GGTWEAK: Gene Tagging with Weak Supervision for German Clinical Text

Sandro Steinwand*, Florian Borchert*, Silvia Winkler, and Matthieu-P. Schapranow

HPI Digital Health Center, Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany
sandro.steinwand@student.hpi.uni-potsdam.de, florian.borchert@hpi.de

Abstract. Accurate extraction of biomolecular named entities like genes and proteins from medical documents is an important task for many clinical applications. So far, most gene taggers were developed in the domain of English-language, scientific articles. However, documents from other genres, like clinical practice guidelines, are usually created in the respective language used by clinical practitioners. To our knowledge, no annotated corpora and machine learning models for gene named entity recognition are currently available for the German language.
In this work, we present GGTweak, a publicly available gene tagger for German medical documents based on a large corpus of clinical practice guidelines. Since obtaining sufficient gold-standard annotations of gene mentions for training supervised machine learning models is expensive, our approach relies solely on programmatic, weak supervision for model training. We combine various label sources based on the surface form of gene mentions and gazetteers of known gene names, with only partial individual coverage of the training data. Using a small amount of hand-labelled data for model selection and evaluation, our weakly supervised approach achieves an F_1 score of 76.6 on a held-out test set, an increase of 12.4 percent points over a strongly supervised baseline.
While there is still a performance gap to state-of-the-art gene taggers for the English language, weak supervision is a promising direction for obtaining solid baseline models without the need to conduct time-consuming annotation projects. GGTweak can be readily applied in-domain to derive semantic metadata and enable the development of computer-interpretable clinical guidelines, while the out-of-domain robustness still needs to be investigated.

Keywords: Clinical NLP · Gene Named Entity Recognition · German Language · Computer Interpretable Guidelines.

1 Introduction

Molecular Tumor Boards (MTBs) become increasingly established in cancer care and necessitate time-intensive research for the latest scientific evidence [13].

* These authors equally share first authorship.

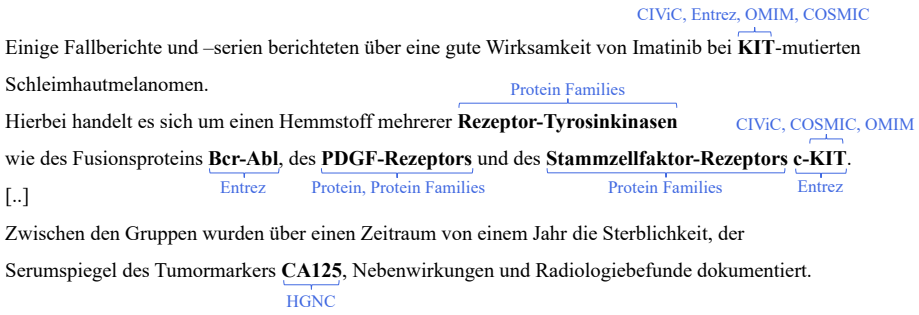

Fig. 1. Examples of gene / protein mentions (**bold**) in German oncology guidelines and labelling functions (blue) matching partially overlapping subsets of these mentions.

Therefore, specialized language technology is needed to extract molecular information from the medical literature and make insights from oncogenetics accessible in a scalable manner. For most downstream processing, Named Entity Recognition (NER) is an essential building block [19]. In this work, we consider the detection of gene and protein mentions, as shown in Fig. 1. We follow common practice in biomedical text mining and treat these entities interchangeably [6].

State-of-the-art approaches for gene NER typically rely on supervised machine learning models, trained on text corpora manually annotated by subject-matter experts. As such strong supervision is expensive to obtain, different sources of external knowledge, heuristics, and other kinds of noisy labels can be exploited in addition. Recently, such weakly supervised methods were successfully applied for clinical NER and can approach the performance of models trained with a comparable amount of strong supervision [9].

