A fine-grained corpus annotation schema of German nephrology records

Roland Roller[⋆], **Hans Uszkoreit**[⋆], **Feiyu Xu**[⋆], **Laura Seiffe**[⋆], **Michael Mikhailov**[⋆◇], **Oliver Staeck**[◇], **Klemens Budde**[◇], **Fabian Halleck**[◇] and **Danilo Schmidt**[◇]
[⋆]Language Technology Lab, DFKI, Berlin, Germany
{firstname.surname}@dfki.de
[◇]Charité Universitätsmedizin, Berlin, Germany
{firstname.surname}@charite.de

Abstract

In this work we present a fine-grained annotation schema to detect named entities in German clinical data of chronically ill patients with kidney diseases. The annotation schema is driven by the needs of our clinical partners and the linguistic aspects of German language. In order to generate annotations within a short period, the work also presents a semi-automatic annotation which uses additional sources of knowledge such as UMLS, to pre-annotate concepts in advance. The presented schema will be used to apply novel techniques from natural language processing and machine learning to support doctors treating their patients by improved information access from unstructured German texts.

1 Introduction

Long-term treatment and follow-up of chronically ill patients result in complex medical data and patient records. Although such data is nowadays to a large extent digitalized in various hospital information systems or clinical databases, information is mostly unstructured and difficult to access. Thus, reliable methods to access useful information in clinical data would clearly support physicians. An information extraction system could be applied in the clinical routine to analyze individual patient records for alarming symptoms, historical events, contraindications or side effects. Furthermore it could help to identify subgroups of patients with special characteristics, identify patients for clinical studies or correlating medication and symptoms in historical patient data. Automated information extraction could allow the development of alert systems, which help the clinicians in their daily routine and thus would increase patients safety. However, the first step towards any information extraction is the definition of information of interest, such as diseases, medications or dosing size. This information is then defined within an annotation schema and is used to manually annotate a gold standard corpus to train and evaluate information extraction methods.

Unfortunately, manual annotation is time consuming (Kim et al., 2008) and expensive (Angeli et al., 2014). In particular in the medical domain, expert knowledge is often required which makes the annotation process even more difficult and costly. Therefore existing schemata and corpora could be used in order to save time and effort for the annotation of new data. On the other hand, existing schemata might not cover the information of interest. Furthermore, most of the existing and assessable clinical data sets are in English language. The existing German-language clinical data sets are not freely available. Consequently, we aim to create a new gold standard corpus for German data. This work introduces an annotation schema for reports of the nephrology domain which is based on the requirements of physicians in our project and is motivated by linguistic aspects of German language. The schema takes into account that current German medical dictionaries (which often support named entity recognition) are much smaller than the English ones. Hence, we include annotations on a fine-grained level, in particular in the context of compound words. Moreover, the annotation process includes an automatic pre-annotation step to decrease the duration of manual annotation and to generally ease the annotation process (Batista-Navarro et al., 2015; Kwon et al., 2014).

The paper is structured as follows: The next section presents related work. An overview of relevant data sources is provided in Section 3. The following Section 4 introduces the annotation schema with a

Proceedings of the Clinical Natural Language Processing Workshop,
pages 69–77, Osaka, Japan, December 11-17 2016.

range of different examples. The semi-automatic annotation process is reported in Section 5. The paper finishes with results and future work.

2 Related Work

Information extraction from clinical data has become an important research topic in recent years. With the increasing amount of medical data (such as clinical notes or discharge summaries), the development of reliable text analytics tools could support physicians to better access patient data. However, annotated data sets are required for the development and testing of information extraction methods. Most of the existing annotated clinical data sets are in English language. There are only a few data sets that have been created for non-English languages, such as for Swedish (Skeppstedt et al., 2014), French (Névéol et al., 2015) or Polish (Mykowiecka et al., 2009). For German, only a few sources and clinical corpora exist and will be introduced in the following.

