Bundesministerium
für Wirtschaft
und Energie

Trusted Cloud

Text-Mining in cloud4health – Ansätze und Ergebnisse

Dr. Philipp Senger, Fraunhofer SCAI

averbis
text analytics

Fraunhofer
SCAI

FAU FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

TMF

RHÖN-KLINIKUM
AKTIENGESELLSCHAFT

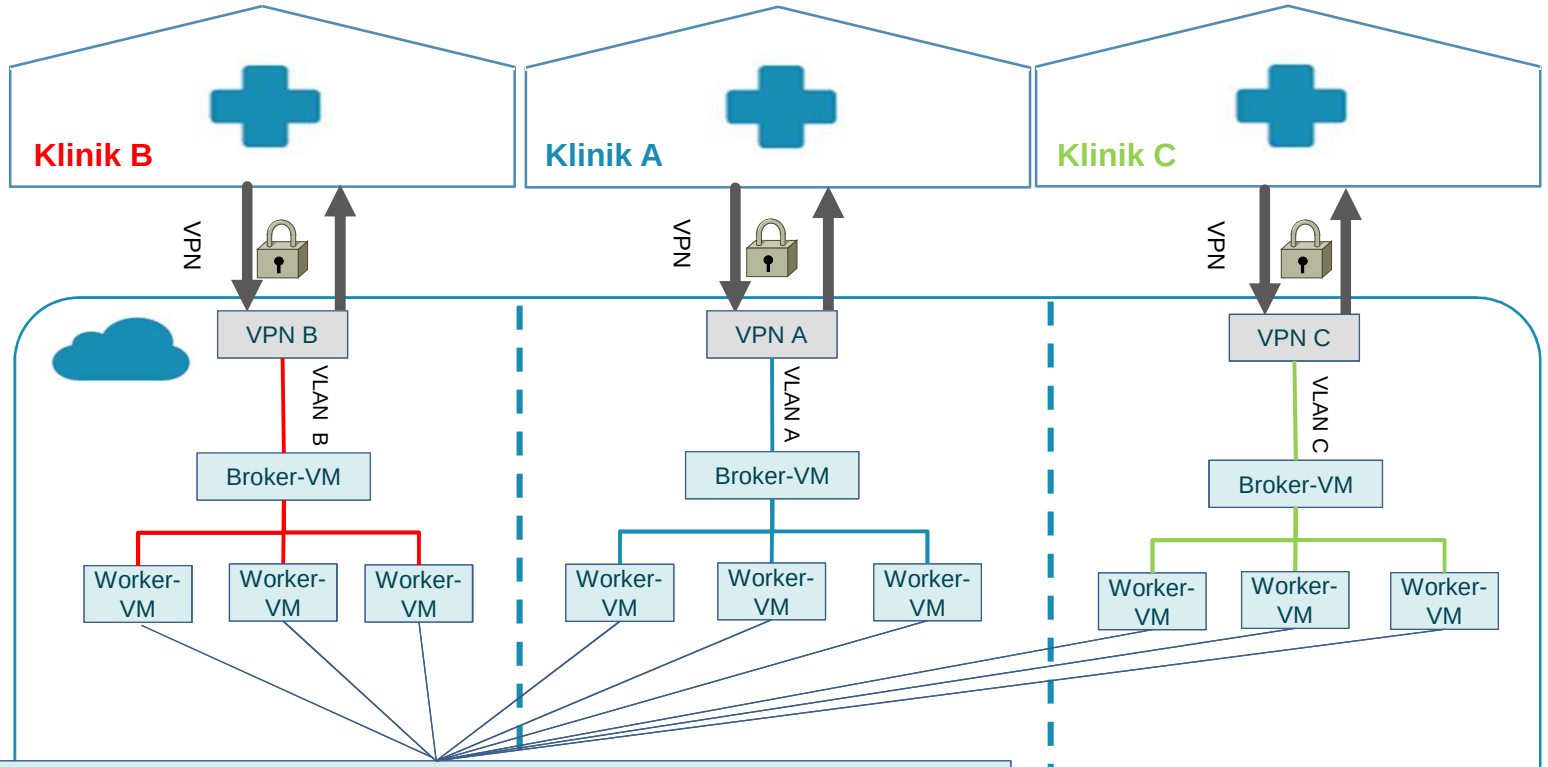
cloud4health

Expertise von Fraunhofer SCAI im Bereich des Text-Minings (Auszug)