In this work, we propose German Gene Tagging with Weak Supervision (GGTweak), the first publicly available NER model for genes and proteins in German medical text. It is based on large amounts of unlabelled text in the German Guideline Program in Oncology NLP Corpus (GGPonc) [3] and requires only a minimal amount of hand-labelled data for model selection and evaluation. We use the skweak framework to implement a range of labelling functions (LFs), and combine their predictions to train a Transformer-based NER model [17]. The contributions of this work are: (1) a dataset with novel gold-standard annotations of gene mentions for a subset of GGPonc, (2) an implementation and detailed analysis of various LFs for finding gene mentions and (3) a freely distributable neural model for gene tagging trained on aggregated weak labels. We make the source code and trained model publicly available [12].

The remainder of this work is structured as follows: In Sect. 3, we share our weak supervision methodology and incorporated data. In Sect. 4, we evaluate our LFs and model performance with respect to a small amount of gold-standard annotations. We discuss our findings in Sect. 5 and conclude our work with an outlook in Sect. 6.

Table 1. A selection of biomedical text corpora annotated on the level of molecular entities as well as the performance of recently published gene taggers evaluated on these corpora. Note: The annotation schemes vary considerably and may include more fine-grained distinctions of subclasses than we use.

Corpus	Lang.	Sent.	NER Model	F_1 Score	Year
CRAFT [1]	EN	21K	HunFlair [25]	0.722	2020
BC2GM [23]	EN	20K	DTranNER [14]	0.845	2020
ProGene [8]	EN	36K	Flair [8]	0.850	2020
JNLPBA [7]	EN	19K	BioELECTRA [20]	0.802	2021
PharmaCoNER [11]	ES	14K	BioBERT v1.1 [24]	0.899	2021

2 Related Work

In Tab. 1, we give an overview of corpora annotated with biomolecular entities and the respective performance of NER taggers. There are a number of English-language corpora based on scientific articles. In addition to such gold-standard corpora, silver-standard annotations can improve the performance of fine-tuned NER taggers when used for transfer learning [10]. Given the large amount of existing annotated gold-standard corpora for the English language, such silver-standard annotations can be obtained by applying existing NER taggers to large, unlabelled corpora. In contrast, non-English corpora with a clinical focus and annotations of biomolecular entities are scarce, with few exceptions, such as Spanish-language clinical case reports in PharmaCoNER [11].

For the German language, there is a general shortage of annotated medical text corpora and none of the few existing ones provides annotations of genes or proteins [27]. We suppose that the genres of clinical texts used so far (often discharge summaries) did not contain any particularly rich molecular information. Moreover, each additional annotation layer complicates annotation and requires specialized domain expertise. Earlier experiments with dictionary-based silver-standard annotations for genes on GGPonc 1.0 resulted in an extremely large number of false positive results [2]. In this work, we aim to alleviate this shortcoming by combining different sources of weak labels instead and train a statistical NER tagger on top of them.

Recently, a number of solutions for integrating multiple label sources as programmatic, weak supervision in structured prediction tasks like NER have been proposed. Fries et al. use the Snorkel framework to integrate labels obtained from a large set of medical terminologies [9,21]. Extensions to the generative label model introduced by Snorkel employ structured probabilistic models, like HMMs, which allow modelling the dependencies of adjacent token labels [22,17]. For this work, we use the skweak framework, as it employs an HMM to model dependencies across labelled tokens. Due to its tight integration with spaCy, the resulting pipeline can be easily shared and integrated into downstream applications [18].

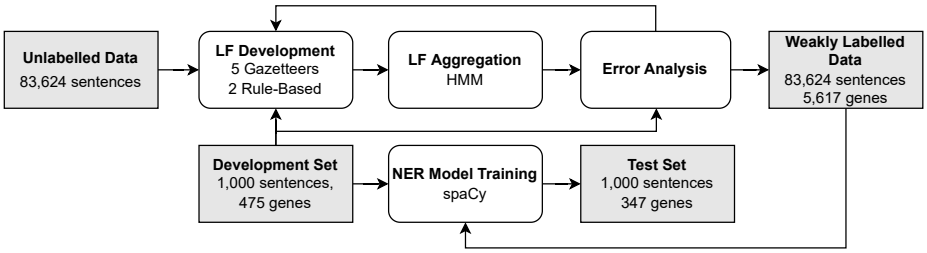

Fig. 2. Overview of GGTweak. Rectangular boxes represent datasets, while round-cornered boxes indicate process steps. Starting from a large unlabelled dataset, we apply seven LFs that cover subsets of gene mentions. Their outputs are aggregated using a Hidden Markov Model (HMM), resulting in a weakly labelled dataset, which is used to train a Transformer-based NER model. A small set of manual annotations are used as development and test data for error analysis and model selection. For comparison, we also train a strongly supervised NER model on the gold standard development set and evaluate it on the test set.

