Named Entities in Medical Case Reports: Corpus and Experiments

Sarah Schulz[1], Jurica Ševa[1], Samuel Rodriguez[1], Malte Ostendorff[2], Georg Rehm[2]
[1]Ada Health GmbH, Karl-Liebknecht-Str. 1, 10178 Berlin, Germany
{first.lastname}@ada.com
[2]DFKI 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 Database[1], or the FSU PRotein GEne corpus[2]. (Grouin et al., 2019) recently presented a corpus with medical entity annotations of clinical cases written in French. The NCBI Disease Corpus (Doğan et al., 2014) contains condition mention annotations along with annotations of symptoms. Compared to this related work, our corpus includes three additional entity types. We annotate *conditions*, *findings* (including medical findings such as blood values), *factors*, and also *modifiers* that indicate the negation of other entities as well as *case entities*, i. e., entities specific to one 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. 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 conditions, findings and factors. 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.

[1] http://ctdbase.org
[2] https://julielab.de/Resources/FSU_PRGE.html

2.2. Annotation Guidelines

We annotate the following entities:

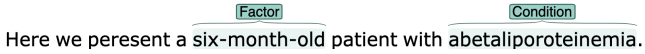
- **case entity** marks the mention of a patient. A case report can contain more than one case description. Therefore, all the findings, factors and conditions related to one patient are linked to the respective case entity. Within the text, this entity is often represented by the first mention of the 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 (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:

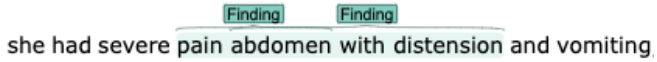
- **has** relations exist between a case entity and factor, finding or condition entities.
- **modifies** relations exist between negation modifiers and findings.
- **causes** relations exist between conditions and findings.

Example annotations are shown in Figure 2.

(a) Factor and case annotation

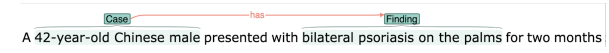


(b) Factor and condition annotation



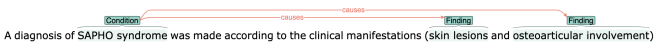
(c) Partially nested finding annotations

Figure 1: Annotated entities (WebAnno)



(a) Annotation of relation *has*

(b) Annotation of relation *modifies*



(c) Annotation of relation *causes*

Figure 2: Annotated relations between entities (WebAnno)

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. Subsequently, we decided to introduce the case entity, marking the patient, which semantically facilitated the annotation of relations of symptoms and factors. 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 WebAnno[3] (Eckart de Castilho 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 distribute[4] the corpus in BioC JSON format[5].

[3] https://webanno.github.io/webanno/, 08/04/2019.

[4] The corpus download URL is subject to acceptance.

[5] http://bioc.sourceforge.net

2.5. Corpus Overview

The corpus consists of 53 documents, which contain an average number of approx. 156 sentences, each with approx. 20 tokens on average. The corpus comprises 8,275 sentences and 167,739 words in total.[6] However, as mentioned above, only case presentation sections, headings and abstracts are annotated. The numbers of annotated entities are summarized in Table 1.

Type	Number	Max	Min	Mean
Documents	53	–	–	–
Sentences	8,275	827	44	156.1
Words	167,739	16,309	1,260	3164.9
Annotated sentences	1063	228	1	19.55
case	69	5	1	3.1
condition	347	9	1	2.0
factor	363	16	1	2.5
finding	3,248	25	1	2.6
modifier	336	18	1	1.4
total annotations	4,363	–	–	–
discontinuous	1,055	–	–	–
multi-label	1,535	–	–	–
discontinuous and multi-label	541	–	–	–
nested	603	–	–	–
fully nested	584	–	–	–
partially nested	19	–	–	–

Table 1: Corpus statistics

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 19 tokens for cases (average length 3.5), nine tokens for conditions (average length 2.0), 16 tokens for factors (average length 2.5), 25 tokens for findings (average length 2.6) and 18 tokens for modifiers (average length 1.4) (cf. Table 1). Examples of rather long entities are given in Table 2.

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

Table 2: Examples of long entities per type

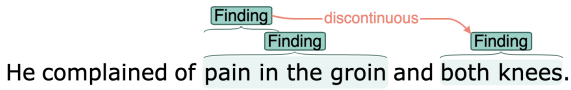


Figure 3: Discontinuous finding annotation

[6] Sentence splitting and tokenization are performed using ScispaCy (https://allenai.github.io/scispacy/) and its *en_core_sci_md* model

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 3). 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.

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

Table 3: Annotated relations between entities. Relations appear within a sentence (intra-sentential) or across sentences (inter-sentential)

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 entities. The relations contained in our corpus are summarized in Table 3.

