2017 IEEE 30th International Symposium on Computer-Based Medical Systems

Automatic extraction of breast cancer information from clinical reports

Claudia Bretschneider
Corporate Technology, Siemens AG
Otto-Hahn-Ring 6
81739 Muenchen, Germany
claudia.bretschneider.ext@siemens.com

Sonja Zillner
Corporate Technology, Siemens AG
Otto-Hahn-Ring 6
81739 Muenchen, Germany
and School of International Business and Entrepreneurship
Steinbeis University
sonja.zillner@siemens.com

Matthias Hammon, Paul Gass
Department of Radiology
Department of Gynaecology
University Hospital Erlangen, Germany
{*matthias.hammon, paul.gass*}*@uk-erlangen.de*

Daniel Sonntag
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
66111 Saarbruecken, Germany
sonntag@dfki.de

***Abstract*—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. The accuracy of the system, developed with nine thousand clinical documents, reaches accuracy levels of 90% for lymph node status and 69% for the structurally most complex feature, the hormone status.**

***Keywords*-information extraction; natural language processing; medical data analysis; electronic health record (EHR)**

I. Introduction

In order to make use of the wealth of unstructured text data captured in electronic health records (EHR) in clinical data intelligence applications, its semantic annotation is required. The automatic detection of medial information extraction (IE) classes is crucial for integrated decision support. Currently, enormous manual effort is put in the acquisition of the structured data needed for clinical data intelligence or as input for quality assurance processes. In order to automate this step, we introduce a semantically enhanced IE pipeline that allows us to extract relevant information from EHRs of mamma carcinoma patients. We extract six medical IE classes, i.e., type of operation conducted, tumour size, grading of tumour, lymph node status, hormone receptor state, HER2 state, and lymphatic spread. These IE classes, together with the age of treated patient in structured format, are considered to be the main influencing indicators for the therapeutic measure to choose. We describe the end-to-end system.

II. Background and Related Work

Medical text processing has gained relevance and maturity during the last two decades [1]. For extracting adverse drug events from text [2] or automatic symptom extraction from texts on rare diseases [3], for example. However, clinical information extraction from patient records is still underrepresented and underdeveloped in clinical settings. Earlier work includes evaluating context features for medical relation mining on medical abstracts; the identification of semantic relations, such as substance A treats disease B, remains a non-trivial task [4]. Recent work and comparative baseline experiments include temporal information extraction [5]. A special trend becomes apparent, the need for ontology modelling of medical terminology and corresponding information extraction results [6]. Because of enormous annotation costs, mainly unsupervised methods are being used [7]. In industry and in the context of reliable clinical relevance, however, very detailed (and labor-intensive) supervised rule-based approaches on manually labelled corpora represent the state-of-the-art. In particular, the investigation of reports on cancer-related diseases is relevant for our work [8]. Similar to our approach to extract relevant textual features from breast cancer reports, there are other systems that aim to extract TNM classification (as classification of “tumour size”, “node”, “metastasis” data) from clinical texts. Most of them employ supervised rule-based approaches on pathology reports to determine the required information on cancer patients [9]. With the recognition of the required information from the standardised and well-defined TNM coding, they reach accuracy values up to 72%, 78%, and 94% for those three classes, respectively. The idea to base IE

2372-9198/17 $31.00 © 2017 IEEE
DOI 10.1109/CBMS.2017.138

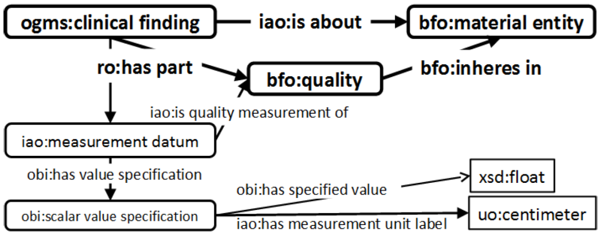

Figure 1. Adapted model for clinical information (MCI)

systems on the specific sublanguage employed by clinicians and introducing a semantic grammar approach has been first introduced by Sager [10] with their system and further applied to radiology reports and biomedical texts [11]. Similar to the semantic model we employ in this work, Coden et al. [12] introduce a model tailored specifically for the representation of cancer-related clinical data.