3 Methods

This section describes the used dataset and the weak supervision approach based on LFs and their aggregation, outlined in Fig. 2.

3.1 Dataset and Annotation

As a dataset, we use the freely distributable GGPonc corpus. Originally, the complete corpus contains 1,877K tokens in 10.19K documents. For compatibility with skweak, sentence segmentation and tokenization was carried out again using the spaCy model `de_core_news_md`.

We randomly sampled 2,000 sentences from the subset of documents that contain at least one gene mention according to the silver-standard annotations in GGPonc 1.0 [2]. These sentences were then manually annotated with gene mentions using the INCEpTION tool [15]. Annotation was performed by a single medical student with extensive experience in linguistic annotation. The amount of hand-labelled data was chosen a priori such that a single annotator can annotate it in around one work week. In total, 822 mentions of genes and proteins were annotated. The manually annotated documents are used as development and test sets (1,000 sentences each). The remaining 83,624 sentences are labelled automatically with weak supervision and used as training data.

3.2 Labelling Function Development

We apply the following LFs based on external knowledge bases, naming conventions for gene names, and other heuristics to programmatically annotate the unlabelled part of the corpus with automatically induced labels. For implementation details, please refer to the interactive notebooks in our source code repository [12].

Gazetteer-based LFs The following LFs are based on *gazetteers*, i.e., they match tokens to entries in a list of known gene and protein names.

- **CIViC:** Gazetteer (case-sensitive) based on canonical gene names in the Clinical Interpretation of Variants in Cancer database, a community knowledge base of cancer genes and variants (*Examples: SMO, VEGF, TP53*).
- **Entrez:** Gazetteer (case-sensitive) based on the aliases of all genes in CIViC that occur in Entrez Gene, the gene-specific database of the National Center for Biotechnology Information [5]. Since many gene aliases can also occur as common German terms in other contexts (e.g., the pronoun "er"), we further filter by the part-of-speech tag of matched tokens (NOUN, PROPN, or X) (*Examples: p16, B-Raf, HER-2*).
- **OMIM:** Gazetteer (case-insensitive) based on the Online Mendelian Inheritance in Man database, a comprehensive catalogue of human genes and genetic disorders (*Examples: PALB2, TNF, BRCA1*).
- **COSMIC:** Gazetteer (case-sensitive) based on the Catalogue of Somatic Mutations in Cancer database, a knowledge base of somatic mutations and additional information associated with cancer in humans (*Examples: BTK, IGHV, BRAF*).
- **Proteins:** Custom gazetteer (case-sensitive) sourced from the German Wikipedia overview page of proteins, manually refined by exploration of the unlabelled training part of the dataset [26] (*Examples: PD-L1, Cyclooxygenase, Uridin-5'-Diphospho-Glucuronosyltransferase*).

Rule-based LFs Another type of LF is based on heuristics that take the surface forms of tokens into account, e.g., a particular composition of uppercase letters and numbers, as well as specific prefixes and suffixes.

- **HGNC:** Heuristic derived from the HUGO Gene Nomenclature Committee naming conventions for genes, using regular expression. As matching short gene names based on this convention would lead to many false positives, we instead rely on a case-insensitive lookup in CIViC for these genes (*Examples: CA125, CYP19, mTORC1*).
- **Protein Families:** Heuristic based on common suffixes describing groups of proteins, e.g., "-rezeptor", "-kinase" or the "-RAS" family (*Examples: Rezeptor-Tyrosinkinasen, MAP-Kinase, k-ras*).

3.3 Labelling Function Aggregation

The LFs were designed such that they cover specific subsets of gene mentions in our corpus (as shown in Fig. 1). Therefore, the partial and potentially conflicting outputs of these LFs are aggregated using the HMM label model from SKWEAK, which emits a single label per token, accounting for correlations and conflicts among the LFs. We fit the HMM on the LF outputs on the training set. The HMM predictions on the 83.6K sentences of the training set result in more than 5,617 automatic annotations of gene mentions.