The two most relevant sources for this work are described in Bretschneider et al. (2013) and Toepfer et al. (2015). Bretschneider et al. (2013) focused on the classification of sentences in radiology reports as either pathological and non-pathological based on the given findings. Toepfer et al. (2015) addressed the extraction of fine-grained information from German transthoracic echocardiography reports. The presented terminology involves three main types: objects, attributes and values. Unfortunately, both data sets are not publicly available.

Another very interesting corpus is the FraMed corpus which is described in Wermter and Hahn (2004). The authors present a German-language medical text corpus containing manually supplied sentence boundary, token segmentation and part-of-speech (POS) tags. Due to the fact that the corpus cannot be legally accessed by a third party, Faessler et al. (2014) present an freely available tool for segmentation and POS tagging for German clinical data, based on models trained on the FraMed corpus. Further relevant sources for German clinical data are for instance the German Specialist Lexicon (Weske-Heck et al., 2002) or the German MeSH[1]. A good overview is also provided in the work of Schulz et al. (2013).

3 Utilized Data Sources

This section presents the two data sources used for this work. Firstly, a biomedical knowledge source is presented which is used to automatically pre-annotate data to reduce annotation time. Secondly, the textual data which is used for the annotation is introduced and then analyzed by its (linguistic) characteristics.

3.1 UMLS

The Unified Medical Language System[2] (UMLS) is a large biomedical knowledge base containing millions of medical terms and relations between them. The core component, the Metathesaurus, unifies more than 120 biomedical knowledge vocabularies, such as the Medical Subject Heading (MeSH), the Medical Dictionary for Regulatory Activities (MedDRA) or the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM).

Medical concepts can be described in different ways with different spellings, different abbreviations and also in different languages. UMLS unifies those variations using the Concept Unique Identifier (CUI). Furthermore, UMLS links each CUI against at least one semantic type, such as 'Finding', 'Sign or Symptom' for instance. Most of the concepts are defined in English. However, more than 200,000 German entries can be found[3] in UMLS.

In this work, UMLS will be used for two different purposes. First of all, German concepts of the Metathesaurus are used to pre-annotate data by aligning semantic types to concepts of our annotation schema (see Section 5). Furthermore, unique CUIs should be assigned to annotated concepts in our corpus (normalization against UMLS) at a later stage. Normalization helps to access data more efficiently. Rather than searching for the string '*Niereninsuffizienz*' ('*renal insufficiency*') we can use its

[1] `http://www.dimdi.de/static/en/klassi/mesh_umls/mesh/index.htm`

[2] `https://www.nlm.nih.gov/research/umls/`

[3] UMLS 2016AA, including all German sources

UMLS-CUI *C1565489* which includes different variations in German, such as '*Insuffizienz der Niere*', '*beeinträchtigte Nierenfunktion*' or '*Nierenfunktionsbeeinträchtigung*'.

3.2 Clinical text of the nephrology domain

The annotation task in this paper is conducted within the MACSS[4] (Medical Allround-Care Service Solutions) project, which focuses on improving the safety of patients after kidney transplantation. A key focus of this project is to improve the communication with the patient via a mobile app and to facilitate data exchange and bilateral communication between physicians. Another important goal of this project is the improvement of drug safety by analysis of potentially dangerous drug-drug interactions. In this context the text annotation aims to generate a corpus for the detection of correlated information in historical patient data (e.g. by correlating medication and symptoms). In addition, we want to analyze individual patient records in order to identify alarming symptoms, contraindications or side effects of medications.

At the current project stage German discharge summaries and clinical notes of a hospital's kidney transplant department are annotated. The content of the data set has two peculiar characteristics compared to clinical data of other domains: First, the topic in the documents is related to kidney transplant patients and second, the patients are under a long-term treatment. Both types of documents (discharge summaries and clinical notes) are generally written by medical doctors and have significant differences. The clinical notes are rather short and are written by doctors during or shortly after a visit of a patient. The currently used documents consider only those sections which are addressed to other physicians outside the hospital, such as family doctors or the physician who transferred the patient.

