Cross-lingual German Biomedical Information Extraction: from Zero-shot to Human-in-the-Loop

Siting Liang[*] and **Mareike Hartmann**[*,1] and **Daniel Sonntag**[*,2]
[*]German Research Center for Artificial Intelligence, Germany
[1]University of Oldenburg, Germany
[2]Saarland University, Germany
siting.liang|mareike.hartmann|daniel.sonntag@dfki.de

Abstract

This paper presents our project proposal for extracting biomedical information from German clinical narratives with limited amounts of annotations. We first describe the applied strategies in transfer learning and active learning for solving our problem. After that, we discuss the design of the user interface for both supplying model inspection and obtaining user annotations in the interactive environment.

1 Introduction

Medical information extraction from the large volume of unstructured medical records has the potential to facilitate clinical research and enhance personalized clinical care. Especially the narrative notes, such as radiology reports, discharge summaries and clinical notes provide a more detailed and personalized history and assessments, offering a better context for clinical decision making (Chen et al., 2015a; Spasic et al., 2020). Name Entity Recognition (NER) task from Natural Language Processing (NLP) studies, have attempted to accurately and automatically extract medical terms from clinical narratives (Sonntag et al., 2016; Sonntag and Profitlich, 2019; Miotto et al., 2018; Lerner et al., 2020; Wei et al., 2020; Kim and Meystre, 2020) using annotated clinical text corpora (Johnson et al., 2016; Henry et al., 2019; Miller et al., 2019; Lee et al., 2020b; Alsentzer et al., 2019).

The large data collection benefits the research community in developing AI applications in processing medical documents in English (Spasic et al., 2020). However, there are several limitations in improving information extraction from medical records with machine learning methods in other languages, like German in our case: few German annotated datasets are publicly available, and research on non-English medical documents is scarce (Starlinger et al., 2017; Kittner et al., 2021). In most cases, domain experts have higher priority commitments and no capacity to annotate large numbers of training examples for use in machine learning applications (Yimam et al., 2015). Our proposed project for extracting medical terms from German clinical narratives with little annotated training data addresses this problem. Two of the most widely studied approaches to this challenge are transfer learning and active learning. In transfer learning, models transfer knowledge learned from data-rich languages or tasks to languages or tasks with less or no annotated data (Wang et al., 2019; Lauscher et al., 2020; Xie et al., 2018a; Yuan et al., 2019; Pires et al., 2019; Xie et al., 2018b; Plank, 2019). Active learning is an approach to maximize the utility of annotations while minimizing the annotation effort on the unlabeled target data (Chen et al., 2015a; Miller et al., 2019; Liu et al., 2020, 2022; Chaudhary et al., 2019; Shelmanov et al., 2019; Zhang et al., 2020; Lauscher et al., 2020). We train a German biomedical NER model building on these two approaches, addressing the following research questions: a) How to transfer knowledge from annotated English clinical narratives corpora to the German NER model? b) In active learning, 1) What is the minimum amount of annotated samples needed for retraining the model? 2) How to evaluate the effectiveness of the query strategies in real-time training (which human (annotator) factors do we have to consider in addition to model performance)?

2 Approach

We frame our research problem as NER task for German text in the biomedical domain, and combine transfer and active learning strategies in order to reduce the need for annotated data. Our proposed framework, which is shown in figure 1, is similar to the work of Chaudhary et al. (2019). First, a base NER model is pre-trained with English source data using transfer learning strategies (see section 3). Second, we fine-tune the model continuously with annotations in the target language using active