3.4 Named Entity Recognition Models

We can use the trained HMM to predict labels for unseen instances, which we do for comparison on the development and test set. However, in order to obtain a model that can potentially generalize beyond our LFs, we train another Transformer-based NER model with the spaCy framework on top of the HMM output. To this end, we use the aggregated, weakly labelled data for model training and the gold-standard development set for model selection. The NER model's encoder is initialized from the BERT checkpoint `bert-base-german-cased`, which was pre-trained on German general domain and legal text. We use an initial learning rate of 10^{-5}, with 250 warmup steps and linear learning rate decay. For all other hyperparameters, we use the default values provided by spaCy. The model was trained for 20,000 optimization steps, which takes about 2 hours on a single Nvidia A40 GPU. As the final model, we choose the checkpoint that achieves the maximal F_1 score on the development set.

For comparison with the traditional setting of building NER taggers, we train another model with the same architecture and hyperparameters using the development set as training data, i.e., with just the small amount of available strongly supervised data. The final evaluation of both models is performed on our initially defined test set.

4 Results

The results of the incorporated LFs, the HMM and NER models are presented in Tab. 2. Since we do not have access to ground truth labels on the training set, we estimate the contributions made by each LFs through coverage and overlap. For the development and test set, we can compare all LFs and aggregated models to gold-standard labels.

4.1 Labelling Function Analysis

All LFs achieve high levels of precision and a coverage of up to 40.6% of the targetted labels. The rule-based LFs show small overlap (38.2% and 36.9%), i.e., more than 60% of the mentions they label are unique to these LFs. While the coverage of the suffix-based LF for protein families is low, it has a non-negligible recall on the development set (7.6%), that, combined with the uniqueness of its labels, has a positive impact on the final model.

Considering synonyms from Entrez Gene drastically improves coverage on the training set compared to CIViC, at the expense of a small decrease in precision. Likewise, OMIM as the biggest database has high coverage and only 50.0% overlap with other LFs. In contrast, CIViC and COSMIC both share high overlap but rather low coverage. After aggregation, the combined labels from the HMM result in a slightly lower precision compared to the individual LFs, but provide a better recall and F_1 score.

Table 2. Performance metrics of each LF and derived statistical models (coverage = number of tokens labelled by one LF divided by the number of tokens labelled by all LFs, overlap = number of tokens labelled by one LF that are also labelled by any other LF divided by the total number of tokens labelled by this LF). The strongly supervised model was trained on the development set and is therefore only evaluated on the test set.

	Training		Development			Test		
	Coverage	**Overlap**	**Pr.**	**Rec.**	**F_1**	**Pr.**	**Rec.**	**F_1**
Gazetteers								
CIViC	.210	.980	.944	.465	.624	.841	.473	.606
Entrez	.406	.686	.902	.503	.646	.890	.608	.722
OMIM	.344	.500	.926	.524	.670	.818	.493	.616
COSMIC	.223	.930	.928	.436	.594	.854	.473	.608
Proteins	.137	.699	.934	.120	.212	**.975**	.112	.200
Rule-based								
HGNC	.365	.382	.833	.305	.446	.836	.280	.420
Protein Families	.018	.369	**1.000**	.076	.142	.250	.012	.022
HMM	-	-	.841	.680	.752	.789	.689	.736
GGTWEAK	-	-	.855	**.720**	**.782**	.819	.718	**.766**
Strong Supervision	-	-	-	-	-	.558	**.758**	.642

4.2 Evaluation Against Gold-Standard Annotations

The final GGTweak NER model achieves an F1 score of 78.2% on the development set and 76.6% on the held-out test set. Moreover, GGTweak performs 12.4 percent points better than the model trained with strong supervision in terms of F_1 score. While the strongly supervised model has slightly higher recall (+4 pp.), GGTweak shows dramatically higher precision (+26.1 pp.). GGTweak also outperforms the HMM consistently by a margin of 3 pp., highlighting the added value of transfer learning through pre-trained Transformer weights.

