tapped on

PHI categories

1	Names
2	Location
3	Dates
4	Phone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Securitiy numbers
8	Medical Record Numbers
9	Health plan beneficary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers
13	Device identifiers
14	URLs
15	IP numbers
16	Biometric identifiers
17	Images
18	Other

International Research

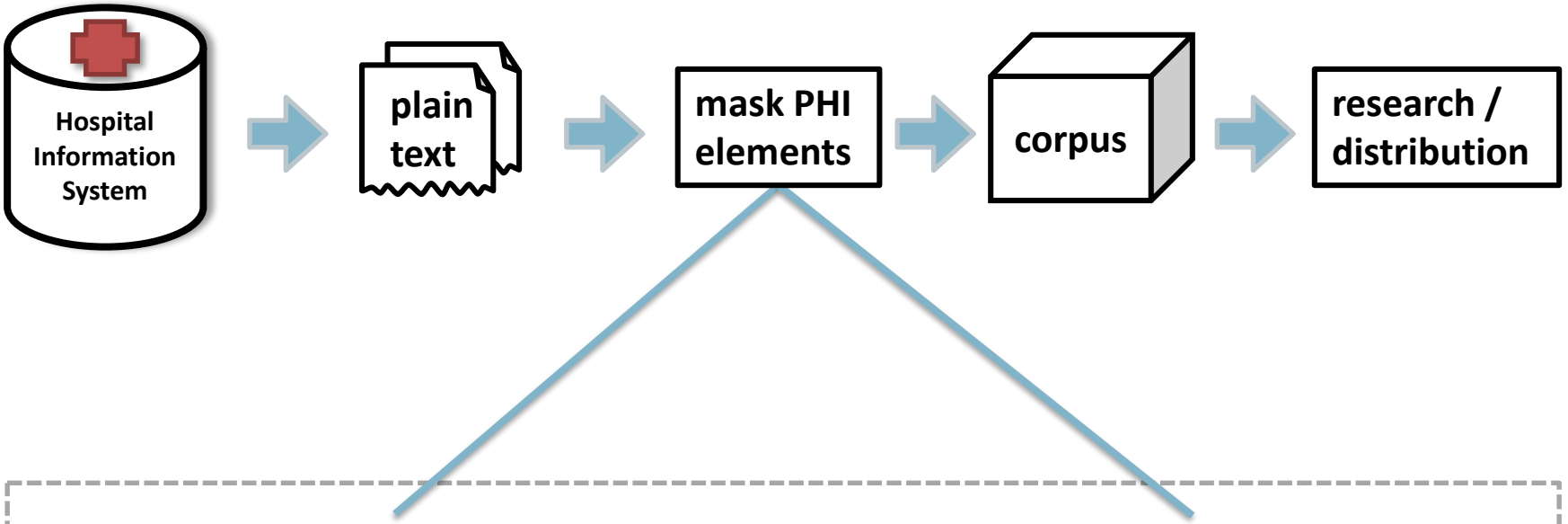
- US Health Insurance Portability and Accountability Act (HIPAA) (1996): **Personal Health Information (PHI) criteria**
- text corpora from **i2b2 de-idenfication challenges** in 2006 and 2014

Situation in Germany

- **data protection defined by law**
- **no specification defined (no equivalent to HIPAA PHI)**