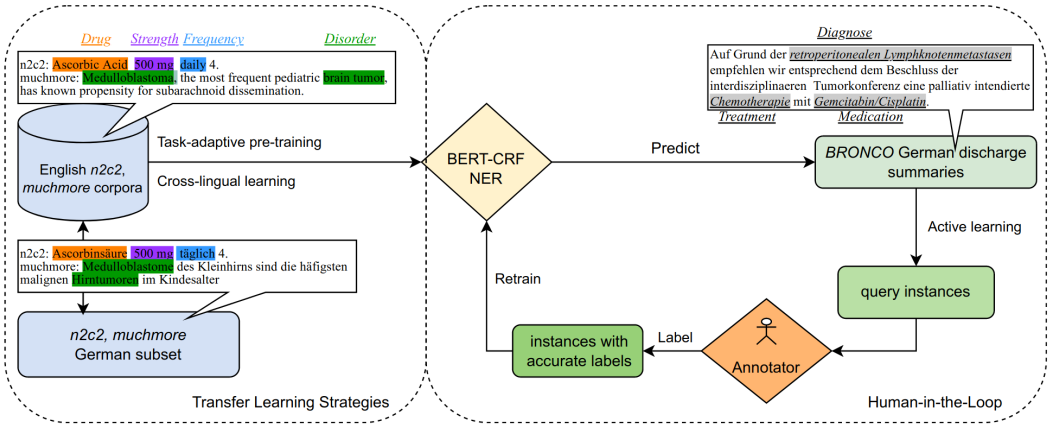

Figure 1: Overview of our research project. We utilize a BERT-CRF architecture as the base NER model. There are three relevant datasets for training and testing which are described in section 2. The box on the left side of the figure illustrates the transfer learning strategies: task-adaptive and cross-lingual learning for preparing our base NER model, which are explained in section 3. Human-in-the-loop shown in the right box is the main part of our project, we detail it in section 4. The design of the user interface for human-in-the-loop is discussed in section 5.

learning involving human-in-the-loop (HITL) (see section 4). In contrast to Chaudhary et al. (2019), we also focus on designing a user interface to obtain human annotations by refining the predictions of the base NER model and pay more attention to the human factors in real-time training (see section 5).

The NER model used in our framework is a BERT-CRF as in Liu et al. (2020), which consists of a BERT-encoder (Devlin et al., 2018) and CRF classifier (Lafferty et al., 2001). The hidden states output from the last layer of the BERT-encoder is fed into the CRF classifier as sequence input to predict the sequence of entity labels.

Training and Test Data Our target task is to extract entities from the German *BRONCO* (Kittner et al., 2021) dataset, which is a collection of discharge summaries with annotated medication terms, i.e. drugs, strength and duration etc, but also other important biomedical information, such as anatomies, diagnosis, and treatments. Two available corpora from the biomedical domain, *n2c2* (Henry et al., 2019) and *muchmore* (Widdows et al., 2002)[1] have relevant context and useful annotations for our task. Hence, we apply these two datasets for pre-training our base model in order to transfer knowledge to the target task. Since neither *n2c2* nor *muchmore* have a matching entity label set for our task, we need to reorganize the training data and define a new joint entity label set. More information about the datasets and defining the entity label set is described in the Appendix A.

[1] http://muchmore.dfki.de

3 Transfer Learning Strategies

We do not train the base NER model from scratch but pre-train it in this part of the work using transfer learning strategies. The aim of pre-training is to transfer the knowledge from annotated English clinical narratives data and German biomedical knowledge to the base NER model for processing the German discharge summaries.

Task-adaptive Pre-training To date, domain-specific language models such as BioBERT (Lee et al., 2020b), clinicalBERT (Alsentzer et al., 2019), medBERT (Rasmy et al., 2021) and BEHRT Li et al. (2020) that are pre-trained on large collections of PubMed abstracts, clinical documents, or electronic health records, are supposed to learn domain knowledge and can directly be applied to downstream tasks in the biomedical domain. However, domain-adaptive pre-training does not lead to much improvement in the downstream tasks over the general BERT model (Gururangan et al., 2020; Laparra et al., 2021). Whereas domain-adaptive pre-training makes use of data from the target domain, task-adaptive pre-training directly uses unlabeled data from the target task to adapt a pre-trained language model using a language modeling objective, e.g. masked language modeling.

Task-adaptive pre-training with text derived from the specific tasks can further benefit the in-domain LMs performing on these tasks in biomedical domain (Laparra et al., 2021). In our work,

we use the available training data for task-adaptive pre-training of the in-domain LMs and find the best pre-trained LM to transfer domain knowledge to our base model. We compare the in-domain LMs and the general BERT model with two criteria: efficiency and accuracy.

Zero- and Few-shot Cross-Lingual Learning Cross-lingual learning is a common approach to alleviate the problem of lacking in-language training data but rich annotated English data is available (Xie et al., 2018b; Pires et al., 2019; Plank, 2019; Wang et al., 2019; Zhao et al., 2020; Lauscher et al., 2020). The zero-shot setup assumes that no annotated training data is available in the target language, and multilingual LMs (Pires et al., 2019; Lample and Conneau, 2019) have shown their cross-lingual generalization capabilities in different NLP tasks across ranges of non-English languages (Hu et al., 2020). However, the impact of linguistic properties of different languages in multilingual models is not yet thorough evaluated (Virtanen et al., 2019). Research in few-shot transfer learning (Zhang et al., 2020; Chaudhary et al., 2019; Lauscher et al., 2020) has the aim of increasing the performance of the cross-lingual model with only a handful of annotated samples in the target languages. We conduct experiments both in a zero- and few-shot setting to investigate the effectiveness of the multilingual LMs in our task compared to the monolingual in-domain LMs. Considering that a multilingual model has ten times larger size than the monolingual variants, we also evaluate its efficiency and computation cost both in training and testing time.

