Named Entities in Medical Case Reports: Corpus and Experiments

Sarah Schulz¹, Jurica Ševa¹, Samuel Rodriguez¹, Malte Ostendorff², Georg Rehm²
¹Ada Health GmbH, Karl-Liebknecht-Str. 1, 10178 Berlin, Germany
{first.lastname}@ada.com
²DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
{first.lastname}@dfki.de

Abstract
We present a new corpus comprising annotations of medical entities in case reports, originating from PubMed Central’s open access library. In the case reports, we annotate cases, conditions, findings, factors and negation modifiers. Moreover, where applicable, we annotate relations between these entities. As such, this is the first corpus of this kind made available to the scientific community in English. It enables the initial investigation of automatic information extraction from case reports through tasks like Named Entity Recognition, Relation Extraction and (sentence/paragraph) relevance detection. Additionally, we present four strong baseline systems for the detection of medical entities made available through the annotated dataset.

Keywords: Named Entity Recognition, Case Reports, Corpus

1. Introduction
The automatic processing of medical texts and documents plays an increasingly important role in the recent development of the digital health area. To enable dedicated Natural Language Processing (NLP) that is highly accurate with respect to medically relevant categories, manually annotated data from this domain is needed. One category of high interest and relevance are medical entities. Only very few annotated corpora in the medical domain exist. Many of them focus on the relation between chemicals and diseases or proteins and diseases, such as the BC5CDR corpus (Li et al., 2016), the Comparative Toxicogenomics Database¹ (Davis et al., 2019), the FSU PRotein GEnecorpus² (Hahn et al., 2010) or the ADE (adverse drug effect) corpus (Gu et al., 2012). The NCBI Disease Corpus (Doğan et al., 2014) contains condition mention annotations along with annotations of symptoms. Several new corpora of annotated case reports were made available recently. Grouin et al. (2019) presented a corpus with medical entity annotations of clinical cases written in French, Ju et al. (2019) presented a corpus focusing on phenotypic information for chronic obstructive pulmonary disease while Smallheiser et al. (2019) presented a corpus focusing on identifying main finding sentences in case reports. The corpus most comparable to ours is the French corpus of clinical case reports by Grouin et al. (2019). They represent a detailed description of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient. We focus on documents freely available through PubMed Central³ (PMC). The presentation of the patient’s case can usually be found in a dedicated section or the abstract. We perform a manual annotation of all mentions of case entities, conditions, findings, factors and modifiers. The scope of our manual annotation is limited to the presentation of a patient’s signs and symptoms. In addition, we annotate the title of the case report.

2. A Corpus of Medical Case Reports with Medical Entity Annotation

2.1. Annotation tasks
Case reports are standardized in the CARE guidelines (Rison et al., 2013). They represent a detailed description of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient. We focus on documents freely available through PubMed Central³ (PMC). The presentation of the patient’s case can usually be found in a dedicated section or the abstract. We perform a manual annotation of all mentions of case entities, conditions, findings, factors and modifiers. The scope of our manual annotation is limited to the presentation of a patient’s signs and symptoms. In addition, we annotate the title of the case report.

2.2. Annotation Guidelines
We annotate the following entities:

- **case entity** marks the mention of a patient. A case report can contain more than one case description. Therefore, all the findings, factors and conditions related to one patient are linked to the respective case entity. Within the text, this entity is often represented by the first mention of the patient and overlaps with the factor annotations which can, e.g., mark sex and age (cf. Figure 1).

- **condition** marks a medical disease such as pneumothorax or dislocation of the shoulder.

- **factor** marks a feature of a patient which might influence the probability for a specific diagnosis. It can be immutable (e.g., sex and age), describe a specific medical history (e.g., diabetes mellitus) or a behaviour (e.g., smoking).