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Corpus Development



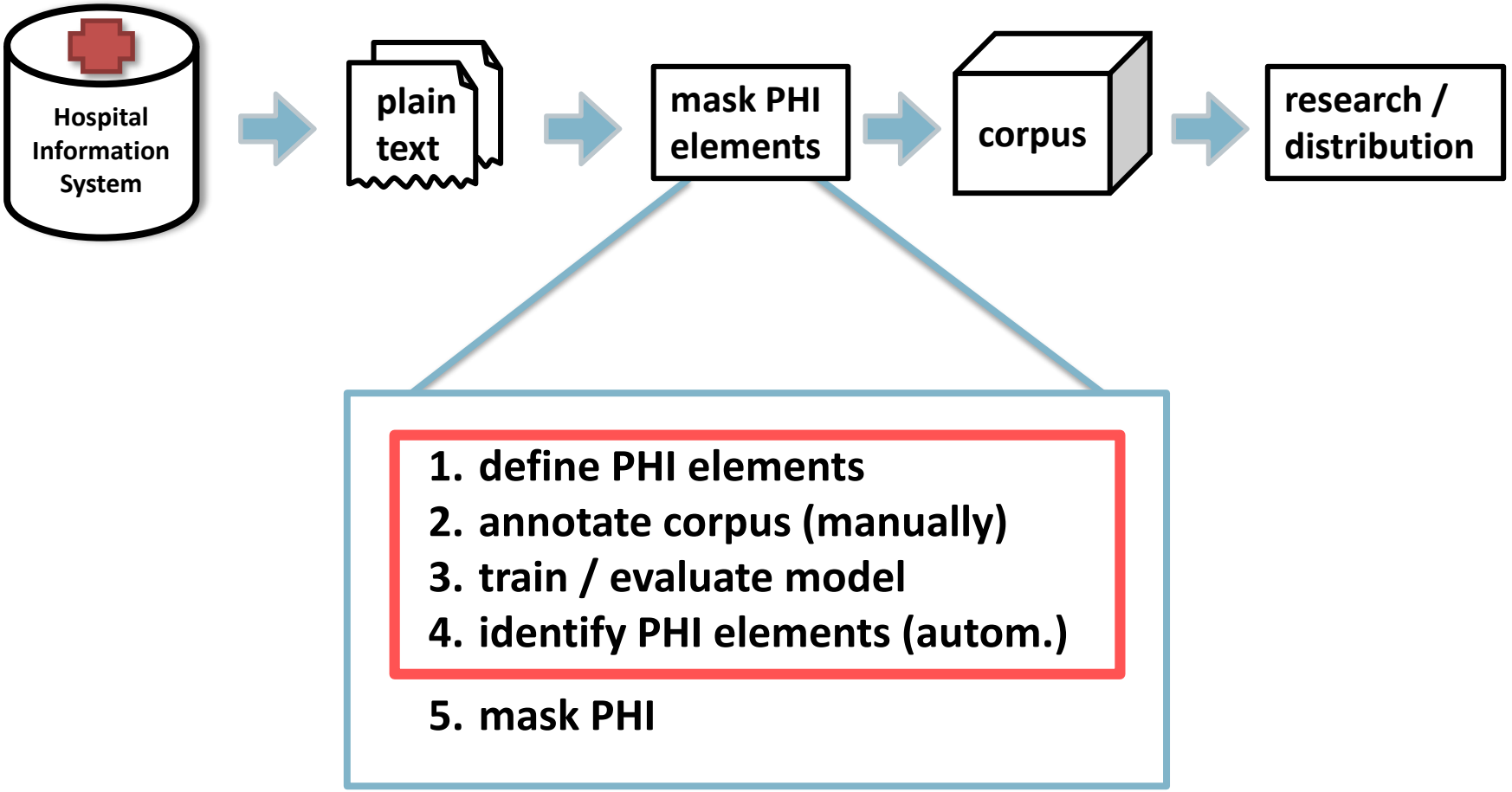
anonymization

- Patient name: **XXX XXXXX**
- In our institution from **XXX XX** until **XXX XX XXXX**