Fraunhofer SCAI

- **Dictionary-based Named Entity Recognition via ProMiner in UIMA**
  - Namenserkennung (Gene/Protein, Krankheiten, Medikamente, ...
- **Machine Learning (teilweise in UIMA)**
  - Named Entity Recognition von komplexen Typen (z.B. chemischen Entitäten, IUPAC)
  - Relationsextraktion (z.B. Protein-Protein-Interaktionen)
  - (Dokumenten-)Klassifikation, Ähnlichkeiten und Clustering
  - Topic Modellierung
  - Sentiment Analysis
- **HPC Prozessierung von großen Dokumentenkorpora in UIMA**
  - PubMed 2015 in unter 6h
  - Patente
  - PDFs
  - Datenbanken
- **Semantische Suche**
  - www.SCAIView.com
  - Semantische Suche mit Dokumenten- und Entitätsranking

Text-Mining in cloud4health – Wo genau?



Text-Mining findet Use-Case-spezifisch in jeder Worker VM statt:

- UIMA Architektur
- Dokumentenbasierte Verteilung über die Worker
- Multithreading innerhalb jeder Worker-VM

Anwendungsmöglichkeiten

cloud4health eröffnet eine Vielzahl von Anwendungsmöglichkeiten für:

- Öffentliche Einrichtungen, Krankenkassen
- Industrie & Mittelstand (Medizin- und Biotechnik, Pharma)
- Krankenhäuser (öffentlich & privat)

Mögliche Bereiche für Anwendungsszenarien:

- Überprüfung klinischer Leitlinien,
- Qualitäts- / Kostenmonitoring, Feasibility studies
- Retrospektives Befüllen von Registern / Studien
- Patientenrekrutierung für Studien
- Strukturierung großer Datenmengen und Überführung in strukturierte Informationssysteme
- ...

cloud4health

Im Projekt wurden 4 konkrete Anwendungsszenarien umgesetzt:

- **Qualitätsmonitoring medizinischer Produkte**
  - Retrospektives Befüllen von Registern am Beispiel des Endoprothesenregisters Deutschland
- **Klinische Leitlinien/Plausibilität von Verordnungen**
  - Zusammenarbeit mit P^3 zur Überprüfung von Verordnungen im Bereich der Psychiatrie
- **Biodatenbanken**
  - Extraktion von Tumorgraduierung (z.B. TNM Kodierungen) aus großen Pathologiedatenbeständen
- **Pharmakovigilanz**
  - Detektion von Nebenwirkungen an verschiedenen Fallbeispielen

cloud4health

Im Projekt wurden 4 konkrete Anwendungsszenarien umgesetzt:

- **Qualitätsmonitoring medizinischer Produkte**
  - Retrospektives Befüllen von Registern am Beispiel des Endoprothesenregisters Deutschland
- **Klinische Leitlinien/Plausibilität von Verordnungen**
  - Zusammenarbeit mit P^3 zur Überprüfung von Verordnungen im Bereich der Psychiatrie
- **Biodatenbanken**
  - Extraktion von Tumorgraduierung (z.B. TNM Kodierungen) aus großen Pathologiedatenbeständen
- **Pharmakovigilanz**
  - Detektion von Nebenwirkungen an verschiedenen Fallbeispielen

Konkrete Anwendungsszenarien – Wie wurde vorgegangen?

cloud4health

Spezifikation des Anwendungsszenarios	• Definition der genauen Aufgabenstellung zusammen mit dem Kunden • Sichtung von prototypischen Datensätzen
Datensichtung und Freigabe	• Datenfreigabe des Kunden • Sichtung der Daten im Detail • Eventuell Anpassung der Aufgabenstellung
Technische Realisierung	• Auswahl der Methodik • Erste Realisierung des Prototypen • Iteration mit Kunden • Finale Version und Export in Cloud
Datenextraktion, Ergebnisse und Fazit	• Abschließende Auswertung • Performanzmessungen • Protokollierung der Ergebnisse • Abschluss

cloud4health

Im Projekt wurden 4 konkrete Anwendungsszenarien umgesetzt:

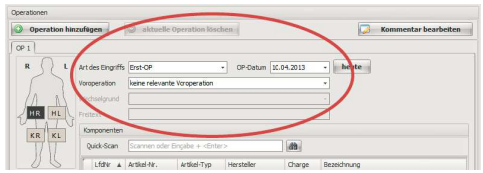
- **Qualitätsmonitoring medizinischer Produkte**
  - Retrospektives Befüllen von Registern am Beispiel des Endoprothesenregisters Deutschland
- **Klinische Leitlinien/Plausibilität von Verordnungen**
  - Zusammenarbeit mit P^3 zur Überprüfung von Verordnungen im Bereich der Psychiatrie
- **Biodatenbanken**
  - Extraktion von Tumorgraduierung (z.B. TNM Kodierungen) aus großen Pathologiedatenbeständen
- **Pharmakovigilanz**
  - Detektion von Nebenwirkungen an verschiedenen Fallbeispielen

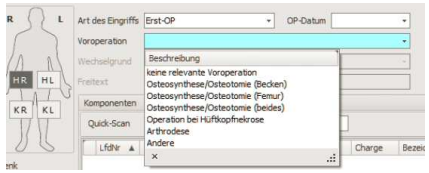
Endoprothesenregister – Spezifikation

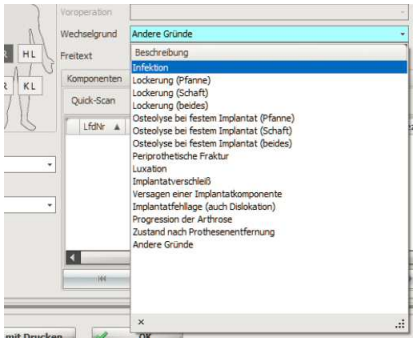
Angelehnt an den Vorgaben des Deutschen Endoprothesenregisters

- Bundesweites Register für Endoprothesen zur Ermöglichung von Qualitätskontrollen von künstlichen Hüft und Knieprothesen

Spezifikation der zu extrahierenden Datenpunkte:







- Gelenk (Knie/Hüfte), Seite (rechts/links)
- Art der Operation (Erst-OP, Wechsel),
- Voroperationen am endoprothetisch versorgten Gelenk
- Standzeiten bei Versorgung im gleichen Haus
- Gründe für Wechsel von Prothesenkomponenten
- Nebendiagnosen
- Komplikationen
- Schmerzinformation

Eigenschaften der unstrukturierten Eingabedaten:

- OP-Berichte und Entlassbriefe
- Insgesamt 70 Dokumente manuell über alle zu extrahierenden Klassen annotiert
- Mehrfach Annotation vermindert Fehlerquote
- Erstellter Goldstandart mit teilweise vielen möglichen Ausprägungen
- Goldstandard deckt Ausprägungen meist gut ab:
  - OP-Art: Erst-OP (41), Wechsel-OP (10), Re-operation (1), 18 unklar
  - Lokation: rechts (42), links (27), unklar (3)
  - Anatomie: Knie (21), Hüfte (36), unklar (13)
- Geringe Größe des Goldstandards rekrutiert sich demnach aus seiner Komplexität

Endoprothesenregister – Technische Realisierung

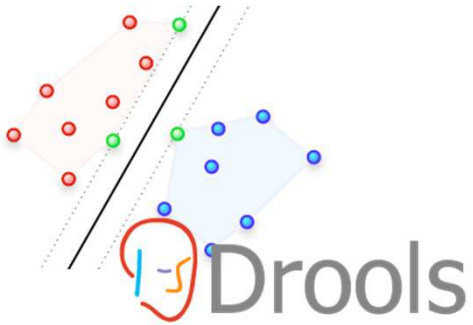
cloud4health

NER und maßgeschneiderte Terminologien:

- EPRD-spezifische, selbstdefinierte Konzepte
  - Lokalisation
  - Material
  - Hersteller
  - Schmerzen
  - ...
- Value-Sets aus Standard-Terminologien:
  - Diagnosen (ICD10)
  - Operationen (OPS/ICD10)
  - Anatomie (Radlex)

Regelbasiertes System und Maschinelles Lernen:

- Regelbasiertes System zur Extraktion bestimmter Datenpunkte anhand annotierter Entitäten
- Maschinelles Lernen zur Extraktion komplexerer/kontextspezifischer Datenpunkte



Endoprothesenregister – Technische Realisierung

cloud4health

```
– <ItemGroupData ItemGroupOID="de.cloud4health.odm.eprd.itemGroup.finding" ItemGroupRepeatKey="11">
    <ItemData ItemOID="de.cloud4health.odm.confidence" Value="1.0"/>
    <ItemData ItemOID="de.cloud4health.odm.evidence" Value="Hüft-TEP Luxation"/>
    <ItemData ItemOID="de.cloud4health.odm.eprd.finding.conceptId" Value="c4h.epr.finding.luxation"/>
  </ItemGroupData>
– <ItemGroupData ItemGroupOID="de.cloud4health.odm.eprd.itemGroup.operation" ItemGroupRepeatKey="12">
    <ItemData ItemOID="de.cloud4health.odm.confidence" Value="1.0"/>
    Value="Hüft-TEP"/>
    ion.conceptId" Value="5-820.0"/>

    eprd.itemGroup.operation" ItemGroupRepeatKey="10">
    Value="1.0"/>
    Value="Osteosynthese"/>
    ion.conceptId" Value="NA"/>
```

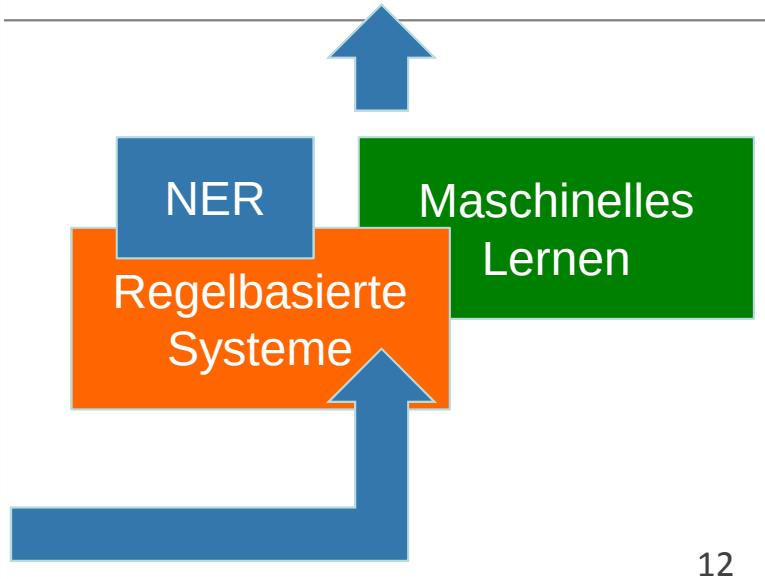
AnatomicPart AnatomicSide BodySide Finding Operation Tense

Eingriffs-Nummer: XXXXXXX
OP-Datum: XXXXXXXXXX 18:55:00
Abteilung: XXXXXXXXXXXXXXX
1. Operateur: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX:
Assistent: XXXXXXXXXXXXXXXXXX; . PJ
Anästhesist: XXXXXXXXXXXX; OAV XXXXXXXXXXXXXXXXX
Anästhesiepflege: XXXXXXXXXXXXX
Instrumentation: XXXXXXXXX
Springer: XXXXXXXXXXXXXX
Lagerung:

Präoperative Diagnose: Luxation einer Hüftgelenktotalendoprothese [Hüft-TEP] (T84.0)
Postoperative Diagnose: Luxation einer Hüftgelenktotalendoprothese [Hüft-TEP] (T84.0)
Therapie: Geschlossene Reposition Endoprothese Hüftgelenk ohne Osteosynthese (8-201.g R)
Freitext Diagnose: Hüft-TEP Luxation rechts
Freitext Therapie: geschlossene Reposition Hüft TEP rechts
Vorgeschichte: Der Verletzte war im Bad gestrauchelt und hatte sich bei diesem Fall bei deutlichem Verdrehen des Beines eine Hüft-TEP-Luxation rechts zugezogen.
Indikation zur Geschlossenen Reposition Hüft-TEP rechts.