- **finding** marks a sign or symptom a patient shows. This can be visible (e.g., rash), described by a patient

¹ http://ctdbase.org
² https://julielab.de/Resources/FSU_PRGE.html
³ https://www.ncbi.nlm.nih.gov/pmc/
Corpus | Annotated entities | Relationships | # documents
--- | --- | --- | ---
BC5CDR | chemicals (4,409), diseases (5,818) | chemical-disease (3116) | 1,500 PubMed articles
FSU PRotein GEne | protein, protein_family_or_group, protein_complex, protein_variant, protein_enum | - | 3,308 PubMed articles
ADE | drugs (5,063), adverse effect (5,776), dosage (231) | drug-adverse effect (6821), drug-dosage (279) | 3,000 PubMed articles
NCBI Disease Corpus | diseases (6,892) | - | 793 PubMed abstracts
Grouin et al. | 31 targets | - | 717 clinical cases
COPD | 16 COPD phenotype targets | - | 30 full texts
Smallheiser et al. | sentence(s) with main finding(s) | - | 416 full texts
Our work | case (69), condition (347), factor (363), finding (3,248), modifier (336) | - | 53 full texts

Table 1: Summary overview of relevant and comparable corpora.

Figure 1: Annotated entities (WebAnno)

- Seasonal allergic conjunctivitis A 7-year-old boy developed
- Here we present a six-month-old patient with abetalipoproteinemia.
- She had severe pain abdomen with distension and vomiting,
- was a 44-year-old white man with hypothyroidism,

(a) Factor and case annotation
(b) Factor and condition annotation
(c) Partially nested finding annotations

Figure 2: Annotated relations between entities (WebAnno)

- Rapidly progressive polyneuropathy due to dry beriberi
- She ultimately had a normal endoscopy,

(e.g., headache) or measurable (e.g., decreased blood glucose level).

- negation modifier explicitly negate the presence of a certain finding usually setting the case apart from common cases.

We also annotate relations between these entities, where applicable. Since we work on case descriptions, the anchor point of these relations is the case that is described. The following relations are annotated:

- has relations exist between a case entity and factor, finding or condition entities.
- modifies relations exist between negation modifiers and findings.
- causes relations exist between conditions and findings.

Example annotations are shown in Figure 2.

2.3. Annotators

We asked medical doctors experienced in extracting knowledge related to medical entities from texts to annotate the entities described above. Initially, we asked four annotators to test our guidelines on two texts. Subsequently, identified issues were discussed and resolved. Following this pilot annotation phase, we asked two different annotators to annotate two case reports according to our guidelines. The same annotators annotated an overall collection of 53 case reports.

Inter-annotator agreement is calculated based on two case reports. We reach a Cohen’s kappa (Cohen, 1960) of 0.68. Disagreements mainly appear for findings that are rather unspecific such as She no longer eats out with friends which can be seen as a finding referring to “avoidance behaviour”.

2.4. Annotation Tools and Format

The annotation was performed using WebAnno⁴ (Eckart de Castro et al., 2016), a web-based tool for linguistic annotation. The annotators could choose between a pre-annotated version or a blank version of each text. The pre-annotated versions contained suggested entity spans based on string matches from lists of conditions and findings synonym lists. Their quality varied widely throughout the corpus. The blank version was preferred by the annotators. We distribute⁵ the corpus in BioC JSON format⁶. BioC was chosen as it allows us to capture the complexities of the annotations in the biomedical domain. It represented each documents properties ranging from full text, individual passages/sentences along with captured annotations and relationships in an organized manner. BioC is based on character offsets of annotations and allows the stacking of different layers.

2.5. Corpus Overview

The corpus consists of 53 documents, which contain an average number of 156.1 sentences per document, each with 19.55 tokens on average. The corpus comprises 8,275 sentences and 167,739 words in total. However, as mentioned above, only case presentation sections, headings and abstracts are annotated. The numbers of annotated entities are summarized in Table 2.

Findings are the most frequently annotated type of entity. This makes sense given that findings paint a clinical picture of the patient’s condition. The number of tokens per entity ranges from one token for all types to 5 tokens for cases (average length 3.1), nine tokens for conditions (average length 2.0), 16 tokens for factors (average length 2.5), 25

⁵ The corpus can be dowloaded here: https://github.com/adahealth/medical_case_report_corpus.
⁶ http://bioc.sourceforge.net
⁷ Sentence splitting and tokenization are performed using Scispacy (https://allenai.github.io/scispacy/) and its en_core_sci_md model.
The most frequently annotated relation in our corpus is the