III. Textual and Semantic Resources

1) Medical Corpus: We use the supervised corpus provided by our clinical partner, the University Hospital Erlangen (Friedrich-Alexander-Universität Erlangen-Nürnberg, FAU). The clinical texts were gathered in the context of the BBCC2 (Bavarian Breast Cancer Cases and Controls) study [13]. The original aim of this case-control study was “the investigation of genetic and non-genetic biomarkers and their influence on breast cancer risk and prognosis” using a cohort of patients with invasive breast cancer. We process an anonymised subset of the overall corpus; a de-identification tool [14] 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.

2) Semantic Model: In addition to the BBCC2, we work with a textual resource that encodes certain features documented in the medical records in a standardised manner. We employ the model for clinical information (MCI) [15]; it is based on RDF and OWL for the descriptions of clinical concepts and their relations. MCI can be applied for integrating and structuring clinical data from heterogeneous sources and incorporates existing medical ontologies for the interpretation of the data. We use the MCI for the representation of the extracted IE classes from the corpus and to model these as clinical findings (figure 1). A clinical finding is modelled by providing information on (1) the general concept (`bfo:material entity`) by linking existing ontology concepts, and (2) its concrete manifestation in form of a `bfo:quality`. For example, the tumour size parameter `T2` is represented by providing the URI to the concept http://purl.org/tnmo/tnmo.owl#MammaryGlandTNM_T in the TNM-O ontology [16] and its respective quality value of `T2`. For this work, the MCI was extended by the following ontology concepts in order to represent the required information about breast cancer patients in a unified way: LOINC [17], TMNO [16] and CPT [18].

3) Clinical Guidelines: We use the German S3 guideline on the treatment of breast cancer [19], which defines how therapeutic measures should be applied based on the clinical picture of the patient, i.e., all the information relating to a disease, disorder, or a patient’s state. For this work, we reduced the relevant therapy suggestions to those covering adjuvant chemotherapy in case of non-metastasised primary breast cancer. These therapy suggestions depend on the variables HER2 status, hormone status, patient age, grading, and number of the invaded lymph nodes. We compiled a set of rules and included them into a context model. These rules suggest a chemotherapy if at least one of the following conditions is met: HER2 status positive, hormone status negative, patient younger than 35, grading level of 3, or more than 4 lymph nodes are invaded.

IV. Implementation

We employ the text corpus processor `Unitex` (http://unitexgramlab.org) for implementing the rule-based grammar approach. We integrate this tool into a `UIMA` (Unstructured Information Management Architecture)-based natural language processing (NLP) pipeline, which consists of four basic steps; the architecture is divided into two major components: the information extraction pipeline (steps 1 and 2) and the data contextualisation component employing the semantic model (steps 3 and 4).

1) Linguistic Preprocessing: The linguistic preprocessing of the corpus includes sentence splitting and tokenisation tasks. The sentence splitting is simplified by the fact that the texts we work with stick to a standardised documentation scheme to imitate a semi-structured document, where most sentences not only end with a period but with a line break. However, an extension of the tokenisation process was required because the reports are populated with hyphenated tokens at every 3rd to 4th line break. Similar to the approach discussed by [20], we re-concatenated hyphenated words at line breaks with a corpus-based approach. In addition, we extended the standard German tokeniser for compounds, which separates the tokens at spaces and special characters, by an algorithm that uses the corpus as baseline to determine whether the re-concatenated multi-term expression is known.

2) Information Extraction Recognition Rules: Based on the discussion with the medical experts and the observations taken from the development corpus, we developed sublanguage recognition rules to recognise the information entities (IE classes) in the texts. This process is simplified

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: XXXXXXXXXXXX, eine XXXXXXXXXX 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