	Discharge Summaries	**Clinical Notes**
#documents available	118	1607
#words (total)	89691	68480
#sentences (total)	16068	11871
avg. words per document (std. deviation)	760.09 (208.62)	42.61 (35.74)

Table 1: Comparison of our Clinical Data Sources

Discharge summaries instead are written during a stay at the hospital. The document is more structured. It contains information about medical history, diagnosis, condition, medication etc. of the patient. Discharge summaries contain much more text compared to clinical notes and are often written by physicians. Furthermore, discharge summaries often contain longer and more well-formed sentences.

Table 1 provides a brief analysis of both document types. Discharge summaries contain a larger average number of words per document compared to the clinical notes[5]. However, the standard deviation of the avgerage word number per document shows that both document types have a large variation in text length. Some clinical notes contain only a few words.

3.2.1 Data Characteristics

The clinical data of this project share the same characteristics as other clinical documents across the world, such as syntactic shortened and reduced semantic complexity. Additionally, the texts contain a large number of Greek- and Latin-rooted words. Often, only keywords are used, together with a lot of abbreviations which are not entirely consistently used over the different texts/authors (Kim et al., 2011). Spelling mistakes and indirect colloquial patient language ('patient reports that legs were tickling') might occur. Besides, texts vary concerning writing style and information density. Due to the nature of German language the documents are also rich in inflection forms and compounds.

Overall, especially linguistic characteristics are of great interest defining our annotation schema: we assume that linguistic resources play a major role in the understanding of the structure of medical data.

[4] `http://macss-projekt.de/`

[5] The information is generated by applying a German tokenizer and a sentence splitter. All non alphabetical tokens are removed.

The German language tends to have a complex sentence- and word structure. While the former varies a lot between different texts and is therefore hard to generalize, the latter is worthy to be considered in more detail.

First of all, characteristics of part-of-speeches (POS) and word formation processes like derivation and composition seem to be important for a deeper understanding. Beside nouns also adjectives and verbs support detailed textual information as presented in Example 1. The example shows, that crucial information can be also expressed by an adjective (1a) or a verb (1b).

(1) a. [Depressive] Episode ('depressive episode')
b. Wir übernahmen den Patienten [intubiert] ('we took over an intubated patient')

In German the POS of a word can be easily changed by derivation processes (Fleischer, 2012) which means, that given concepts are not limited to a specific word category. For this reason it is necessary to not solely rely on the POS distribution and to keep concepts open to various POS. Example 2 illustrates the described situation. 'Delirant' and 'im Delirium', both mean that the patient is in an acute confusional state ('delirium'). While the former is grammatically used as (predicative) adjective, the latter is used as noun.

(2) a. Der Patient war [delirant] ('the patient was delirous')
b. Der Patient war im [Delirium] ('the patient had delirium')

The same situation applies to changes from noun to verb (or vice versa):

(3) a. Es erfolgte die [Sedierung] ('sedation was undertaken')
b. Wir [sedierten] den Patienten ('we sedated the patient')

Compounds like those presented in Example 4 are a very typical phenomenon of the German language and work really productive: They can be built by nearly every POS, yet compounds can be formed by other compounds. This grammar device is frequently used in our corpus.

(4) a. Niereninsuffizienz ('Renal insufficiency')
b. Aortenklappenstenose ('Aortic valve stenosis')

'Niereninsuffizienz' can be paraphrased as 'Insuffizienz der Niere' (literally: 'insufficiency of the kidney'). The given example shows that a fine-grained examination of lexemes help gaining more information than a simple review of the surface does. In Example 4a, a body part in combination with a medical condition might span a new and more specific medical condition, whereas the body part expresses the location of the condition.

4 Annotation Schema

For this work information related to the patient, the disease pattern and the treatment are of interest. In order to answer these superordinate questions, relevant concepts are created that structure the information supporting entities: therefore, focus is on the elements that express medical conditions, their treatments, and further diagnostic procedures. Consequently, the concepts '*Medical_Condition*', '*Treatment*', and '*Diagnostic/Lab_Procedure*' are the most important and the most frequent ones. However, also other concepts, such as '*Body_Part*' or '*Medication*' for instance are important information and considered for the annotation.