<table>
<thead>
<tr>
<th>Type of Relation</th>
<th>Intra-sentential</th>
<th>Inter-sentential</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>case has condition</td>
<td>28 18.1%</td>
<td>127 81.9%</td>
<td>155 4.0%</td>
</tr>
<tr>
<td>case has finding</td>
<td>169 7.2%</td>
<td>2180 92.8%</td>
<td>2349 61.0%</td>
</tr>
<tr>
<td>case has factor</td>
<td>153 52.9%</td>
<td>136 47.1%</td>
<td>289 7.5%</td>
</tr>
<tr>
<td>modifier modifies finding</td>
<td>994 98.5%</td>
<td>15 1.5%</td>
<td>1009 26.2%</td>
</tr>
<tr>
<td>condition causes finding</td>
<td>44 3.6%</td>
<td>3 6.4%</td>
<td>47 1.2%</td>
</tr>
</tbody>
</table>

Table 4: Annotated relations between entities. Relations appear within a sentence (intra-sentential) or across sentences (inter-sentential)

3. Baseline systems for Named Entity Recognition in medical case reports

We evaluate the corpus using Named Entity Recognition (NER), i.e., the task of finding mentions of concepts of interest in unstructured text. We focus on detecting cases, conditions, factors, findings and modifiers in case reports (cf. Section 2.2.). We approach this as a sequence labeling problem. Four systems were developed to offer comparable robust baselines.

The original documents are pre-processed (sentence splitting and tokenization with ScispaCy). We do not perform stop word removal or lower-casing of the tokens. The BIO labeling scheme is used to capture the order of tokens belonging to the same entity type and enable span-level detection of entities. Detection of nested and/or discontinuous entities is not supported. The annotated corpus is randomized and split in five folds using scikit-learn (Buitinck et al., 2013). Each fold has a train, test and dev split with the test split defined as .15% of the train split. This ensures comparability between the presented systems.

3.1. Conditional Random Fields

Conditional Random Fields (CRF) (Lafferty et al., 2001) are a standard approach when dealing with sequential data in the context of sequence labeling. We use a combination of linguistic and semantic features\(^8\), with a context window of size five, to describe each of the tokens and the dependencies between them. Hyper-parameter optimization is performed using randomized search and cross validation. Span-based F1 score is used as the optimization metric.

3.2. BiLSTM-CRF

Prior to the emergence of deep neural language models, BiLSTM-CRF models (Huang et al., 2015) had achieved state-of-the-art results for the task of sequence labeling. We use a BiLSTM-CRF model with both word-level and character-level input. BioWordVec\(^9\) (Chen et al., 2018) pre-trained word embeddings are used in the embedding layer for the input representation. A bidirectional LSTM layer is applied to a multiplication of the two input representations. Finally, a CRF layer is applied to predict the

\(^8\) https://github.com/allenai/scispacy/releases/tag/v0.2.2 (Neumann et al., 2019) and the en_core_sci_md model.

\(^9\) A list of features will be published with the corpus to guarantee reproducibility of the results.

BioWordVec_PubMed_MIMICIII_d200.bin

Table 2: Corpus statistics

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>8,275</td>
<td>827</td>
<td>44</td>
<td>156.1</td>
</tr>
<tr>
<td>Words</td>
<td>167,739</td>
<td>16,309</td>
<td></td>
<td>3164.9</td>
</tr>
<tr>
<td>Annotated sentences</td>
<td>1063</td>
<td>228</td>
<td>1</td>
<td>19.55</td>
</tr>
<tr>
<td>case</td>
<td>69</td>
<td>5</td>
<td>1</td>
<td>3.1</td>
</tr>
<tr>
<td>condition</td>
<td>347</td>
<td>9</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>factor</td>
<td>363</td>
<td>16</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>finding</td>
<td>3,248</td>
<td>25</td>
<td>1</td>
<td>2.6</td>
</tr>
<tr>
<td>modifier</td>
<td>336</td>
<td>18</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>total annotations</td>
<td>4,363</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discontinuous</td>
<td>1,055</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>multi-label</td>
<td>1,535</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discontinuous and multi-label</td>
<td>541</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nested</td>
<td>603</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fully nested</td>
<td>584</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>partially nested</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Examples of long entities per type