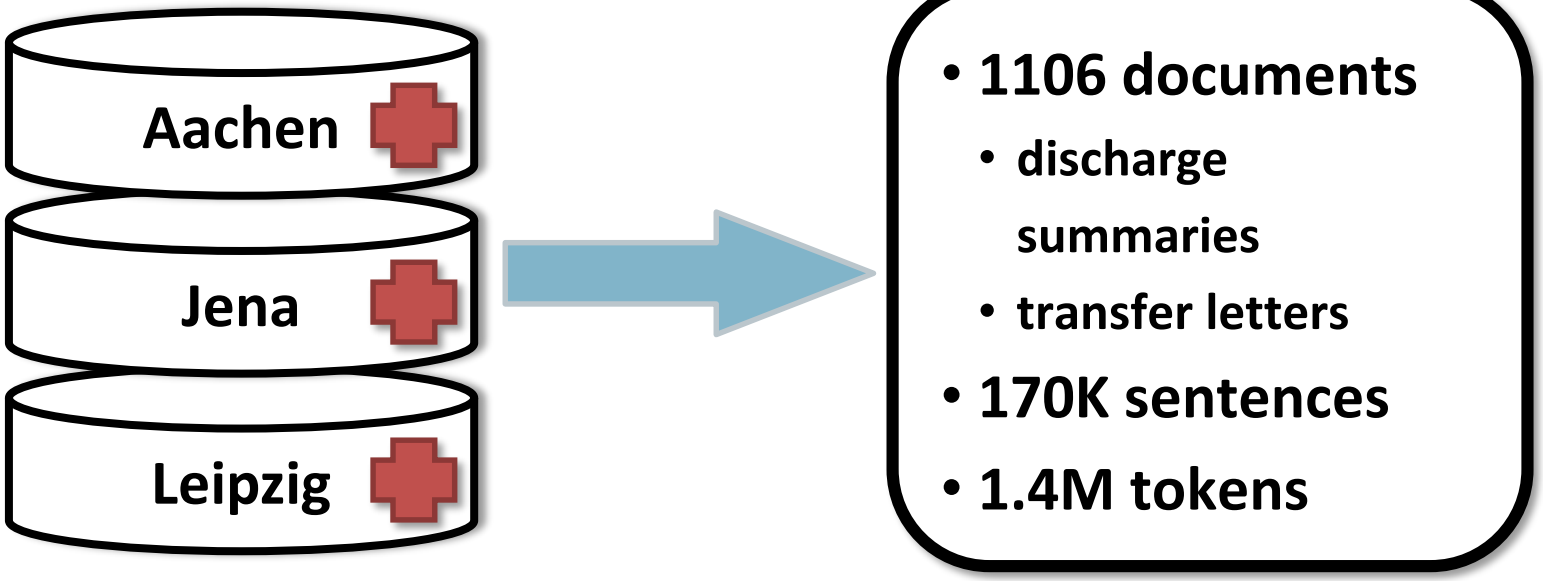
pseudonymization

- Patient name: **John DOE**
- In our institution from **AUG 25** until **AUG 29 2019**

Corpus Development



3000PA Corpus (Hahn et al., MIE 2018)



German Clinical Reference Text Corpus

- 2010-2015
- internistic or ICU units
- patients deceased

Redefinition of PHI categories

1	Names
2	Location
3	Dates
4	Phone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Securitiy numbers
8	Medical Record Numbers
9	Health plan beneficary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers
13	Device identifiers
14	URLs
15	IP numbers
16	Biometric identifiers
17	Images
18	Other

Adopted and new sub-categories

- ✓ Person/Name
  - ➔ Patient
  - ➔ Relative
  - ➔ Staff

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Redefinition of PHI categories

1	Names
2	Location
3	Dates
4	Phone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Securitiy numbers
8	Medical Record Numbers
9	Health plan beneficary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers
13	Device identifiers
14	URLs
15	IP numbers
16	Biometric identifiers
17	Images
18	Other

- ✓ Person/Name
  - ➔ Patient
  - ➔ Relative
  - ➔ Staff

Adopted

- ✓ Date
- ✓ Location
- ✓ Other

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Redefinition of PHI categories

1	Names
2	Location
3	Dates
4	Phone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Securitiy numbers
8	Medical Record Numbers
9	Health plan beneficary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers
13	Device identifiers
14	URLs
15	IP numbers
16	Biometric identifiers
17	Images
18	Other

- ✓ Person/Name
  - → Patient
  - → Relative
  - → Staff
- ✓ Date
- ✓ Location
- ✓ Other

Reorganized categories

- → Contact
- → Id

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Redefinition of PHI categories

1	Names
2	Location
3	Dates
4	Phone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Securitiy numbers
8	Medical Record Numbers
9	Health plan beneficary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers
13	Device identifiers
14	URLs
15	IP numbers
16	Biometric identifiers
17	Images
18	Other

- ✓ Person/Name
  - → Patient
  - → Relative
  - → Staff
- ✓ Date
- ✓ Location
- ✓ Other
- → Contact
- → Id

Excluded

- ✗ Biometric identifiers
- ✗ Images

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Redefinition of PHI categories

1	Names
2	Location
3	Dates
4	Phone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Securitiy numbers
8	Medical Record Numbers
9	Health plan beneficary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers
13	Device identifiers
14	URLs
15	IP numbers
16	Biometric identifiers
17	Images
18	Other

- Person/Name
  - Patient
  - Relative
  - Staff
- Date
- Location
- Other
- Contact
- Id

New categories

➔ Age
➔ Birthday
➔ Typist
➔ MedicalUnit

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Two-phased annotation

Prerequisites

- annotation tool: BRAT
- annotators: 8 medical students and 2 physicians

1st Phase – *Preparation*

- generic annotation of PHI items – 1 annotation category
- pre-annotation of **DATE** by regular expressions
- no agreement calculation