5 Discussion and Limitations

The foundation of our work is GGPOnc, a corpus of German oncology guidelines. While extensive in volume, it is imbalanced regarding the presence of molecular entities, i.e., most sentences do not contain mentions of genes or proteins. We note that the HMM implementation provided by skweak is particularly sensitive to false positives. For these reasons, it is challenging to develop high-precision LFs while maintaining high coverage. Although we have not performed exhaustive ablation experiments, we notice that additional LFs increase the recall of the final model, usually at the expense of decreased precision.

Interestingly, the performance of the final model drops only slightly when evaluating it on the test set in comparison to the development set, although the latter was used during LF development. This indicates a certain generalizability of the model beyond the scope of the LFs. However, we note that the considered text genre provides only a partial representation of the different notations for

gene names that may occur in clinical documents. Both aspects impact the generalization performance of our model and need to be further investigated.

There still remains a performance gap to state-of-the-art gene taggers for English biomedical literature, which often achieve F_1 scores significantly larger than 80% (see Tab. 1). However, we have to bear in mind that the research community has had access to annotated English-language corpora for a much longer time. Furthermore, the underlying problem might be intrinsically harder for the German language due to its grammatical intricacies, such as the prevalence of compound nouns. We rely on several upstream components in spaCy for basic linguistic tasks, such as tokenization and POS tagging. Although these general-domain solutions appear to work reasonably well on GGPOnc, errors introduced by them might influence the performance of our LFs. Lastly, we have not performed any optimization of hyperparameters of both the HMM label model and the Transformer-based NER training, which would likely have a positive impact on model performance.

6 Conclusion and Future Work

In this work, we presented a novel approach for gene tagging in German medical text. With an F_1 score of 76.6 %, we could demonstrate the viability of weak supervision for this task with substantially decreased demand for labels from human experts. Importantly, GGTweak outperforms a model that was trained on the same amount of gold-standard labels that we used for model selection only.

As future work, we plan to add more diversely targetting LFs and explore other Transformer checkpoints, e.g., domain-specialized models for the German [16,4] or other languages, as shown by Sun et al. [24]. An important downstream task is the normalization of gene mentions to identifiers in knowledge bases, such as Entrez Gene [5]. We expect that this will be challenging, as German terms relating to genes and more generally to groups of genes might not have easily identifiable aliases in such knowledge bases.

We believe that more annotated language resources in conjunction with weak supervision can support the development of high-quality gene taggers for clinical documents. Our findings should be readily applicable to other languages, as clinical guidelines are a widely available text genre and most of our LFs do not rely on language-specific resources.

Acknowledgements

Parts of this work were generously supported by a grant of the German Federal Ministry of Research and Education (01ZZ1802H).

References

1. Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A., Cohen, K.B., Verspoor, K., Blake, J.A., Hunter, L.E.: Concept annotation in the CRAFT corpus. BMC bioinformatics **13**(1), 1–20 (2012)
2. Borchert, F., Lohr, C., Modersohn, L., Langer, T., Follmann, M., Sachs, J.P., Hahn, U., Schapranow, M.P.: GGPONC: A corpus of German medical text with rich metadata based on clinical practice guidelines. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. pp. 38–48 (2020)
3. Borchert, F., Lohr, C., Modersohn, L., Witt, J., Langer, T., Follmann, M., Gietzelt, M., Arnrich, B., Hahn, U., Schapranow, M.P.: GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline ner taggers. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 3650–3660 (2022)
4. Bressem, K.K., Papaioannou, J.M., Grundmann, P., Borchert, F., Adams, L.C., Liu, L., Busch, F., Xu, L., Loyen, J.P., Niehues, S.M., Augustin, M., Grosser, L., Makowski, M.R., Aerts, H.J., Löser, A.: MEDBERT.de: A comprehensive German BERT model for the medical domain. arXiv (2023). https://doi.org/10.48550/ARXIV.2303.08179
5. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R., Murphy, T.D.: Gene: a gene-centered information resource at NCBI. Nucleic Acids Research **43**(D1), D36–D42 (2015)
6. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings in Bioinformatics **6**(1), 57–71 (03 2005)
7. Collier, N., Ohta, T., Tsuruoka, Y., Tateisi, Y., Kim, J.D.: Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). pp. 73–78. Geneva, Switzerland (2004)
8. Faessler, E., Modersohn, L., Lohr, C., Hahn, U.: ProGene - a large-scale, high-quality protein-gene annotated benchmark corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4585–4596 (2020)
9. Fries, J.A., Steinberg, E., Khattar, S., Fleming, S.L., Posada, J., Callahan, A., Shah, N.H.: Ontology-driven weak supervision for clinical entity classification in electronic health records. Nature communications **12**(1), 1–11 (2021)
10. Giorgi, J.M., Bader, G.D.: Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics **34**(23), 4087–4094 (2018)
11. Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., Krallinger, M.: PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks. pp. 1–10. Association for Computational Linguistics, Hong Kong, China (2019)
12. Hasso Plattner Institute's Digital Health Center on GitHub: GGTWEAK source code repository. `https://github.com/hpi-dhc/ggponc_molecular` [retrieved: Mar 17, 2023] (2023)
13. Henkenjohann, R., Bergner, B., Borchert, F., Bougatf, N., Hund, H., Eils, R., Schapranow, M.P.: An engineering approach towards multi-site virtual molecular tumor board software. In: International Conference on ICT for Health, Accessibility and Wellbeing. pp. 156–170. Springer (2021)