---

1. Bericht: Larynxmaskennarkose.
Unter Durchleuchtung erfolgt zunächst der Ausgleich durch Längenverkürzung, dann Eindrehen des Hüftkopfes auf die Pfanne. Unter Durchleuchtungskontrolle wird das Repositionsergebnis überprüft sowie das korrekte Eintreten des Kopfes in die Pfanne.



Endoprothesenregister – Ergebnisse (Auswahl)

cloud4health

Information	Precision	Recall	F-Score
Anatomie	1,0	1,0	1,0
Wechselgrund	0,89	0,89	0,89
Komplikation	0,85	0,9	0,88
Bewegungseinschränkung	0,83	0,84	0,82
Schmerz	0,83	0,81	0,82

cloud4health

Im Projekt wurden 4 konkrete Anwendungsszenarien umgesetzt:

- **Qualitätsmonitoring medizinischer Produkte**
  - Retrospektives Befüllen von Registern am Beispiel des Endoprothesenregisters Deutschland
- **Klinische Leitlinien/Plausibilität von Verordnungen**
  - Zusammenarbeit mit P^3 zur Überprüfung von Verordnungen im Bereich der Psychiatrie
- **Biodatenbanken**
  - Extraktion von Tumorgraduierung (z.B. TNM Kodierungen) aus großen Pathologiedatenbeständen
- **Pharmakovigilanz**
  - Detektion von Nebenwirkungen an verschiedenen Fallbeispielen

Biodatenbanken – Spezifikation

cloud4health

Aufgabendefinition:

- Datenmigration von unstrukturierten textuellen medizinischen Berichten zu hochstrukturierter Information in klinischen Datenbanksystemen

Spezifikation der zu extrahierenden Datenpunkte:

- TNM Klassifikation
  - (T)umor
  - (N)ode
  - (M)etastasis
- Prä- und Suffixe
- Tumorgraduierung
- Residualtumor

Eigenschaften der unstrukturierten Eingabedaten:

- 400 Beispieldokumente wurden ausgewählt, anonymisiert und manuell annotiert
  - 300 Erlangen
  - 100 Rhön Kliniken Berg
- Diese Dokumente wurden als Goldstandard für die Evaluierung herangezogen
- RegEx Komponente wurde mit allen Daten getestet
- Machine Learning Komponente (überwachtes Lernen) wurde mit entsprechenden Untermengen trainiert und evaluiert (10-fache Kreuzvalidierung)

cloud4health

Datenmigration von unstrukturierten textuellen medizinischen Berichten zu hochstrukturierter Information in klinischen Datenbanksystemen

...

– Daten und Datenfreigabe

In Zusammenschau mit E XXXXXXXX und E XXXXXXXX und sowie unter Berücksichtigung der vor Beginn der neoadjuvanten Chemotherapie diagnostizierten Lebermetastase (vergl. E XXXXXXXX) ergibt sich bei klinisch angegebenem Z.n. neoadjuvanter Chemoth

Klassifikation nach TNM 2010:

Stadieneinteilung:	ypT2 ypN1a (2/4) pM1(HEP) L1 V0 Pn0
Tumorgraduierung:	G2 (2 + 2 + 2 nach Elston & Ellis, vgl. E XXXXXXXX)
R-Klassifikation:	R0 (medial: 0,3 cm),
RDCIS0(lateral: 0,3 cm)	
Regressionsgrading nach Sinn:	Grad 1
ICD-O:	8500/3

