Evaluation der computer-linguistischen Texterschließung neuro-radiologischer Befunde im Berliner BFG-Projekt

KIS-RIS-PACS- und DICOM-Treffen Mainz – Schloß Waldthausen 21. Juni 2013

Josef Schepers, Peter Geibel, Thomas Tolxdorff (Charité)

unter Mitwirkung von
Gabriel Curio, Hebun Erdur, Kati Jegzentis,
Stefan Krüger, Frank Müller, Christian Nolte,
Thorsten Schaaf, Lothar Zimmermann (Charité)
Martin Trautwein et al. (Vivantes)
Christian Seebode et al. (ORTEC medical)

Project Funded by TSB Technologiestiftung Berlin
-Zukunftsfonds Berlin
-Co-financed by the European Union – European Fund for Regional Development

BFG-Projektpartner

- Charité – Universitätsmedizin Berlin
  - größtes Universitätsklinikum Deutschlands (4 Standorte in Berlin)
  - stat. Fälle 140.000 / amb. Fälle 560.000 / Betten 3.213 / MA 13.500

- Vivantes – Netzwerk für Gesundheit GmbH, Berlin
  - größter kommunaler Krankenkenhauskonzern (9 Häuser in Berlin)
  - stat. Fälle 210.000 / amb. Fälle 270.000 / Betten 5.329 / MA 13.500

- ORTEC medical GmbH
  - Berliner IT-Unternehmen

Aufgabenstellungen des BFG-Projektes

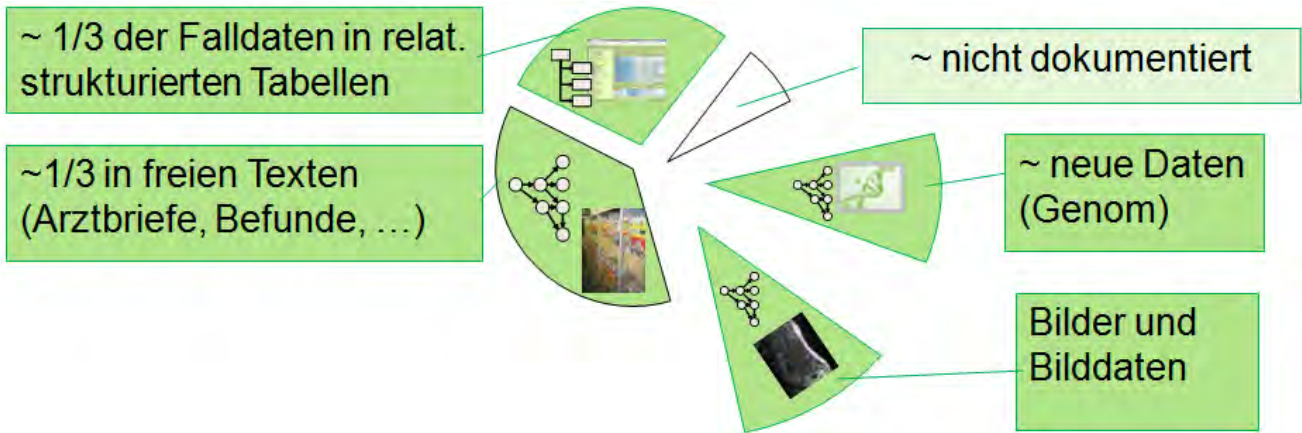
- Erforschung medizinisch-linguistischer Texterschließung (Arztbriefe, Befunde, Epikrisen, Verlegungsberichte…)
- Entwicklung kommerzieller Studiensoftware
- Sekundäre Datennutzung für Machbarkeitsanalysen klinischer Studien
- Sekundäre Datennutzung für Studienrekrutierung
- Vorfüllung von Case Report Forms (eCRFs) für klinische Studien