2nd Phase – *Categories*

- use pre-annotation from 1st Part
- PHI items
- 2 training iterations
- 12-25 agreement documents, 50 final agreement documents

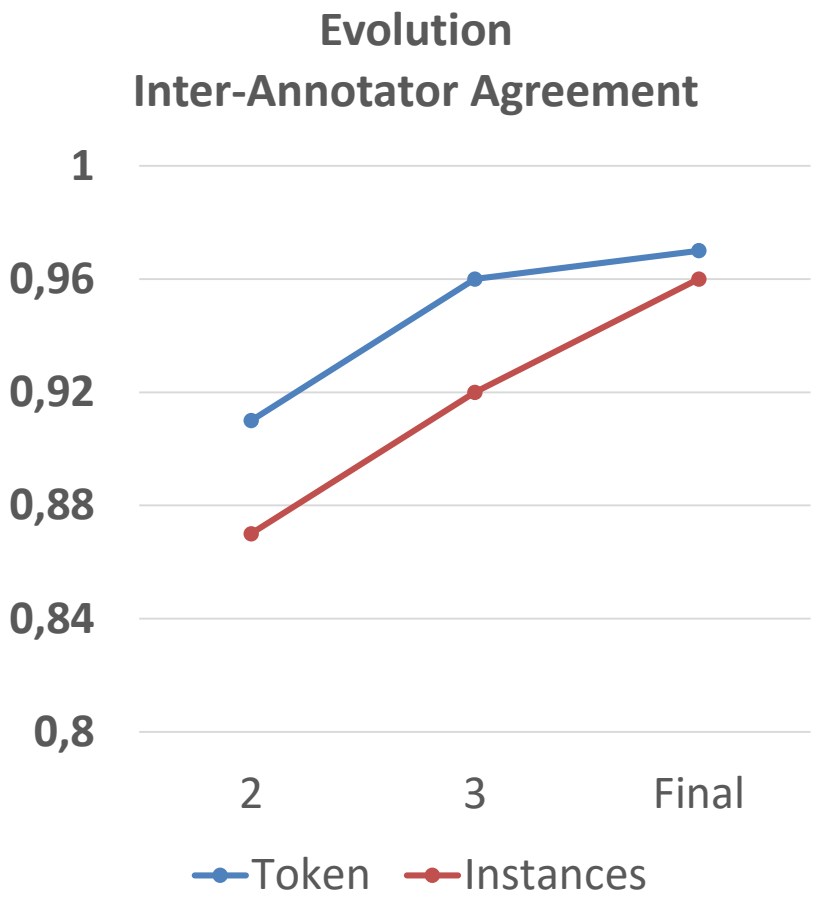
Inter-Annotator Agreement (IAA)

- pair-wise average F-Score
  - partial match: tokens
  - exact match: instances

- ***"Jane Smith"***
  - 1 instance
  - 2 tokens

➢ Ann. 1: Jane ***"Smith"***

➢ Ann. 2: ***"Jane Smith"***
  - 1 partial match
  - 0 exact match



Final Annotation: Corpus and Results

Agreement

- 50 documents (≈ 5%)
- used annotations from person with highest IAA

Final Corpus

- 1.4M token
- 44,167 ann. instances
- 69,042 ann. tokens
  (≈ 5% of all tokens)

Category	Instances		
	Frequency		IAA
Date	20,603	46,6%	0.98
MedicalUnit	6,189	14,0%	0.90
Location	5,429	12,3%	0.98
Staff	5,231	11,8%	0.95
Patient	3,180	7,2%	0.99
Birthdate	1,103	2,5%	1.00
misc.*	2,432	5,5%	0.28—1.00

* Typist, Contact, Age, ID, Other, Relative, Person

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Final Annotation: Corpus and Results

Date

- highest frequency
- different formats, e.g.:
  - 29.08.
  - 8/2019
  - 29.08.2019/9:10 Uhr

Birthdate

- IAA = 1.0
- not mentioned in all documents

Category	Instances		
	Frequency		IAA
Date	**20,603**	**46,6%**	**0.98**
MedicalUnit	6,189	14,0%	0.90
Location	5,429	12,3%	0.98
Staff	5,231	11,8%	0.95
Patient	3,180	7,2%	0.99
Birthdate	**1,103**	**2,5%**	**1.00**
misc.*	2,432	5,5%	0.28—1.00