%

TNM Klassifikation

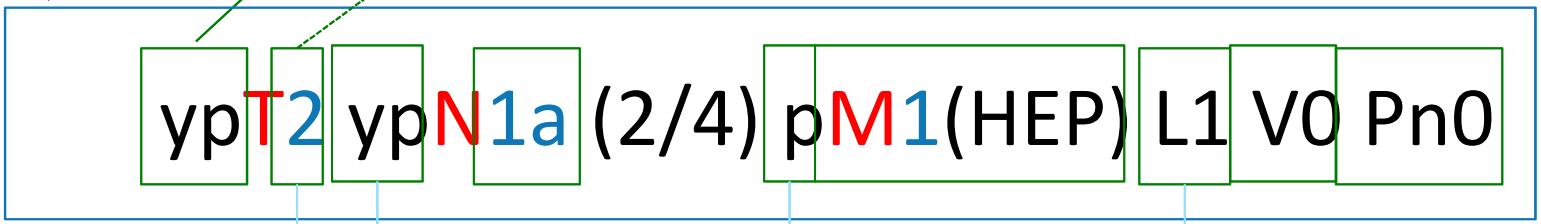
Tumorgraduierung

Residualtumor

cloud4health

1. Schritt: Erkennung von TNM Codes über entsprechende Reguläre Ausdrücke

2. Schritt: Identifikation der relevanten Prä- und Suffixe über Reguläre Ausdrücke



3. Schritt: Strukturierung nach TNM

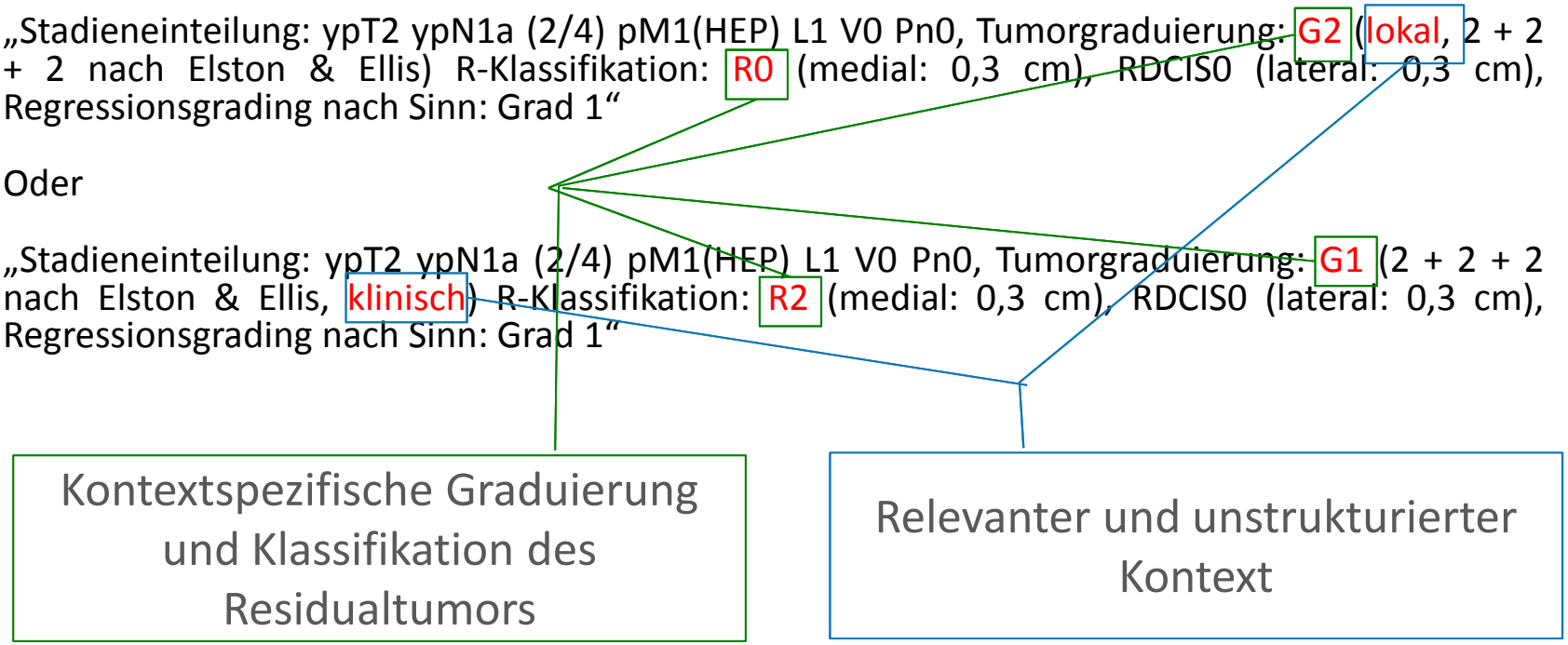
Code	Präfix	Suffix
T	yp	2
N	yp	1a
M	p	1
L	-	1
...	...	...

