Automatic extraction of breast cancer information from clinical reports

German Research Center for Artificial Intelligence

Claudia Bretschneider (Siemens), Matthias Hammon, Paul Gass (University Hospital Erlangen), Sonja Zillner (Siemens), Daniel Sonntag (DFKI)

Background

The majority of clinical data is only available in unstructured text documents. Thus, their automated usage in data-based clinical application scenarios, like quality assurance and clinical decision support by treatment suggestions, is hindered because it requires high manual annotation efforts. In this work, we introduce a system for the automated processing of clinical reports of mamma carcinoma patients that allows for the automatic extraction and seamless processing of relevant textual features. Its underlying information extraction pipeline employs a rule-based grammar approach that is integrated with semantic technologies to determine the relevant information from the patient record.

Our approach

We tested the system on our use case of mammography. We process an anonymised subset of the overall corpus; a de-identification tool was used. This corpus comprises **8,766 clinical texts reporting on 2,096 patients**, where the types of texts range from pathological texts (n=6,884) to operation reports (n=274) and radiology reports (n=1,608) over a time period of 15 years. **For the evaluation, we selected a high quality subset of 92 patients, for which the text records are complete in the corpus**, so that the final development corpus consisted of 6,932 reports and the evaluation corpus of 1,834 texts.

Results

Using this integrated approach of information extraction, semantic modelling and rule-based decision support, we can extract the textual features with an accuracy of 0.69 for the most complex feature, the HER2 status, and up to 0.90 for the lymph node status. Using this information as input, it is possible to predict therapeutic measures with an accuracy of 0.59. Further error reduction for the IE results is planned to be integrated into the interactive facetted search application based on the IE results.

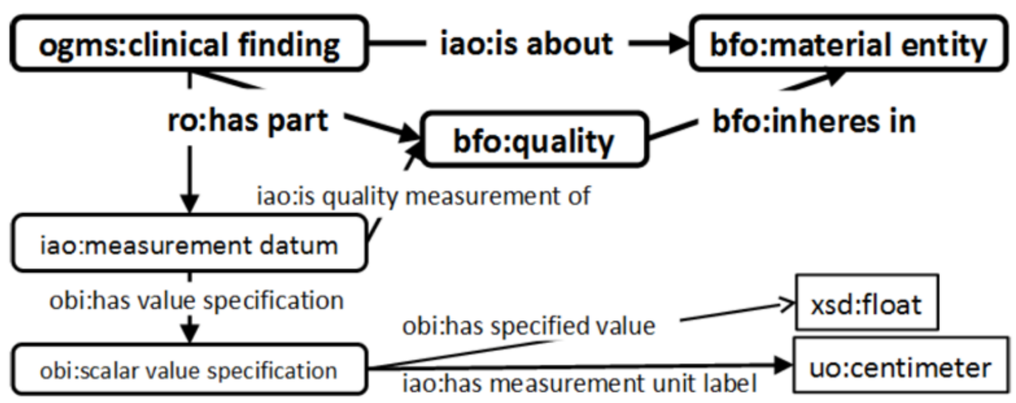

Figure 1. Adapted model for clinical information (MCI)