* Typist, Contact, Age, ID, Other, Relative, Person

Final Annotation: Corpus and Results

MedicalUnits

- different types of medical units
  - Station 123
  - colleagues of oncology
  - Clinic of Surgery II

Location

- data highly diverse, e.g.:
  - Fürstengraben 27/30
    D-07743 Jena

Category	Instances		
	Frequency		IAA
Date	20,603	46,6%	0.98
MedicalUnit	**6,189**	**14,0%**	**0.90**
Location	**5,429**	**12,3%**	**0.98**
Staff	5,231	11,8%	0.95
Patient	3,180	7,2%	0.99
Birthdate	1,103	2,5%	1.00
misc.*	2,432	5,5%	0.28—1.00

* Typist, Contact, Age, ID, Other, Relative, Person

Final Annotation: Corpus and Results

Staff

- a lot of titles
  - Prof. Dr. med. John Smith
  - OA Miller

Patient Name

- mostly in head of document
  - John Smith
  - Mister Smith

Category	Instances		
	Frequency		IAA
Date	20,603	46,6%	0.98
MedicalUnit	6,189	14,0%	0.90
Location	5,429	12,3%	0.98
Staff	**5,231**	**11,8%**	**0.95**
Patient	**3,180**	**7,2%**	**0.99**
Birthdate	1,103	2,5%	1.00
misc.*	2,432	5,5%	0.28—1.00

* Typist, Contact, Age, ID, Other, Relative, Person

Neural Network Baseline Classifier

- randomly sampled
  - 80% training
  - 20% test
- unidirectional LSTM learning word representations based on character embeddings
- bidirectional LSTM:
  - input: character-based representations concatenated with word embeddings
- avg. F1-Score: **0.952**

	Instances		
Type	**Frequency**	**IAA**	**F1-Score**
Birthdate	1,103	**1.00**	**0.975**
Staff	5,231	**0.95**	**0.968**
Date	20,603	**0.98**	**0.964**
Location	5,429	**0.98**	**0.958**
Patient	3,18	**0.99**	**0.957**
Contact	613	**0.97**	**0.956**
MedicalUnit	6,189	**0.90**	**0.952**
misc.*	1819		0.4–0.9

* Typist, Contact, Age, ID, Other, Relative, Person

Conclusion

- HIPAA PHI adapted annotation schema for de-identification of German discharge summaries
- annotation on other heterogeneous data
- prerequisite for a pseudonymization engine

	i2b2 2006	i2b2 2014	3000 PA (Jena)
documents	**889**	**1304**	**1106**
IAA instances	**–**	**.892**	**.96**
IAA token	**–**	**.928**	**.97**
classifier	**.967**	**.9586**	**.952**

MEDINFO 2019 www.medinfo-lyon.org CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