Biodatenbanken – Technische Realisierung

„Stadieneinteilung: ypT2 ypN1a (2/4) pM1(HEP) L1 V0 Pn0, Tumorgraduierung: G2 (lokal, 2 + 2 + 2 nach Elston & Ellis) R-Klassifikation: R0 (medial: 0,3 cm), RDCIS0 (lateral: 0,3 cm), Regressionsgrading nach Sinn: Grad 1“

Oder

„Stadieneinteilung: ypT2 ypN1a (2/4) pM1(HEP) L1 V0 Pn0, Tumorgraduierung: G1 (2 + 2 + 2 nach Elston & Ellis, klinisch) R-Klassifikation: R2 (medial: 0,3 cm), RDCIS0 (lateral: 0,3 cm), Regressionsgrading nach Sinn: Grad 1“



- Erkennung und Interpretation der Graduierungen und der Klassifikation des Residualtumors sind kontextabhängig (z.B. lokal vs. klinisch), daher sehr schwer über Regeln abbildbar
- Machine Learning System wurde auf den annotierten Daten trainiert
- Kontext wurde über spezielle Features mit in den Lernprozess aufgenommen

Biodatenbanken – Ergebnisse Reguläre Ausdrücke

cloud4health

Datenquelle	Kennzahl	Ergebnisse (F-Score)
300 Dokumente aus dem Uniklinikum Erlangen	Korrekte Prädiktion der T Komponente	0,98
	Korrekte Prädiktion der N Komponente	0,93
	Korrekte Prädiktion der M Komponente	0,98
	Korrekte Prädiktion der Anzahl der Nachbefunde	0,95
100 Dokumente aus RHÖN	Korrekte Prädiktion der T Komponente	0,96
	Korrekte Prädiktion der N Komponente	0,99
	Korrekte Prädiktion der M Komponente	0,96
Gesamtkorpus 400 Dokumente	Gemittelte Gesamtprädiktion TNM	**0,97**

Biodatenbanken – Ergebnisse Maschinelles Lernen

cloud4health

Datenquelle	Kennzahl	Ergebnisse (F-Score)
300 Dokumente aus dem Uniklinikum Erlangen	Korrekte Prädiktion der Graduierung	0,95
	Korrekte Prädiktion der Residualtumorklassifikation	0,93
100 Dokumente aus RHÖN	Korrekte Prädiktion der Graduierung	1,0
	Korrekte Prädiktion der Residualtumorklassifikation	0,96
Gesamtkorpus 400 Dokumente	Gemittelte Gesamtprädiktion der Graduierung	**0,96**
	Gemittelte Gesamtprädiktion der Residualtumorklassifikation	**0,93**

Fazit aller Anwendungsszenarien

cloud4health

Fazit

- Es konnten vollständige funktionelle Workflows für alle Anwendungsszenarien erstellt werden
- Hohes Niveau in der Genauigkeit der erzielten Ergebnisse
- Durch die verschiedenen Anwendungen wurde eine Bandbreite an Werkzeugen mit hoher Fähigkeit zur Adaption entwickelt
- Datenschutzkonformität abhängig vom Anwendungsfall und Standort der Klinik
- Enge Zusammenarbeit mit den Kliniken garantierte Projekterfolg
- Großes öffentliches Interesse sichert die Weiterentwicklung

Vielen Dank für Ihre Aufmerksamkeit!

cloud4health

Kontakt

Dr. Juliane Fluck und Dr. Philipp Senger

Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI)

Department of Bioinformatics

Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Phone: +49-2241-14-2280

{juliane.fluck, philipp.senger}@scai.fraunhofer.de

www.scai.fraunhofer.de