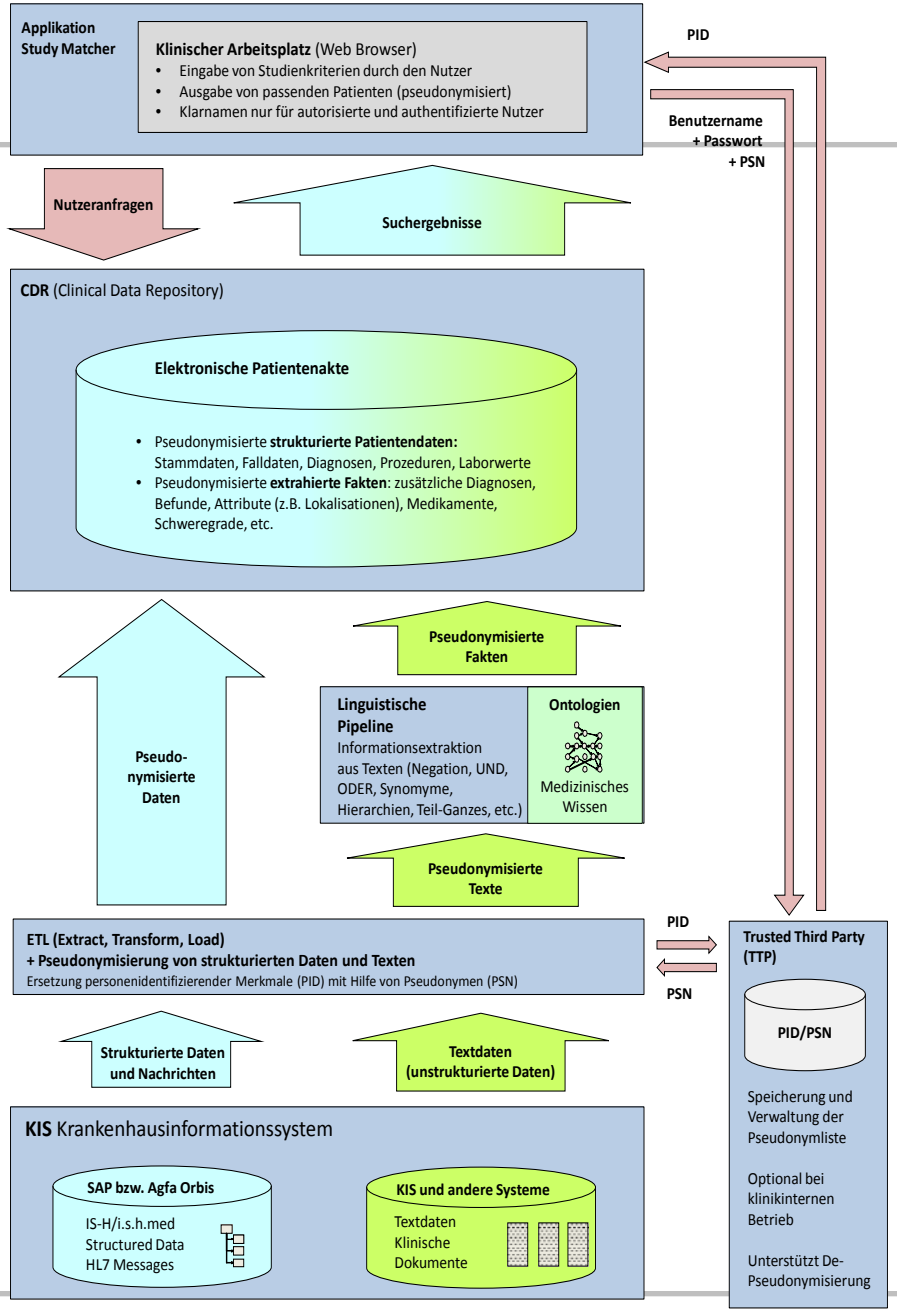
- … *Beitrag zum Senatsziel „Gesundheitsmetropole Berlin“*

„Secondary Use“: Episodenwissen / Faktenwissen der Patientenversorgung als Quelle der klinischen Forschung

- **Episoden-/Faktenwissen umfasst:**
  - Diagnosen, Symptome, Befunde, Prozeduren, Medikamente…
  - Morphologie, Topologie, Komplikationen, Heilungsdauer…
  - Aufenthaltsdaten, Demografische Merkmale...

- **Episodenwissen vorhanden in:**
  ~ ein Drittel in strukturierten Daten (KIS, Labor, etc.)
  ~ ein Drittel in Freitexten (Arztbriefe, Befunde…)
  ~ ein Drittel durch Nacherhebung





Forschungsdatawarehouse
„Clinical Data Repository CDR“

Projektschwerpunkt
„Linguistische Pipeline“

Datenquellen
- Strukturierten Daten und
- Freitextinformationen

Kernelement des BFG-Projektes: die linguistische Verarbeitungskette

„Im MRT kommt ein akuter MCA-Verschlus rechts und ein chronischer ACI-Infarkt lniks zur Darstellung.“