Besides further information such as time and location, negations/speculations and some structural data is of interest. Thus, all those elements often serve as specification of the preceding concepts. The development of the concepts took place by manually examining example corpora.

Table 2 presents the list of entities we currently annotate. The relevant entities are grouped into different categories such as *time information* or *person/body*. Furthermore the table provides a brief explanation of each entity. Note, *Biomedical_Chemistry* is currently grouped into the category *therapy*. However, depending on the context the concept can also occur in the category *Person/Body*.

Category	Entity	Explanation
Time Information	Date	Point in time; date
	Temporal_Course	Temporal courses; other temporal information
Person/Body	Person	Mentions of individuals
	Body_Part	Body parts; organs
	Tissue	Body's own tissues
	Body_Fluids	Body's own fluids
	Local_specification	Anatomical descriptions of position and direction
Process	Process	Body's own biological processes
Condition	State_of_Health	Positive, wanted finding; contrary to *Medical_Condition*
	Medical_Condition	Symptom, Diagnosis and observation
	Diagnostic/Lab_Procedure	All types of tests used to diagnose a disease or to assess the patients' state
	Medical_Specification	Closer definition; describing lexemes, often adjectives
	Degree	State of degrees, e.g. degree of a tumor disease
	Type	Closer definition/specification
Therapy	Medical_Device	Medical devices, utilities and material
	Medication	Drugs, medicine
	Biological_Chemistry	Biochemical substances
	Treatment	Therapeutic procedures, treatments
	Measurement	Measurements and the corresponding units
Structure	Structure_Element	Text structuring elements
Truth	Modality_Positive	Explicitly positive lexeme
	Modality_Negation	Negation particle
	Modality_Vagueness	Vagueness expressing elements

Table 2: Relevant concepts

Medical_Condition comprises a wide range of entities. In fact, entities describing findings, diseases and syndromes are all covered by that single concept. Even professionals cannot always distinguish for certain between a disease and a symptom, for instance in case of hypertension. Hypertension can be categorized as a disease or as a symptom, e.g. of a chronic renal insufficiency. By normalizing concepts to UMLS, a distinction can be achieved in later working steps, if required.

As mentioned above, the development of the concepts does not base on the lexeme's grammatical structure (e.g. the POS) but on its semantic value. Thus, also other aspects of the surface structure may vary: the concept *temporal_course* can occur as a word strings (5a), as a scheme for the dosing of medication (5b), or as an prefix within a lexeme (5c).

(5) a. [Seit drei Tagen] ('For three days')
c. Urbason 4 mg [1 - 0 - 0 - 0] ('Urbason 4 mg 1 - 0 - 0 - 0')
b. [Post]extubationem ('after the extubation')

As illustrated in Section 3.2.1, concepts like *Medical_Condition* are not limited to a certain POS. Conversely, there are some exceptions which appear exclusively in adjectival form: *Medical_Specification* and *Local_Specification* occur only in describing, thus in adjectival position. They do not contain the main information (the patient's medical condition and treatments) but serve as further specification. The concept *State_of_Health* is also a special case regarding its POS-structure. Due to its contrary meaning to *Medical_Condition* it might be assumed, that it occurs within the same position and same context. However, *State_of_Health* is actually only used as adjective. Similar to that, the concept *Type* occurs only in one certain position, namely as the first constituent of a compound, see Example 6:

(6) a. [Druck]schmerz ('tenderness and/or pain on palpation')

While most of the concepts base on their semantic value, *Structure_Element* is an exception because its use does not rely on its meaning but on its function. These entities occur as kind of headlines that structure the texts. Additional information throughout the paragraph can be gained by accentuating these elements. Further examples are given in Table 3.