4 Active Learning with Human-in-the-Loop

We apply transfer learning to prepare our base NER model with the source data: *n2c2* and *muchmore* corpora. To adapt the NER model to the *BRONCO* data, we use active learning to query samples for which we obtain accurate human labels, and improve the accuracy on the target data by retraining the model with the human feedback on these samples. In this part of the work, we do not only analyse the query strategies suitable for the BERT-based deep learning architecture, but also consider the human factors that are expected to strongly affect the human-computer interaction in a real-time training scenario.

Query Strategies from Active Learning In active learning, we attempt to cope with the problem of little annotation resource by measuring how informative each unlabeled instance is and only labeling the most informative instances with the least effort. The representative query strategies for selecting the samples to label fall into two main categories: uncertainty-based sampling (Lewis and Catlett, 1994) and query-by-committee (Seung et al., 1992). When applying active learning to sequence labeling tasks, there are two main issues that we have to address: structured output space and variable-length input. According to results from previous research, sequence level measures are superior to aggregating token-level information for sequence-labeling with CRF models (Settles and Craven, 2008; Chen et al., 2015b; Shen et al., 2017; Liu et al., 2020). We incorporate the following most representative query methods that are explored in prior work for NER tasks (Settles and Craven, 2008; Chen et al., 2015b; Shen et al., 2017; Chen et al., 2017; Siddhant and Lipton, 2018; Shelmanov et al., 2019; Chaudhary et al., 2019; Grießhaber et al., 2020; Shui et al., 2020; Ren et al., 2021; Liu et al., 2020, 2022; Agrawal et al., 2021), in our experiments:

- Lowest Token Probability (LTP) from Liu et al. (2020) as uncertainty-based sampling method;
- Batch Bayesian Active Learning Disagreement (BatchBALD) (Houlsby et al., 2011; Kirsch et al., 2019) with Monta Carlo Dropout (MC) (Gal and Ghahramani, 2016);
- Information Density (ID) (Settles and Craven, 2008; Shen et al., 2017) for addressing the outliers' problem.

We detail the mathematical formulations of the query strategies in Appendix C.

Human Factors Collaboration between the human and the model in real-time training is challenging. Most of the previous work in deep active learning only experiment with the query strategies in a simulated scenario (Culotta and McCallum, 2005a) without measuring the real-time labeling cost and the quality of annotations in practice (Haertel et al., 2008; Settles, 2011a; Wallace et al., 2018; Qian et al., 2020; Lertvittayakumjorn and Toni, 2021; Wang et al., 2021; Ding et al., 2021; Wu et al.,

2021). Due to the large number of model parameters, deep learning methods can be slow when retraining and force annotator to wait for the next query instance to be labeled (Settles, 2011b; Arora et al., 2009; Zhang et al., 2019). The more uncertain the predicted labels of the queried instance is, the more corrections are required from the annotator and may lead to inconsistencies among annotators (Chaudhary et al., 2019). Thus, in addtion to measuring the model performance, we need to consider the following human factors when evaluating the effectiveness of the query strategies in real-time training: i) *annotation workload of each query instance*; ii) *consistency between annotators*.

5 User Interface Design

The user interface is critical to the success of the HITL collaboration, as it can affect both user experience as well as the human factors listed above (Gajos et al., 2008; Kangasrääsiö et al., 2015). Hence, we aim to implement a user interface informed by recommendations from the human-computer interaction literature, in particular by addressing the four central components of user interfaces for interactive machine learning systems identified by Dudley and Kristensson (2018). In the following, we describe how we plan to realize each of these components in our system, where possible building on existing interfaces for HITL NER.

Sample review The *sample review* component allows the user to assess the state of the model on a given sample. Several available interfaces display the predictions of the current model on the sample to be labeled, with the goal to speed up the feedback assignment step (Yang et al., 2018; Lin et al., 2019; Lee et al., 2020a; Trivedi et al., 2019). In contrast, our sample review component focuses on increasing the user's understanding of the state of the model, for example by providing explanations of model predictions along with the predicted labels (Stumpf et al., 2009; Amershi et al., 2014). To this end, we will experiment with applying gradient- and occlusion-based explainability methods previously studies for sequence classification tasks (Atanasova et al., 2020).