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>case</td>
<td>42-year-old poorly prepared mountaineer</td>
</tr>
<tr>
<td>condition</td>
<td>Salter–Harris type II epiphysiolysis at the proximal left humerus</td>
</tr>
<tr>
<td>factor</td>
<td>5 ml of paracetamol was given to child every 4 hours for the past 6 days</td>
</tr>
<tr>
<td>finding</td>
<td>nests and sheets of cells with moderate-to-abundant cytoplasm, eccentrically placed nuclei surrounded by dense pink homogeneous material</td>
</tr>
<tr>
<td>modifier</td>
<td>height and weight were at the 25th - 50th percentile and the 50th - 75th percentile</td>
</tr>
</tbody>
</table>

Figure 3: Discontinuous finding annotation

Entities can appear in a discontinuous way. We model this as a relation between two spans which we call “discontinuous” (cf. Figure 3). Especially findings often appear as discontinuous entities, we found 543 discontinuous finding relations. The numbers for conditions and factors are lower with seven and two, respectively. Entities can also be nested within one another. This happens either when the span of one annotation is completely embedded in the span of another annotation (fully-nested; cf. Figure 1a), or when there is a partial overlapping between the spans of two different entities (partially-nested; cf. Figure 1c). There is a high number of inter-sentential relations in the corpus (cf. Table 4). This can be explained by the fact that the case entity occurs early in each document; furthermore, it is related to finding and factor annotations that are distributed across different sentences.

The most frequently annotated relation in our corpus is the has relation between a case entity and the findings related to that case. This correlates with the high number of finding

<table>
<thead>
<tr>
<th>Type of Relation</th>
<th>Intra-sentential</th>
<th>Inter-sentential</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>case has condition</td>
<td>28 18.1%</td>
<td>127 81.9%</td>
<td>155 4.0%</td>
</tr>
<tr>
<td>case has finding</td>
<td>169 7.2%</td>
<td>2180 92.8%</td>
<td>2349 61.0%</td>
</tr>
<tr>
<td>case has factor</td>
<td>153 52.9%</td>
<td>136 47.1%</td>
<td>289 7.5%</td>
</tr>
<tr>
<td>modifier modifies finding</td>
<td>994 98.5%</td>
<td>15 1.5%</td>
<td>1009 26.2%</td>
</tr>
<tr>
<td>condition causes finding</td>
<td>44 3.6%</td>
<td>3 6.4%</td>
<td>47 1.2%</td>
</tr>
</tbody>
</table>
### 3.3. Multi-Task Learning

Multi-Task Learning (MTL) (Ruder, 2017) has become popular with the progress in deep learning. This model family is characterized by simultaneous optimization of multiple loss functions and transfer of knowledge achieved this way. The knowledge is transferred through the use of one or multiple shared layers. Through finding supporting patterns in related tasks, MTL provides better generalization on unseen cases and the main tasks we are trying to solve.

We rely on the model presented by Bekoulis et al. (2018b) and reuse the implementation provided by the authors. The model jointly trains two objectives supported by the dataset: the main task of NER and a supporting task of Relation Extraction (RE). Two separate models are developed for each of the tasks. The NER task is solved with the help of a BiLSTM-CRF model, similar to the one presented in Section 3.2. The RE task is solved by using a multi-head selection approach, where each token can have one or more relationships to in-sentence tokens. Additionally, this model also leverages the output of the NER branch model (the CRF prediction) to learn label embeddings. Shared layers consist of a concatenation of word and character embeddings followed by two bidirectional LSTM layers. We keep most of the parameters suggested by the authors and change (1) the number of training epochs to 100 to allow the comparison to other deep learning approaches in this work, (2) use label embeddings of size 64, (3) allow gradient clipping and (4) use $d = 0.8$ as the pre-trained word embedding dropout and $d = 0.5$ for all other dropouts. $\eta = 1^{-3}$ is used as the learning rate with the Adam optimizer and tanh activation functions across layers. Although it is possible to use adversarial training (Bekoulis et al., 2018a), we omit from using it. We also omit the publication of results for the task of RE as we consider it to be a supporting task and no other competing approaches have been developed.