14. Hong, S., Lee, J.G.: DTranNER: Biomedical named entity recognition with deep learning-based label-label transition model. BMC bioinformatics **21**(1), 1–11 (2020)
15. Klie, J.C., Bugert, M., Boullosa, B., Eckart de Castilho, R., Gurevych, I.: The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9 (2018)
16. Lentzen, M., Madan, S., Lage-Rupprecht, V., Kühnel, L., Fluck, J., Jacobs, M., Mittermaier, M., Witzenrath, M., Brunecker, P., Hofmann-Apitius, M., et al.: Critical assessment of transformer-based ai models for German clinical notes. JAMIA open **5**(4), ooac087 (2022)
17. Lison, P., Barnes, J., Hubin, A.: skweak: Weak supervision made easy for NLP. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. pp. 337–346. Association for Computational Linguistics, Online (2021)
18. Montani, I., Honnibal, M., Landeghem, S.V., Boyd, A., Peters, H., McCann, P.O., Geovedi, J., O'Regan, J., Samsonov, M., Altinok, D., Orosz, G., de Kok, D., Kristiansen, S.L., Miranda, L., Bot, E., Roman, Baumgartner, P., Fiedler, L., Hudson, R., Kannan, M., Edward, Howard, G., Phatthiyaphaibun, W., Tamura, Y., Bozek, S., murat, Daniels, R., Flusskind: explosion/spaCy: v3.4.1: Fix compatibility with CuPy v9.x (Jul 2022). https://doi.org/10.5281/zenodo.6907665
19. Perera, N., Dehmer, M., Emmert-Streib, F.: Named entity recognition and relation detection for biomedical information extraction. Frontiers in cell and developmental biology p. 673 (2020)
20. Raj Kanakarajan, K., Kundumani, B., Sankarasubbu, M.: BioELECTRA: pretrained biomedical text encoder using discriminators. In: Proceedings of the 20th Workshop on Biomedical Language Processing. pp. 143–154 (2021)
21. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid training data creation with weak supervision. In: Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases. vol. 11, p. 269 (2017)
22. Safranchik, E., Luo, S., Bach, S.: Weakly supervised sequence tagging from noisy rules. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5570–5578 (2020)
23. Smith, L., Tanabe, L.K., nee Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., et al.: Overview of BioCreative II gene mention recognition. Genome biology **9**(2), 1–19 (2008)
24. Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., Wang, J.: Deep learning with language models improves named entity recognition for PharmaCoNER. BMC bioinformatics **22**(1), 1–16 (2021)
25. Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., Akbik, A.: HunFlair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics **37**(17), 2792–2794 (2021)
26. Wikipedia: Kategorie:Protein. `https://de.wikipedia.org/wiki/Kategorie:Protein` [retrieved: Mar 17, 2023] (2023)
27. Zesch, T., Bewersdorff, J.: German medical natural language processing–a data-centric survey. In: Applications in Medicine and Manufacturing. pp. 137–142 (2022)