Examples (German; English)	**Annotation**
Sonographie der **Leber** (Ultrasound examination of the **liver**)	Body_part
Ovarialzyste (**ovarial** cyst)	Body_part
inhomogenem Nieren**parenchym** (inhomogeneous renal **parenchyma**)	Tissue
laterales Weichteilrelease (**lateral** soft tissue release)	Local_specification
Sonographie der **linken** Niere (ultrasound examination of the **left** kidney)	Local_specification
Reaktion auf Licht (**Reaction** to light)	Process
physiologische Darm**geräusche** (physiologic bowel **sounds**)	Process
Haut **warm** und **trocken** (skin **warm** and **dry**)	State_of_Health
terminale Nieren**insuffizienz** (terminal renal **insufficiency**)	Medical_Condition
EKG vom 24.01.2000 (**ECG** from 24.01.2000)	Diagnostic/Lab_Procedure
Röntgen Thorax in zwei Ebenen (chest **radiography** in two projections)	Diagnostic/Lab_Procedure
chronische NTx-Glomerulonephritis (**Chronic** glomerulonephritis of the renal allograft)	Medical_Specification
Transplantatversagen nach chronischer NTx-Glomerulonephritis (Renal **allograft** failure after chronic glomerulonephritis of the renal allograft)	Medical_Device
chronische Niereninsuffizienz **Stadium III** (Chronic kidney disease **stage 3**)	Degree
Primärimplantation (**primary** implantation)	Type
Transaminasenanstieg (Elevation of **transaminases**)	Biological_Chemistry
Wir übernahmen den Patienten **sediert**, **intubiert** und **beatmet** (We took over the **sedated**, **intubated** and mechanically **ventilated** patient.)	Treatment
Nephr**ektomie** (Nephr**ectomy**)	Treatment
Tumorausdehnung beträgt **4,5 x 3 x 6 cm**. (Tumor dimensions are **4,5 x 3 x 6 cm**)	Measurement
keine Ödeme (**no** oedemas)	Modality_Negative

Table 3: Annotation Schema - Concept Examples

4.1 Annotation Process

The annotation process aims at a detailed annotation level. This means, that the annotation attempts to detect many information in the documents, but also to consider a fine-granularity. The following example in Figure 1 motivates the granularity. The term 'terminale Niereninsuffizienz' ('terminal renal insufficiency') will be annotated on different levels:

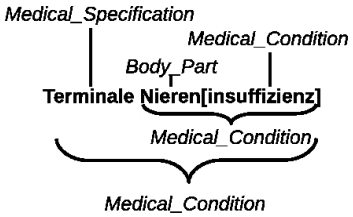

Figure 1: Annotation Granularity

First of all the complete term 'terminale Niereninsuffizienz' will be annotated as *medical_condition*, which is closest to the UMLS entry. Besides also 'Niereninsuffizienz' and 'insuffizienz' will be annotated as *Medical_Condition* in order to achieve a fine-granularity. Furthermore strings such as 'terminale' ('terminal') will be annotated as *Medical_Specification* and 'Niere' ('renal/kidney') as *Body_Part*.

There are different reasons for the detailed annotation level. Firstly, 'terminale Niereninsuffizienz' is the most specific term which includes all other information. Often NER systems target the longest and most specific match. However, UMLS might not cover necessarily all variants. Even more problematic is the fact, that medical terms of interest might be not covered by the German subset of UMLS. A fine granularity might help at a later stage to learn larger constructs (e.g. adjective + compound noun) which are not in the dictionary.

The fine-granularity can be carried to extremes: Some *Local_Specifications* provide special information due to derivation processes, see Example 7:

(7) intrapulmonal ('intrapulmonary')

The first constituent of the lexeme, 'intra-' comes from a finite set of Latin-rooted prefixes which name directional and locational information. 'Pulmonal' is also a Latin-rooted element which can be translated as 'concerning the lungs' ('pulmo' is Latin for 'lung' and the suffix '-al' indicates an adjective). In combination with the prefix 'intra', which means 'inside sth.', the lexeme's meaning is 'inside the lungs' what would be annotated as *Local_Specification*. Since UMLS has no entry for the German lexeme, this fine-grained analysis can provide a deeper understanding.