[hatBeobachtung:Infarkt[hatLokalisation:ArteriaCarotisInterna[],hatBeobachtungsAttribut:chronisch[],hatBeobachtungsAttribut:MRT[],hatBeobachtungsAttribut:Sinister[]],hatBeobachtung:Verschluss[hatBeobachtungsAttribut:Akut[],hatBeobachtungsAttribut:Dexter[],hatBeobachtungsAttribut:MRT[],hatLokalisation:ArteriaCerebriMedia[]]]

Tippfehlerkorrektur, Lemmatisierung, Stemming, Ellipsen, …

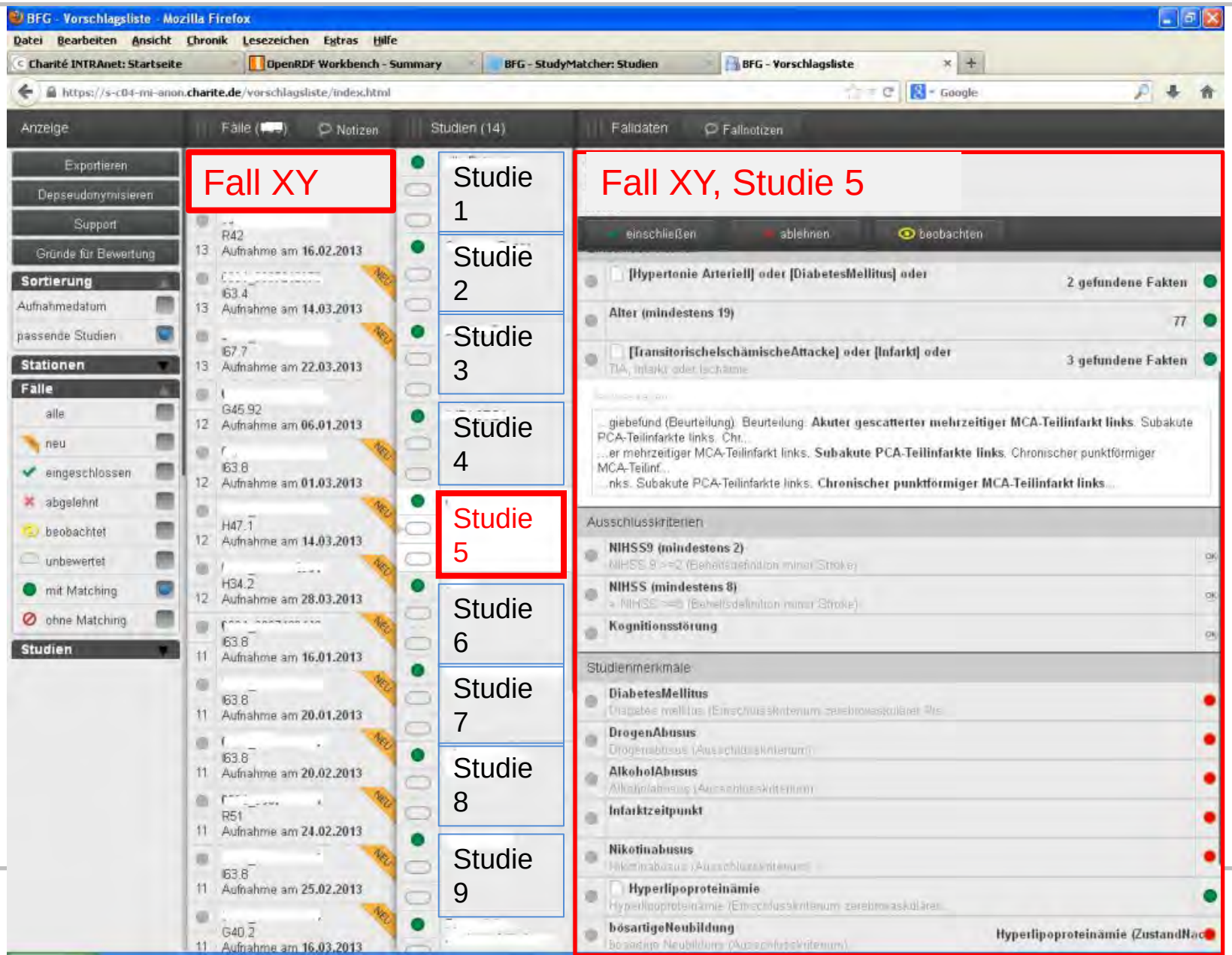
Hierarchische Phrasenstruktur

Mapping von Phasen-Segmenten auf Substrukturen der Ontologie

Elemente der morpho-syntaktischen Analyse und der semantischen Annotation:

- **Identifikation von Negationen**
- **Skopus der Negation** (nicht einfach und komplex)
- **Instantiierung = konkrete Realisierungen eines Wortes im Kontext**
  Krampf: Krampf, Krampfs, Krampfes, Krämpfen, ...
- **Auflösung von Komposita**
  Segmentierung von „Muskelkrämpfe“
  in /muskel/ und /krampf/
- **Zuweisung von Wortarten durch einen Tagger**
  Muster /ART ADJA* NN/ => Nominalphrase
  - die/ART rechte/ADJA Clavicula/NN
  - keine/ART Residuen/NN
- **Erkennung von syntaktischer Struktur durch einen Parser**
- **Parser nutzt Wortarten-Tags zu Phrasenerkennung**

Identifikation von Studien für Patienten



Evaluationsformen

- Subjektive Evaluation der Funktionen & Features (Faktor f)
  - Zufriedenheit mit Funktionsumfang
  - Zufriedenheit mit Benutzerfreundlichkeit etc.
- **Quantitative Evaluation des Informationsgehaltes (Faktor i)**
  - Intrinsische Evaluation (Syntax, Morphologie, Semantik)

    Trefferquote und Genauigkeit der Erkennung von Absätzen, Sätzen, Nominalphrasen, Worten (Instanzen), Komposita, Wortarten, Skopi und Bedeutung
  - **Extrinsische Evaluation (Patientenidentifikation)**

    Hier: Trefferquote und Genauigkeit der Erkennung von geeigneten Studien für Patienten / Patienten für Studien

Wahrheitsmatrix der Patientenidentifikation (extr. Evaluation)

		Tatsächliche (wahre) Klasse		
		positiv (krank, studien-kompatibel)	negativ (gesund, nicht-kompatibel)	Summe
	Summe	wP = TP + FN	wN = FP + TN	Gesamt
Ermittelt durch BFG-System	positiv/ identifiziert	TP: richtig positiv	FP: falsch positiv	iP = TP+FP
	negativ/ nicht-ident.	FN: falsch negativ	TN: richtig negativ	iN = FN+TN

Trefferquote, Sensitivität = Recall = TP / (TP + FN) = TP / wP

Genauigkeit, Pos. Präd. W.= Precision = TP / (TP + FP) = TP / iP

F-Score: gewichtetes Kombinationsmaß aus Recall und Precision

Evaluation

10 Peter Geibel et al.

Precision & Recall

Trial	n	TP	FP	TN	FN	P	R	S	N	F_1	F_5
Studie 1	280	1	106	171	2	0.01	0.33	0.62	0.99	0.02	[illegible]
Studie 2	280	50	132	87	11	0.27	0.82	0.40	0.89	[illegible]	0.76
Studie 3	280	0	5	275	0	0.00	NaN	0.98	[illegible]	NaN	NaN
Studie 4	280	1	15	262	2	0.06	[illegible]	0.95	0.99	0.11	0.29
Studie 5	169	1	21	147	[illegible]	[illegible]	1.00	0.88	1.00	0.09	0.55
Studie 6	273	57	[illegible]	[illegible]	15	0.35	0.79	0.46	0.86	0.48	0.75
Studie 7	[illegible]	39	77	135	28	0.34	0.58	0.64	0.83	0.43	0.57
Studie 8	279	17	19	243	0	0.47	1.00	0.93	1.00	0.64	0.96
Studie 9	279	13	25	236	5	0.34	0.72	0.90	0.98	0.46	0.69
Studie 10	280	1	28	251	0	0.03	1.00	0.90	1.00	0.07	0.48
Studie 11	280	1	235	44	0	0.00	1.00	0.16	1.00	0.01	0.10

Vorläufige Bewertung

- Gutes Zwischenergebnis
- System bereits heute brauchbar für:
  - Patientensuche / Studienvorschläge
  - Befundrecherche
- Verbesserungen notwendig für:
  - Machbarkeitsstudien (Feasibility)
  - Vorfüllung eCRFs (Case Report Form)
- Evaluation erfolgt kontinuierlich und differenziert
- Verbesserungen orientieren sich an Evaluationsergebnissen

Vielen Dank für Ihre Aufmerksamkeit

josef.schepers@charite.de

Kontakte BFG-Projekt:

Charité (Studien)	Vivantes (Studien)	ORTEC medical (Software)
bfg-info@charite.de ++49 (0) 30 - 450 544 544	alfred.holzgreve@vivantes.de ++49 (0) 30 - 130 142 901	info@bfg-berlin.de; info@ortec.org ++49 (0) 30 - 889 28 88-0 Am Sandwerder 37, 14109 Berlin