by the fact that some of the features are part of a domain-defined documentation standard, which is documented in a structured manner, such as the TNM classification [9] for the tumour. The variety in documentation of the numerous text features to be extracted are illustrated in table I. **OP** means the type of operation. For its recognition we rely on a term list that has been manually compiled in cooperation with the experts. Furthermore, actual operations have to be differentiated from the options discussed with the patient, but not conducted. **Size**: One indicator for the tumour size is the T parameter in the TNM classification. Second, the tumour size can be extracted. **Grading**: The three-level grading information (G1-3) is either documented as part of the TNM classification or in free-text to indicate the grading level. We explicitly apply rules to avoid false positives, here the grading documented from the gravia/para information on pregnancies and births of the patient that uses similar codes for documentation (e.g., G1/P1). **LK**: The lymph node status is documented as part of the coding system in the TNM classification using the N parameter. **Hormone**: The pathology records use an own standardised documentation scheme for the hormone status. A set of lexical regular grammar rules extracts the documented indicators (as binary positive/negative information, percentage indicator, or with outdated IRS value). **HER2**: Similar to the hormone status, the HER2 status is also documented: either by a numeric coding scheme (0,1, 2+, or 3+), the binary information (positive/negative), or as free text phrase of the possible amplification of the respective gene. The latter requires the consideration of negated concepts, which documents the absence of a positive HER2 gene amplification. **Lymph**: The lymphatic spread is documented as L parameter as part of the TNM classification. Nevertheless, it needs to be differentiated from the (internal) pathological documentation of the lamella and capsule examined, which is indicated as L1. The rules are encoded using the Unitex-based graphs that employ transducers for creating output for each matching sequence. We developed a set of 21 graphs and subgraphs for the recognition and disambiguation of the described textual features. These graphs also include the definitions of the output created after their processing. Each manifestation and its properties are embedded in an XML structure that represents the text annotations.

3) Data Contextualisation: The annotations in RDF format are transformed into the schema and semantics of MCI. As the annotations are already represented in RDF format, we can use SPARQL queries to facilitate the required schema transformation. The resulting representation in the MCI models the extracted information on a patient level. Additionally, it is attributed with provenance information, such as the creation date of the source document, as well as the type of record the information was extracted from (e.g., pathology report or radiology report or operation report). Since the course of the disease is documented over a long period of time, numerous values can be extracted for a feature from the multitude of texts for a single patient. The provenance information of each extracted value in the semantic model allows us to determine the relevant one. Based on the clinical documentation specifics, we define three prioritisation rules that allow us to identify the document from which the extracted information about a specific variable is to be taken: the first rule regards any pathology record as primary source of information, thus more relevant than the radiology or operation records. The second rule prefers the information from most recent reports. Finally, the third rule ignores all reports after the creation date of the BBCC2 standard, which was in June 2014.

4) UIMA / SOLR Integration and Data Usage: In order to integrate the XML structures in this pipeline, we integrate an annotator that parses the XML tags and creates the UIMA annotations for the defined annotation types. Finally, the information extraction component needs to consume the created text annotations in a standardised, machine-readable format. For this purpose we employ the UIMA2LOD component [21] to transform the proprietary UIMA annotations into a semantic format understood by the MCI (or any other model based on Semantic Web technologies). Since the UIMA2LOD component is also build on the UIMA framework, we can seamlessly integrate it into our pipeline and extract necessary RDF resources into the file format understood by the subsequent semantic model. The annotated texts are transferred in XMI format[1] and stored in a local database at DFKI (see figure 2). Important additional components are the SOLR search platform and the facetted search and presentation user interface modules. Solr[2] is an open source enterprise search platform used in many large websites and applications and is one of the most popular enterprise search engines.[3] Solr runs as a standalone full-text search server and uses the Lucene Java search library at its core for full-text indexing and (facetted) search. A software module extracts the relevant medical tags and features and stores these in a database structure similar to the i2b2 star structure (in order to simplify the updates of the target system i2b2[4]). The user interface to search and explore the annotated text database by using facets is built as a web service based on the SOLR extension "solarium" for PHP systems.[5]

V. Evaluation

To illustrate their clinical usage, we describe two use cases: quality assurance and treatment decisions towards integrated decision support. As a first usage scenario, we envision the application of the extracted information in the context of the certification of breast centres. The clinic administration regularly has to compare the conducted therapy with those suggested by the guidelines. In order to pass the quality gate, the suggested therapy has to be used in at least 80 percent of the cases. Currently, the data required for the documentation of the conducted treatments has been collected manually for the BBCC2 standard. Our proposed pipeline can be used to automatically collect textual information necessary to for the certification process, which leads to less manual effort and faster data collection and interactive presentation (facetted search). In a second use case, we plan to use the extracted data in a system supporting the clinical routine. Based on the extracted information from the text and the rule set compiled based on the clinical guidelines, we implemented a first prototype that derives therapy suggestions for a given patient. In order to assist the clinical decision process, this rule-based system is able to suggest and validate the initial guideline based suggestion.