Annotating German Clinical Documents for De-Identification

christina.lohr@uni-jena.de

www.julielab.de

Jena University Language & Information Engineering Lab

SMITH Smart Medical Information Technology for Healthcare

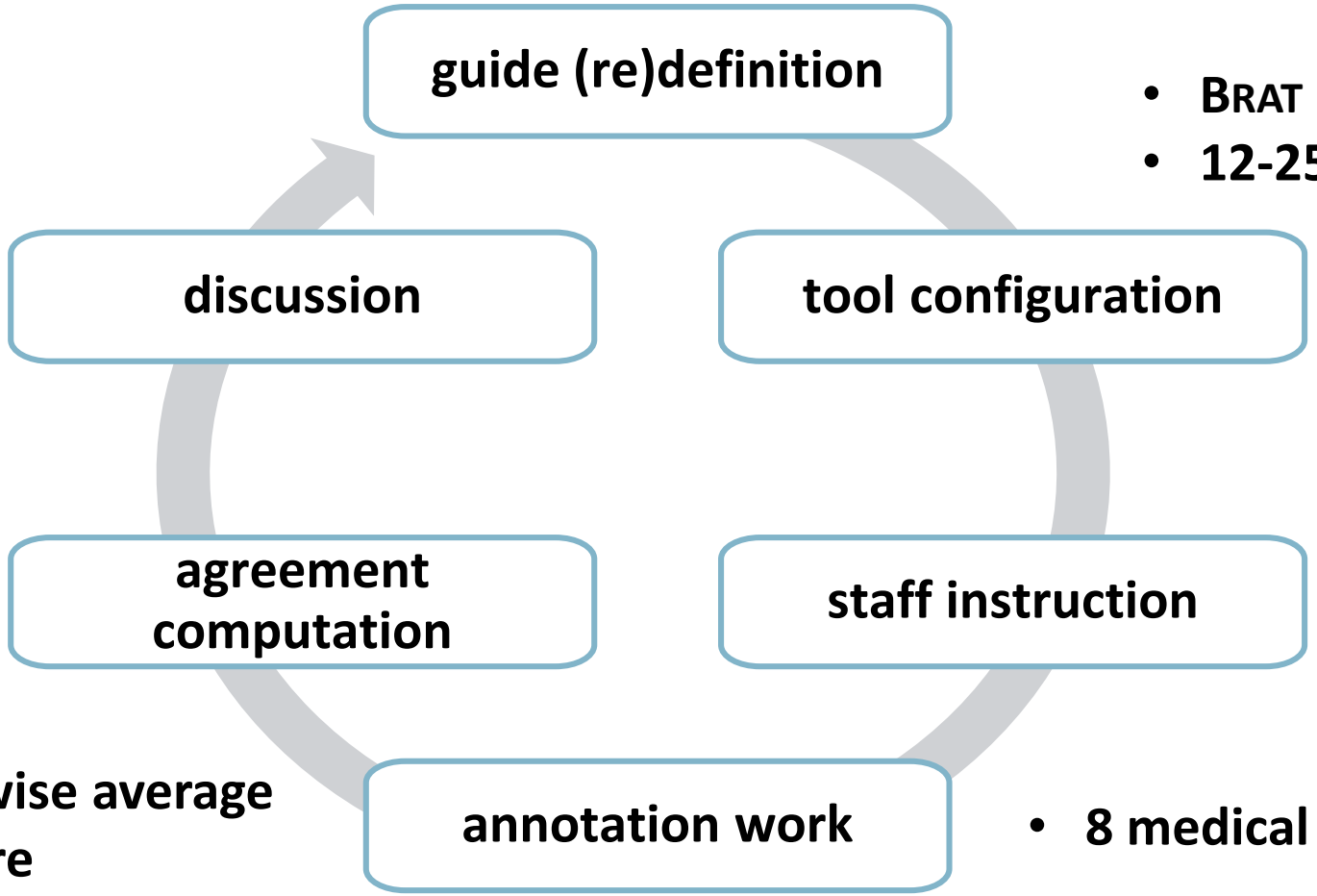
Federal Ministry of Education and Research

DFG

This work was funded by the German Federal Ministry of Education and Research (BMBF) within the SMITH project and the Deutsche Forschungsgemeinschaft (DFG) under within *STAKI²B²* project.

	i2b2 2006	i2b2 2014	3000 PA (Jena)
	discharge summaries	longitudinal clinical narratives	discharge summaries transfer letters
token	472331	805118	1477506
token per file	531	617	1336
PHI token	28204	–	69042
PHI instances	19498	28872	44167
IAA instances	**–**	**0.892**	**0.96**
IAA token	**–**	**0.928**	**0.97**
documents	**889**	**1304**	**1106**
PHI instances per file	22	22	40
Classificator	**.967**	**.9586**	**.952**
	Decision Tree	**CRF-based**	**Neural Network**

Annotation – Iterative Training Process



- BRAT
- 12-25 docs

- pair-wise average F-Score

- 8 medical students

Inter-Annotator Agreement

Category	Type	Instances		Tokens	
		Frequency	Avg. F1	Frequency	Avg. F1
Age	Age	498	1.00	500	1.00
Contact	Contact	613	0.97	2,009	0.98
Date	Date	20,603	0.98	24,277	0.99
	Birthdate	1,103	1.00	1,103	1.00
ID	ID	398	0.81	424	0.82
	Typist	655	0.86	1,418	0.93
Location	Location	5,429	0.98	11,286	0.99
MedicalUnit	MedicalUnit	6,189	0.90	12,499	0.95
Person	Person	14	-	23	-
	Patient	3,180	0.99	5,167	1.00
	Relative	36	0.80	62	0.88
	Staff	5,231	0.95	10,003	0.97
Other	Other	218	0.28	271	0.26
Total	*	44,167	0.96	69,042	0.97