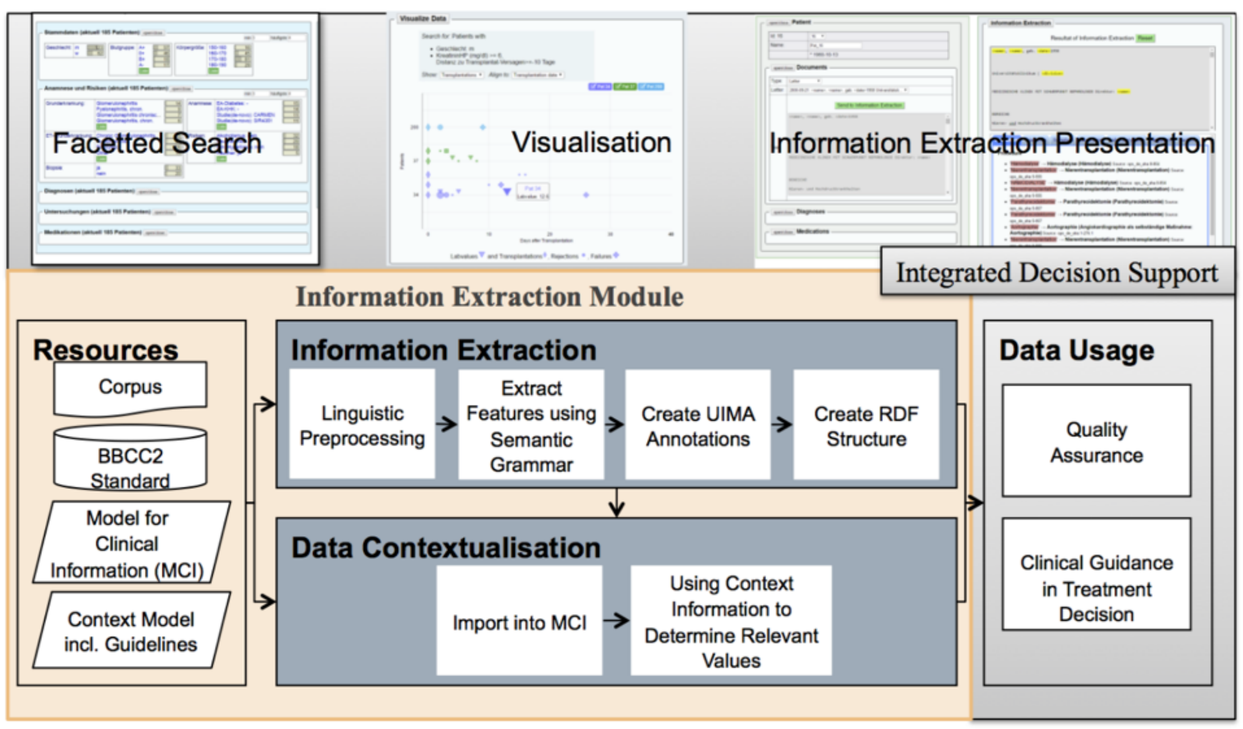

Figure 2. UIMA and SOLR integration for decision support

Table I
FEATURES TO BE EXTRACTED (UNDERLINED) FROM THE CLINICAL REPORTS AND TO AVOID FALSE POSITIVES (CROSSED OUT).

(1) Type of operation conducted (OP) Freitext Therapie: Ablatio re, SNB (blau) ggf Axilla Axilladissektion besprochen, da XXXXXXXX die primaer empfohlene ~~Ablatio mammae~~ strikt ablehnte.
(2) Tumor size (Size) Klinik: Mammakarzinom links, cT1 cN0, 1,7 cm. I.: Maximal 1,2 cm großes mäßig differenziertes invasiv-duktales Mammakarzinom Nach kaudal Abstand ~~0,3cm~~.
(3) Grading of tumor (Grading) Anteile eines mäßig differenzierten, vorwiegend solide wachsenden Mammakarzinoms (linke Mamma, lt. Klinik). Klinik: Mammakarzinom links, IDC G2. Vorgeschichte: XXXXXXXXXXX, eine XXXXXXXXX Patientin (~~G1~~/P1)
(4) Lymph node status (LK) pT2 pN1mi (1/13) L0 V0 Pn0
(5) Hormone receptor state (Hormone) Östrogenhormonrezeptoren: > 80 % (IRS 12/12) Progesteronhormonrezeptoren: negativ
(6) HER2 state (HER2) HER 2-Onkogen-Protein-Expression: 0 (negativ). HER 2-Onkogen-Protein-Expression: Score 1+, somit negativ. Ratio HER2/CEN17 = > 5
(7) Lymphatic spread (Lymph) pT3 pN0 L0 V0 Pn1 G2 R1 Kapsel ~~L1~~, Kapsel 2+3 L2, Kapsel 4+5 L3, Kapsel 6+7 [...] Lymphangiosis carcinomatosa

Project:

KDI – Clinical Data Intelligence

Contact:
DFKI GmbH
Research Department
Intelligent User Interfaces

Dr. Daniel Sonntag
E-mail: Daniel.Sonntag@dfki.de
Website: www.dfki.de/MedicalCPS/?page_id=8