Table II
Results for each BBCC2 Information Extraction class

	OP	Grad	Size	Lymph	HER2	LK	Horm
TP	58	68	30	69	60	82	78
FP	32	24	62	23	24	9	14
FN	2	0	0	0	8	1	0
TN	13	13	13	13	13	13	13
Acc	**0.67**	**0.77**	**0.41**	**0.78**	**0.69**	**0.90**	**0.86**
Sens	**0.96**	**100**	**100**	**100**	**0.88**	**0.98**	**100**
Spec	**0.28**	**0.35**	**0.17**	**0.36**	**0.35**	**0.59**	**0.48**

Table III
Results for Conducted Therapy vs. Suggested Therapy

		Chemotherapy Suggested	
		Yes	**No**
Chemotherapy Conducted	**Yes**	25	12
	No	25	30

We conduct a two-step evaluation, which includes an evaluation of the information extraction results and an evaluation of the therapy suggestions derived from the first set of results.

Information Extraction: The results of the comparison of extraction results with the BBCC2 standard are listed in Table II (see http://www.hutchon.net/EPRval.htm for statistics definitions). We use the substitution error counts against the BBCC2 standard of $\{5, 22, 8, 22, 16, 4, 4\}$ for these classes respectively to approximate true negatives $(81/6 \geq 13)$. This means there is a disregarded feature annotation in the corpus (true negative) which we have annotated by our own method (as FP count). As can be seen, the features documented as part of the TNM classification are easy to extract (high sensitivity/specificity values) with relatively high accuracy. This encoded schema does not require linguistic context to be considered. Those feature values with high linguistic variation are more error-prone. Additionally, some of the values and their disease characteristics need further, detailed consideration in the grammar. For example, for breast cancer, the tumour can have more than one center, hence, affect both breasts, and each of these centres can be graded. Our system currently does not consider this detail, which we consider as possible future enhancement of the system. As the system comes up with high sensitivity for all

[1] http://www.omg.org/spec/XMI/
[2] http://lucene.apache.org/solr/
[3] http://db-engines.com/en/ranking/search+engine
[4] https://www.i2b2.org/about/intro.html
[5] http://www.solarium-project.org/

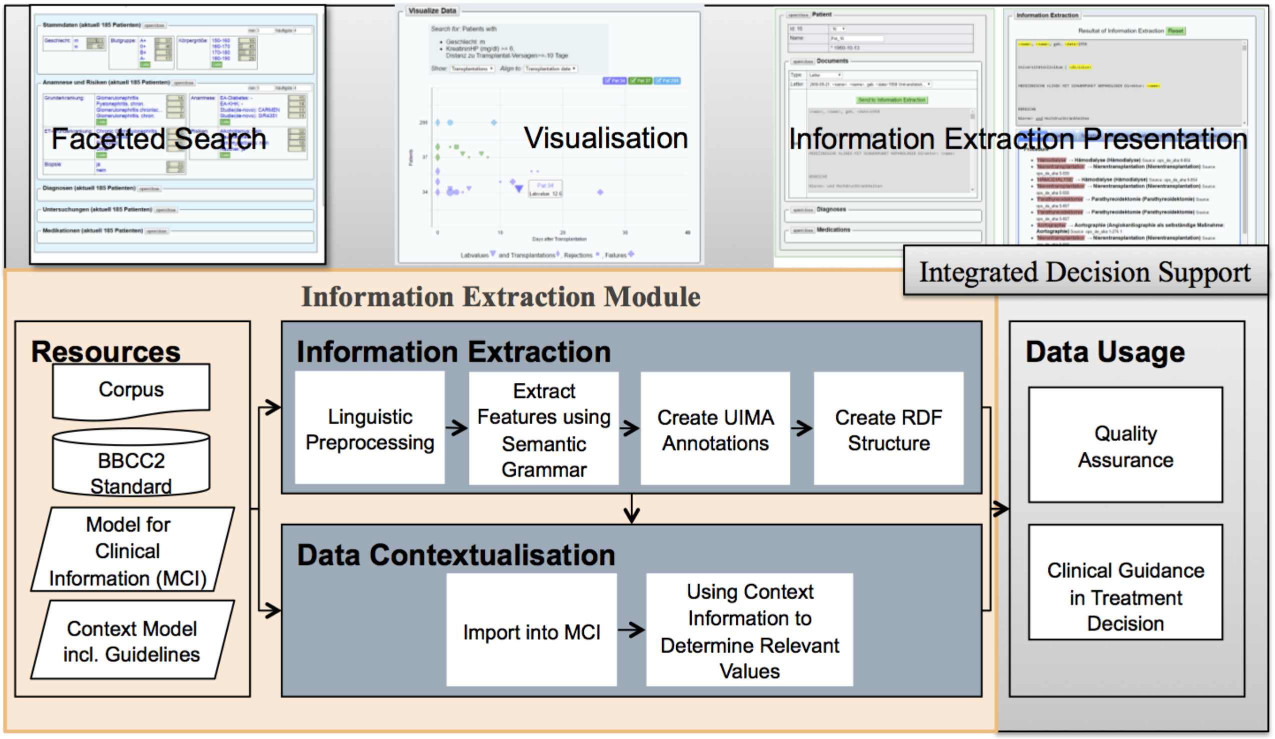
Facetted Search
Visualisation
Information Extraction Presentation
Integrated Decision Support
Information Extraction Module
Resources
Corpus
BBCC2 Standard
Model for Clinical Information (MCI)
Context Model incl. Guidelines
Information Extraction
Linguistic Preprocessing
Extract Features using Semantic Grammar
Create UIMA Annotations
Create RDF Structure
Data Contextualisation
Import into MCI
Using Context Information to Determine Relevant Values
Data Usage
Quality Assurance
Clinical Guidance in Treatment Decision