Feedback assignment The *feedback assignment* component allows the user to provide the model with feedback, which can take various forms and constrains the type of interface needed to efficiently collect it. The above mentioned works displaying current model predictions collect label-level feedback by recording the user's binary decision on the correctness of the models suggestions, and a drop-down menu displaying the available label set in case the model prediction is incorrect. Lin et al. (2020) allow the users to mark spans of the input sequence that serve as explanations for a specific prediction. Lee et al. (2020a) additionally collect natural language explanations, using auto-completion to ensure the user provides phrases that can be handled by the system's semantic parser. They find that feedback in the form of important input spans is most efficient, and we plan to focus on this feedback format in combination with label-level annotations.

Model inspection The *model inspection* component provides the user with a compact summary of global model performance, e.g. by visualizing performance scores on the validation data. Erdmann et al. (2019) define two complementary evaluation frameworks for active learning models: *Exclusive* evaluation measures model performance on held-out data and indicates how well the model will generalize to additional unlabeled data. *Inclusive* evaluation measures annotation accuracy on the target corpus annotated by user and model jointly. We plan to implement the component such that the user can choose the appropriate metric for the task at hand.

Task overview The *task overview* component gives information about additional task-related decisions, e.g. termination criteria, that determine when to stop the annotation process for an optimal cost-benefit trade-off (Zhu and Hovy, 2007; Laws and Schütze, 2008).

6 Conclusion

We have presented our current ongoing work on extracting German biomedical information with a limited number of training sources. To evaluate the applied strategies in an experimental setting, we define our target task based on the training and test data at hand and first engage non-expert annotators in the computer-human interaction. Our long-term goal in this project is to apply our findings and generalize the tool on more diverse types of clinical documents in German. In order to evaluate the effectiveness of HITL in our system from the perspective of other stakeholders, we will be working with physicians in the future.

Acknowledgements

The proposed research is funded by the pAItient project (BMG, 2520DAT0P2).

References

Ankit Agrawal, Sarsij Tripathi, and Manu Vardhan. 2021. Active learning approach using a modified least confidence sampling strategy for named entity recognition. *Progress in Artificial Intelligence*, 10(2):113–128.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120.

Shilpa Arora, Eric Nyberg, and Carolyn Rose. 2009. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 18–26.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274.

Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime G Carbonell. 2019. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5164–5174.

Yukun Chen, Thomas A Lask, Qiaozhu Mei, Qingxia Chen, Sungrim Moon, Jingqi Wang, Ky Nguyen, Tolulola Dawodu, Trevor Cohen, Joshua C Denny, et al. 2017. An active learning-enabled annotation system for clinical named entity recognition. *BMC medical informatics and decision making*, 17(2):35–44.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015a. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015b. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Aron Culotta and Andrew McCallum. 2005a. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751.

Aron Culotta and Andrew McCallum. 2005b. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213.

John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Johann Frei and Frank Kramer. 2021. Gernermed – an open german medical ner model.

Krzysztof Z Gajos, Katherine Everitt, Desney S Tan, Mary Czerwinski, and Daniel S Weld. 2008. Predictability and accuracy in adaptive user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1271–1274.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning bert for low-resource natural language understanding via active learning. *arXiv preprint arXiv:2012.02462*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Robbie Haertel, Eric Ringger, Kevin Seppi, James Carroll, and Peter McClanahan. 2008. Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 65–68.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Antti Kangasrääsiö, Dorota Glowacka, and Samuel Kaski. 2015. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 247–251.

Youngjun Kim and Stéphane M Meystre. 2020. Ensemble method–based extraction of medication and related information from clinical texts. *Journal of the American Medical Informatics Association*, 27(1):31–38.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, et al. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*, 4(2):ooab025.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Egoitz Laparra, Aurelie Mascio, Sumithra Velupillai, and Timothy A. Miller. 2021. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearbook of Medical Informatics*, 30:239 – 244.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Florian Laws and Hinrich Schütze. 2008. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 465–472.

Dong-Ho Lee, Rahul Khanna, Bill Yuchen Lin, Seyeon Lee, Qinyuan Ye, Elizabeth Boschee, Leonardo Neves, and Xiang Ren. 2020a. Lean-life: A label-efficient annotation framework towards learning from explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 372–379.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020b. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ivan Lerner, Jordan Jouffroy, Anita Burgun, and Antoine Neuraz. 2020. Learning the grammar of prescription: recurrent neural network grammars for medication information extraction in clinical texts. *arXiv preprint arXiv:2004.11622*.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *ICML*.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. Triggerner: Learning with entity triggers as explanations for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8503–8511.