### 3.4. BioBERT

Deep neural language models have recently evolved to a successful method for representing text. In particular, Bidirectional Encoder Representations from Transformers (BERT) outperformed previous state-of-the-art methods by a large margin on various NLP tasks (Devlin et al., 2019). For our experiments, we use BioBERT, an adaptation of BERT for the biomedical domain, pre-trained on PubMed abstracts and PMC full-text articles (Lee et al., 2019). The BERT architecture$^{12}$ for deriving text representations uses 12 hidden layers, consisting of 768 units each. For NER, token level BIO-tag probabilities are computed with a single output layer based on the representations from the last layer of BERT. We fine-tune the model on the entity recognition task during four training epochs with batch size $b = 32$, dropout probability $d = 0.1$ and learning rate $\eta = 2^{-5}$. These hyper-parameters are proposed by Devlin et al. (2019) for BERT fine-tuning.

### 3.5. Evaluation

To evaluate the performance of the four systems, we calculate the span-level precision (P), recall (R) and F1-scores (F1) on four distinct baseline NER systems. All scores are computed as average over five-fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>CRF</th>
<th>BiLSTM CRF</th>
<th>MTL</th>
<th>BioBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.59</td>
<td>0.40</td>
<td>0.55</td>
<td>0.51</td>
</tr>
<tr>
<td>R</td>
<td>0.76</td>
<td>0.22</td>
<td>0.38</td>
<td>0.64</td>
</tr>
<tr>
<td>F1</td>
<td>0.66</td>
<td>0.28</td>
<td>0.44</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>case</td>
<td>condition</td>
<td>factor</td>
<td>finding</td>
</tr>
<tr>
<td>P</td>
<td>0.45</td>
<td>0.40</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>R</td>
<td>0.18</td>
<td>0.05</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>F1</td>
<td>0.26</td>
<td>0.09</td>
<td>0.40</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>modifier</td>
<td>micro avg.</td>
<td>macro avg.</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.74</td>
<td>0.52</td>
<td>0.51</td>
<td>0.31</td>
</tr>
<tr>
<td>R</td>
<td>0.32</td>
<td>0.31</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>F1</td>
<td>0.45</td>
<td>0.39</td>
<td>0.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 5: Span-level precision (P), recall (R) and F1-scores (F1) on four distinct baseline NER systems. All scores are computed as average over five-fold cross validation.

$^{11}$https://github.com/bekou/multihead_joint_entity_relation_extraction

$^{12}$We use the PyTorch version by HuggingFace (Wolf et al., 2019).

$^{13}$https://github.com/chakki-works/seqeval
to other common NER corpora, like CoNLL 2003\textsuperscript{14}, even the best baseline system only achieves relatively low scores. This outcome is due to the inherent difficulty of the task (annotators are experienced medical doctors) and the small number of training samples.

4. Conclusion

We present a new corpus, developed to facilitate the processing of case reports. The corpus focuses on five distinct entity types: cases, conditions, factors, findings and modifiers. Where applicable, relationships between entities are also annotated. Additionally, we annotate discontinuous entities with a special relationship type (*discontinuous*). The corpus presented in this paper is the very first of its kind and a valuable addition to the scarce number of corpora available in the field of biomedical NLP. Its complexity, given the discontinuous nature of entities and a high number of nested and multi-label entities, poses new challenges for NLP methods applied for NER and can, hence, be a valuable source for insights into what entities “look like in the wild”. Moreover, it can serve as a playground for new modelling techniques such as the resolution of discontinuous entities as well as multi-task learning given the combination of entities and their relations. We provide an evaluation of four distinct NER systems that will serve as robust baselines for future work but which are, as of yet, unable to solve all the complex challenges this dataset holds. A functional service based on the presented corpus is currently being integrated, as a NER service, in the QURATOR platform (Rehm et al., 2020).

Acknowledgments

The research presented in this article is funded by the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (Unternehmen Region, Wachstumskern, grant no. 03WKDA1A), see http://qurator.ai. We want to thank our medical experts for their help annotating the data set, especially Ashlee Finckh and Sophie Klopfenstein.

5. Bibliographical References