Figure 2. UIMA and SOLR integration for decision support

those variables, its application might be highly beneficial for the facetted search application. However, further evaluations are needed to support this statement.

Therapy Suggestions: In addition, we evaluated how the actually conducted therapy relates to the proposed therapy using the extracted variables (details are shown in Table III). The accuracy is 59.9%, sensitivity 67.57%, and specificity 54.55%. There are two main factors that influence the error between conducted and suggested therapy. First the error propagation from the IE step, and second, the fact that the suggestion includes the cases where the patient did not follow the suggestion. The patient can still choose if he or she wants to have the chemotherapy performed or not. This also explains why the error rate is clearly higher when a chemotherapy is suggested. Conversely, that a patient insists on a chemotherapy even though it is not suggested, can also happen but is less likely according to the doctors. We plan to further investigate this difference between suggested and conducted therapy and want to include social and demographic details on the patient in this empirical study.

VI. Conclusion and Outlook

We proposed a new system that extracts textual features from reports on mamma carcinoma patients employing a lexical grammar-based approach; second, we derive therapy suggestions from the extracted information. For this purpose a set of seven variables have been defined which need to be extracted from the reports. With the help of a semantic model the multitude of multi-value variables can be handled and the most relevant manifestation can be determined. We tested the system on our use case of mammography. 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. Currently, this system is restricted to the described rules and works on the patient-level. As we discussed in the evaluation, in some of the cases the treatment decision deviates from the actual suggestions according to the guidelines. We plan to evaluate this gap by conducting an empirical study and incorporating social and demographic data. Based on open-source software tools and exchangeable information extraction modules, a suitable information extraction pipeline for the doctor has been created: this type of a knowledge based system provides physicians with a practicable tool for the analysis of medical data and decision support. The use case of mammography in a project of individualised medicine [22] shows a decision support on the integration of existing unstructured information about patients and treatments. Additional case studies will be necessary to characterise the effectiveness. We recognise that there are limitations to our approach, in terms of accuracy for the lexical scope of numerical measurements and values.

References

[1] S. Meystre and P. J. Haug, "Natural Language Processing to Extract Medical Problems from Electronic Clinical Documents: Performance Evaluation," *J. of Biomedical Informatics*, vol. 39, no. 6, pp. 589–599, Dec. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2005.11.004

[2] P. Odom, V. Bangera, T. Khot, D. Page, and S. Natarajan, "Extracting adverse drug events from text using human advice," in *Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings*, 2015, pp. 195–204.

[3] J. Métivier, L. Serrano, T. Charnois, B. Cuissart, and A. Widlöcher, "Automatic symptom extraction from texts to enhance knowledge discovery on rare diseases," in *Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings*, 2015, pp. 249–254.

[4] S. Vintar, L. Todorovski, D. Sonntag, and P. Buitelaar, "Evaluating context features for medical relation mining," in *Proceedings of the ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics*, 2003.

[5] T. Mkrtchyan and D. Sonntag, "Deep parsing at the CLEF2014 IE task," in *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, 2014, pp. 138–146.

[6] D. Sonntag, P. Wennerberg, P. Buitelaar, and S. Zillner, "Pillars of ontology treatment in the medical domain," *J. Cases on Inf. Techn.*, vol. 11, no. 4, pp. 47–73, 2009.