Bill Yuchen Lin, Dong-Ho Lee, Frank F Xu, Ouyu Lan, and Xiang Ren. 2019. Alpacatag: An active learning-based crowd annotation framework for sequence tagging. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations, pages 58–63.

Mingyi Liu, Zhiying Tu, Zhongjie Wang, and Xiaofei Xu. 2020. Ltp: a new active learning strategy for bert-crf based named entity recognition. *arXiv preprint arXiv:2001.02524*.

Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022. Ltp: A new active learning strategy for crf-based named entity recognition. *Neural Processing Letters*, pages 1–22.

Timothy Miller, Alon Geva, and Dmitriy Dligach. 2019. Extracting adverse drug event information with minimal engineering. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2019, page 22. NIH Public Access.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Barbara Plank. 2019. Neural cross-lingual transfer and limited annotated data for named entity recognition in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 370–375, Turku, Finland. Linköping University Electronic Press.

Kun Qian, Poornima Chozhiyath Raman, Yunyao Li, and Lucian Popa. 2020. Partner: Human-in-the-loop entity name understanding with deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13634–13635.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40.

Burr Settles. 2011a. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1467–1478.

Burr Settles. 2011b. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA. Association for Computing Machinery.

Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V Dylov. 2019. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 482–489. IEEE.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*, pages 1308–1318. PMLR.

Aditya Siddhant and Zachary C Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*.

Daniel Sonntag and Hans-Jürgen Profitlich. 2019. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. *Artificial intelligence in medicine*, 93:13–28.

Daniel Sonntag, Volker Tresp, Sonja Zillner, Alexander Cavallaro, Matthias Hammon, André Reis, Peter A Fasching, Martin Sedlmayr, Thomas Ganslandt, Hans-Ulrich Prokosch, et al. 2016. The clinical data intelligence project. *Informatik-Spektrum*, 39(4):290–300.

Irena Spasic, Goran Nenadic, et al. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, and Ulf Leser. 2017. How to improve information extraction from german medical records. *it - Information Technology*, 59(4):171–179.

Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies*, 67(8):639–662.

Gaurav Trivedi, Esmaeel R Dadashzadeh, Robert M Handzel, Wendy W Chapman, Shyam Visweswaran, and Harry Hochheiser. 2019. Interactive nlp in clinical care: identifying incidental findings in radiology reports. *Applied clinical informatics*, 10(04):655–669.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2018. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *arXiv preprint arXiv:1809.02701*.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. *arXiv preprint arXiv:2103.04044*.

Qiang Wei, Zongcheng Ji, Zhiheng Li, Jingcheng Du, Jingqi Wang, Jun Xu, Yang Xiang, Firat Tiryaki, Stephen Wu, Yaoyun Zhang, et al. 2020. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, 27(1):13–21.

Dominic Widdows, Beate Dorow, and Chiu-Ki Chan. 2002. Using parallel corpora to enrich multilingual lexical resources. In *LREC*, pages 240–245.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2021. A survey of human-in-the-loop for machine learning. *arXiv preprint arXiv:2108.00941*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018a. Neural cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1808.09861*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime G. Carbonell. 2018b. Neural cross-lingual named entity recognition with minimal resources. *CoRR*, abs/1808.09861.

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. Yedda: A lightweight collaborative text span annotation tool. *ACL 2018*, page 31.

Seid Muhie Yimam, Chris Biemann, Ljiljana Majnaric, Šefket Šabanović, and Andreas Holzinger. 2015. Interactive and iterative annotation for biomedical entity recognition. In *International Conference on Brain Informatics and Health*, pages 347–357. Springer.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2019. Interactive refinement of cross-lingual word embeddings. *arXiv preprint arXiv:1911.03070*.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020. Seqmix: Augmenting active sequence labeling via sequence mixup. *arXiv preprint arXiv:2010.02322*.

Shanshan Zhang, Lihong He, Eduard Dragut, and Slobodan Vucetic. 2019. How to invest my time: Lessons from human-in-the-loop entity extraction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2305–2313.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2020. A closer look at few-shot crosslingual transfer: The choice of shots matters. *arXiv preprint arXiv:2012.15682*.

Jingbo Zhu and Eduard Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 783–790.