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, available through PubMed Central, are pre-processed (sentence splitting and tokenization with ScispaCy.[7] We do not perform stop word removal or lowercasing 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). 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) pretrained 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 sequence of labels. Dropout and L1/L2 regularization is used where applicable. He (uniform) initialization (He et al., 2015) is used to initialize the kernels of the individual layers. As the loss metric, CRF-based loss is used, while optimizing the model based on the CRF Viterbi accuracy. Additionally, span-based F1 score is used to serialize the best performing model. We train for a maximum of 100 epochs, or until an early stopping criterion is reached (no change in validation loss value grater than 0.01 for ten consecutive epochs). Furthermore, Adam (Kingma and Ba, 2014) is used as the optimizer. The learning rate is reduced by a factor of 0.3 in case no significant increase of the optimization metric is achieved in three consecutive epochs.

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.[10] 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 none 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

[7] https://github.com/allenai/scispacy/releases/tag/v0.2.2 (Neumann et al., 2019) and the *en_core_sci_md* model.

[8] A list of features will be published with the corpus to guarantee reproducibility of the results.

[9] https://ftp.ncbi.nlm.nih.gov/pub/lu/Suppl/BioSentVec/BioWordVec_PubMed_MIMICIII_d200.bin

[10] https://github.com/bekou/multihead_joint_entity_relation_extraction

	CRF			BiLSTM CRF			MTL			BioBERT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
case	0.59	0.76	**0.66**	0.40	0.22	0.28	0.55	0.38	0.44	0.43	0.64	0.51
condition	0.45	0.18	0.26	0.00	0.00	0.00	0.62	0.62	**0.62**	0.33	0.37	0.34
factor	0.40	0.05	0.09	0.23	0.04	0.06	0.6	0.53	**0.56**	0.17	0.10	0.12
finding	0.50	0.33	0.40	0.39	0.26	0.31	0.62	0.61	**0.61**	0.41	0.53	0.46
modifier	0.74	0.32	0.45	0.60	0.42	0.47	0.66	0.63	**0.65**	0.51	0.52	0.50
micro avg.	0.52	0.31	0.39	0.41	0.23	0.30	0.52	0.44	**0.47**	0.39	0.49	0.43
macro avg.	0.51	0.31	0.38	0.37	0.23	0.28	0.61	0.58	**0.59**	0.40	0.49	0.44

Table 4: 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.

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[11] 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, along with corresponding micro and macro scores. The reported values are shown in Table 4 and are averaged over five folds, utilising the seqeval[12] framework.

With a macro avg. F1-score of 0.59, MTL achieves the best result with a significant margin compared to CRF, BiLSTM-CRF and BERT. This confirms the usefulness of jointly training multiple objectives (minimizing multiple loss functions), and enabling knowledge transfer, especially in a setting with limited data (which is usually the case in the biomedical NLP domain). This result also suggest the usefulness of BioBERT for other biomedical datasets as reported by (Lee et al., 2019). Despite being a rather standard approach, CRF outperforms the more elaborated BiLSTM-CRF, presumably due to data scarcity and class imbalance. We hypothesize that an increase in training data would yield better results for BiLSTM-CRF but not outperform transfer learning approach of MTL (or even BioBERT). In contrast to other common NER corpora, like CoNLL 2003[13], 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.

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.

[11] We use the PyTorch version by HuggingFace (Wolf et al., 2019).

[12] https://github.com/chakki-works/seqeval

[13] https://www.clips.uantwerpen.be/conll2003/ner/

5. Bibliographical References

Bekoulis, G., Deleu, J., Demeester, T., and Develder, C. (2018a). Adversarial training for multi-context joint entity and relation extraction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2830–2836.

Bekoulis, G., Deleu, J., Demeester, T., and Develder, C. (2018b). Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pages 108–122.

Chen, Q., Peng, Y., and Lu, Z. (2018). Biosentvec: creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT, pages 4171–4186, oct.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus. *J. of Biomedical Informatics*, 47(C):1–10, February.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 76–84, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Grouin, C., Grabar, N., Claveau, V., and Hamon, T. (2019). Clinical case reports for NLP. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 273–282, Florence, Italy, August. Association for Computational Linguistics.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, pages 1026–1034, Washington, DC, USA. IEEE Computer Society.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. pages 1–8.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 05.

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 319–327, Florence, Italy, August. Association for Computational Linguistics.

Rison, R. A., Kidd, M. R., and Koch, C. A. (2013). The care (case report) guidelines and the standardization of case reports. *Journal of Medical Case Reports*, 7(1):261, Nov.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. oct.