5 Semi-Automatic Annotation

The annotation is carried out by three students: two linguists, which are familiar with the domain and one medical expert. The medical student is responsible to annotate data and to support the other two students. The annotation task is conducted using the Brat[6] annotator tool. As seen in (Batista-Navarro et al., 2015) or (Kwon et al., 2014), an automatic pre-annotation can help to decrease the duration of manual annotation and to generally ease the annotation process. For this reason an automatic annotation step will be also included into this annotation. In the following the automatic pre-annotation and the preliminary manual annotation will be described.

5.1 Pre-Annotation

To decrease the duration of manual annotation and to generally ease the annotation process, the corpus is pre-annotated automatically[7] beforehand. In this way, falsely tagged elements can be easily corrected and missing annotations included.

The pre-annotation reads in the text documents and applies a tokenization. Currently up to four tokens are considered and matched to the German and English subset of UMLS. Furthermore also substring matches are allowed in order to detect the different components of compound words. The pre-annotation can be divided into three parts: regex, dictionary-lookup and UMLS dictionary lookup. Concepts which are less likely to be found in UMLS are covered by the first two steps. This information usually describes descriptive information of main concepts.

The regex annotation covers the concepts *Measurement*, *Date*, *Temporal_Course*, and *Structure_Element*. Whereas the first three concepts include numbers, in combination with some measurements or month, such as 'mg', 'ml' or 'January', the concept *Structure_Element* detects text spans followed by a colon (':'). These structuring elements usually define the topic of the following text or section and can be used to build up relations to the concepts found in the follow-up text.

The dictionary lookup considers words which are less likely to be found in UMLS as single concepts. Many of the concepts considered here are used to further specify concepts such as *Body_Part* or *Medical_Condition*. In German, many of those concepts (in particular *Medical_Specification* and *Local_Specification*) occur as adjectives or adverbs. In contrast to our approach, UMLS assigns those specifications directly into the surrounding concept, such as 'akute Blutungsanaemie' ('acute haemorrhagic anaemia') or 'papilläres Schilddrüsenkarzinom' ('papillary thyroid carcinoma') and not necessarily as a single concept. This dictionary is manually generated.

word	**substring**
'Empfehlungen' ('suggestions')	'Lunge' ('lung')
'Behandlung' ('Treatment')	'Hand' ('hand')

Table 4: Substring Matching Errors

The UMLS dictionary lookup searches within a window of 4 tokens for German, stemmed German and English words in UMLS. In order to avoid additional errors only capitalized words are considered for English. This pre-annotation component bases on aligning semantic types of UMLS to concepts of our annotation schema. The mapping schema is presented in Table 5. It means, that if a mention can be found in UMLS, its semantic type is examined and if the type matches to one of our concepts, the

[6] http://brat.nlplab.org/

[7] The tool will be made available here: http://macss.dfki.de

Concept name	STY-Name
Person	Human; Patient or Disabled Group
Body_Part	Body Part, Organ, or Organ Component; Body Location or Region
Tissue	Tissue
Body_Fluids	Body Substance
Local_Specification	Spatial Concept
Process	Biologic Function; Physiologic Function; Organism Function; Mental Process; Organ or Tissue Function; Cell Function
State_of_Health	Qualitative Concept
Medical_Condition	Anatomical Abnormality; Congenital Abnormality; Acquired Abnormality; Finding; Sign or Symptom; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Neoplastic Process; Injury or Poisoning
Diagnostic/Lab_Procedure	Laboratory Procedure; Diagnostic Procedure
Medical_Specification	Organism Attribute; Clinical Attribute; Qualitative Concept
Medical_Device	Medical Device
Medication	Clinical Drug; Pharmacologic Substance; Antibiotic
Biological_Chemistry	Biomedical or Dental Material; Biologically Active Substance; Hormone; Enzyme; Vitamin; Immunologic Factor; Receptor; Organic Chemical; Nucleic Acid, Nucleoside, or Nucleotide; Amino Acid, Peptide, or Protein; Inorganic Chemical; Element, Ion, or Isotope; Gene or Genome
Treatment	Therapeutic or Preventive Procedure
Measurement	Quantitative Concepts

Table 5: Mapping Semantic Types to our Annotation Schema

string will be pre-annotated. Additionally the annotation will be extended by its definitions (if defined in UMLS) and its source vocabularies.