[7] A. Alicante, "Unsupervised entity and relation extraction from clinical records in italian," *Computers in Biology and Medicine*, vol. 72, no. 1, pp. 263–275, 2016.

[8] I. Spasi, J. Livsey, J. A. Keane, and G. Nenadi, "Text mining of cancer-related information: Review of current status and future directions," *International Journal of Medical Informatics*, vol. 83, no. 9, pp. 605 – 623, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386505614001105

[9] A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and S. Colquist, *Journal of the American Medical Informatics Association: JAMIA*, vol. 17, pp. 440–445, 2010.

[10] S. N., *Natural language processing: a computer grammar of english and its applications*. Reading, MA: Addison-Wesley, 1981.

[11] J. Fan and C. Friedman, "Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 805–814, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2011.04.006

[12] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, and P. C. de Groen, "Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model," *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 937 – 949, 2009, biomedical Natural Language Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046408001585

[13] C. Rauh, C. C. Hack, L. Häberle, A. Hein, A. Engel, M. G. Schrauder, P. A. Fasching, A. B. E. S. M. Jud and, C. R. Loehberg, M. Meier-Meitinger, S. Ozan, R. Schulz-Wendtland, M. Uder, A. Hartmann, D. L. Wachter, M. W. Beckmann, and K. Heusinger, "Percent Mammographic Density and Dense Area as Risk Factors for Breast Cancer," *Geburtshilfe Und Frauenheilkunde*, vol. 72, pp. 727–733, 2012.

[14] K. Tomanek, F. Enders, P. Daumke, M. L. Mller, M. Sedlmayr, and H.-U. Prokosch, "Ein System zur De-Identifikation medizinischer Rohdaten," *eGMS*, 2012.

[15] H. Oberkampf, S. Zillner, B. Bauer, and M. Hammon, "An OGMS-based Model for Clinical Information (MCI)," in *Proceedings of the 4th International Conference on Biomedical Ontology, ICBO 2013, Montreal, Canada, July 7-12, 2013.*, ser. CEUR Workshop Proceedings, M. Dumontier, R. Hoehndorf, and C. J. O. Baker, Eds., vol. 1060. CEUR-WS.org, 2013, pp. 97–100. [Online]. Available: http://ceur-ws.org/Vol-1060/icbo2013_submission_56.pdf

[16] M. Boeker, R. Faria, and S. Schulz, "A Proposal for an Ontology for the Tumor-Node-Metastasis Classification of Malignant Tumors: a Study on Breast Tumors," *Ontologies and Data in Life Sciences (ODLS 2014)*, 2014.

[17] A. Srinivasan, N. Kunapareddy, P. Mirhaji, and S. W. Casscells, "Semantic Web Representation of LOINC: an Ontological Perspective," in *AMIA Annual Symposium Proceedings*, vol. 2006. American Medical Informatics Association, 2006, p. 1107.

[18] C. G. Kirschner and R. C. Burkett, *Cpt '95: Physicians' Current Procedural Terminology*. American Medical Association Press, 1994.

[19] R. Kreienberg, W. Jonat, T. Volm, V. Moebus, D. Alt, V. Heilmann, and R. Kreienberg, "S3-Leitlinie Mammakarzinom," *Management des Mammakarzinoms*, pp. 7–11, 2006.

[20] K. Tomanek, J. Wermter, and U. Hahn, "A Reappraisal of Sentence and Token Splitting for Life Sciences Documents," K. A. Kuhn, J. R. Warren, and T.-Y. Leong, Eds., vol. 129. Amsterdam: IOS Press, 2013, pp. 524–528.

[21] C. Bretschneider, H. Oberkampf, and S. Zillner, "UIMA2LOD: Integrating UIMA Text Annotations into the Linked Open Data Cloud," in *Knowledge Engineering and Semantic Web - 6th International Conference, KESW 2015, Moscow, Russia, September 30 - October 2, 2015, Proceedings*, 2015.

[22] D. Sonntag, V. Tresp, S. Zillner, A. Cavallaro, M. Hammon, A. Reis, A. P. Fasching, M. Sedlmayr, T. Ganslandt, H.-U. Prokosch, K. Budde, D. Schmidt, C. Hinrichs, T. Wittenberg, P. Daumke, and G. P. Oppelt, "The clinical data intelligence project," *Informatik-Spektrum Journal*, pp. 1–11, 2015.