The substring matcher also relies on the UMLS dictionary lookup and searches for tokens longer than 3 characters in the German sources. The substring matcher produces various errors as seen in Table 4. However, during the annotation process models and exceptions will be updated to improve the pre-annotation gradually.

Another component of the annotation is an additional synonym dictionary. During the annotation process newly annotated and frequently occurring concepts should be examined in more detail. In this case annotators search for synonyms or English translations in order to find a corresponding entry in UMLS and to extend the German UMLS dictionary.

5.2 Current Annotation Process

At the current stage of the annotation, many files are annotated by at least two different annotators. Annotation differences are then discussed together in a group in order to find the best solution and to ensure a mutual understanding of the annotation task. Using the new annotations the pre-annotation can be successively improved by including new knowledge and addressing frequent errors (such as described in Table 4).

6 Results and Future Work

In this work we presented a fine-grained annotation schema for German clinical text, used for the domain of nephrology. The schema is motivated by linguistic aspects and addresses the needs of clinicians and medical professionals in our project. Furthermore we presented a semi-automatic annotation process in order to ease the annotation procedure. After finishing the concept annotations, the corpus will be normalized against UMLS and extended by relations. The corpus serves as baseline for further information access of patient data in a hospitals' transplant center.

Acknowledgements

This research was supported by the German Federal Ministry of Economics and Energy (BMWi) through the project MACSS (01MD16011F).

References

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning. 2014. Combining Distant and Partial Supervision for Relation Extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics.

Riza Batista-Navarro, Jacob Carter, and Sophia Ananiadou. 2015. Semi-automatic curation of chronic obstructive pulmonary disease phenotypes using argo. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pages 403–408.

Claudia Bretschneider, Sonja Zillner, and Matthias Hammon. 2013. Identifying pathological findings in german radiology reports using a syntacto-semantic parsing approach. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 27–35, Sofia, Bulgaria, August. Association for Computational Linguistics.

Erik Faessler, Johannes Hellrich, and Udo Hahn. 2014. Disclose models, hide the data - how to make use of confidential corpora without seeing sensitive raw data. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Wolfgang Fleischer. 2012. *Wortbildung der deutschen Gegenwartssprache*. Walter de Gruyter.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1).

Youngjun Kim, John Hurdle, and Stphane M Meystre. 2011. Using UMLS lexical resources to disambiguate abbreviations in clinical text. *Annual Symposium proceedings*, 2011:715–722.

Dongseop Kwon, Sun Kim, Soo-Yong Shin, Andrew Chatr-aryamontri, and W. John Wilbur. 2014. Assisting manual literature curation for proteinprotein interactions using bioqrator. *Database*, 2014.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients clinical data. *Journal of Biomedical Informatics*, 42(5):923 – 936. Biomedical Natural Language Processing.

Aurélie Névéol, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. 2015. CLEF ehealth evaluation lab 2015 task 1b: Clinical named entity recognition. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

Stefan Schulz, Josef Ingenerf, Sylvia Thun, and Philipp Daumke. 2013. German-language content in biomedical vocabularies. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*

Maria Skeppstedt, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49:148 – 158.

Martin Toepfer, Hamo Corovic, Georg Fette, Peter Klügl, Stefan Störk, and Frank Puppe. 2015. Fine-grained information extraction from german transthoracic echocardiography reports. *BMC Medical Informatics and Decision Making*, 15(1):1–16.

Joachim Wermter and Udo Hahn. 2004. An annotated german-language medical text corpus. In *Proceedings of the Forth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Gesa Weske-Heck, Albrecht Zaiss, Matthias Zabel, Stefan Schulz, Wolfgang Giere, Michael Schopen, and Rüdiger Klar. 2002. The german specialist lexicon. *Proceedings AMIA Annual Symposium*, pages 884–888.