A Customized Entity Labels

muchmore is a parallel corpus of English-German PubMed abstracts with UMLS (Unified Medical Language System)[2] annotations for training German UMLS vector models. The UMLS term annotations in *muchmore* are about more than thousands of label types but can be concluded into 15 UMLS concept categories. After comparing the entities between *muchmore* and *BRONCO*, four concepts (*Anatomy*, *Procedure*, *Disorder* and *Chemical*) are relevant to our task. We convert two of these labels into entity types required for our target task: (*Procedure -> Treatment*), (*Disorder -> Diagnosis*). Beyond the English *n2c2* corpus, we use the *GERNERMED* (Frei and Kramer, 2021), which was created by automatically translating a subset of English sentences from *n2c2*. We refer to it as German *n2c2*. The dataset can be used as a resource of parallel English and German sentences. Original entity annotations in *n2c2* included: Drug, Strength, Duration, Route, Form, ADE (Adverse Drug Effect), Dosages, Reason and Frequency. German *n2c2* subset does not contain the ADE and Reason labels and focuses on medication administration information, which is more close to our task setting. We apply the same medication labels as German *n2c2* to our task. As a result, our NER task contains an entity label set of $\mathcal{L} = \{$Drug, Strength, Duration, Route, Form, Dosages, Frequency, Diseases, Anatomy, Treatment, Diagnosis, Chemical$\}$. Table 1 shows the details about each dataset and its utilization in our project.

B Sequence Labeling with Subtokens

The tokenization of the BERT model is based on the WordPiece algorithm (Wu et al., 2016). Sequence labeling tasks with the BERT model are as a result done at the sub-token level. We follow the instructions of Devlin et al. (2018) and only label the first sub-tokens of each word with the BIO tags for training the CRF classifier. The remaining sub-tokens receive the same tags as the [PAD] tokens and are excluded in the loss calculation when predicting over the predefined NER label set. One example is shown in table 2

[2] http://umls.nlm.nih.gov

C Mathematical Formulations of Query Strategies

For a predicted sequence label $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}^1, ..., \tilde{\mathbf{y}}^T)$ and sequence input $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_T)$ with a length of T, we define a query strategy as $\phi\mathbf{x}$ on the predicted sequence of label $\tilde{\mathbf{y}}$ given the input x.

1. **Lowest Token Probability (LTP)**. The traditional least confidence strategy measuring the uncertainty of $\tilde{\mathbf{y}}$ by CRF model , i.e. the Viterbi parse (Culotta and McCallum, 2005b; Settles and Craven, 2008). Normally, LC is calculated based on the posterior probability $\tilde{\mathbf{y}}$:

$$\phi^{LC}(\mathbf{x}) = 1 - max_{\mathbf{y}^*} P(\mathbf{y}^*|\mathbf{x}; \theta) \quad (1)$$

   We adopt the variant of LC proposed by Liu et al. (2020) that measures the least confidence of the tokens in the in CRF:

$$\phi^{LTP}(\mathbf{x}) = 1 - min_{\mathbf{y}_i^* \in \mathbf{y}^*} P(\mathbf{y}_i^*|\mathbf{x}; \theta) \quad (2)$$

2. Bayesian Active Learning Disagreement (Houlsby et al., 2011). Gal and Ghahramani (2016) proposed **Monta Carlo Dropout (MC)**for approximating uncertainty in deep learning models. The dropout regularization techniques in deep neural networks are considered as disagreement strategy in bayesian deep learning (BALD) (Houlsby et al., 2011; Shen et al., 2017; Siddhant and Lipton, 2018; Kirsch et al., 2019; Liu et al., 2020; Shui et al., 2020; Ren et al., 2021). It estimate the mutual information between the model parameters and model outputs (Ren et al., 2021).

$$\phi^{BALD}(\mathbf{x}) = 1 - \frac{\max(count(\tilde{\mathbf{y}}^1, ..., \tilde{\mathbf{y}}^T))}{T} \quad (3)$$

   T represents the times of performing stochastic forward pass through the network applying different dropout masks that cause individual regularized output $\tilde{\mathbf{y}}$ at each time. The average result indicates how confident is the model predicting on the current input instance. Kirsch et al. (2019) adapted BALD in deep learning with batch input: $\phi^{BatchBALD}(x_{i...b})$, where $x_{i...b}$ is a batch of input with size of b .

Data	*n2c2*	*muchmore*	*BRONCO*
Properties	505 English discharge summaries (303 in training set, 202 in test set), only contain medication annotations, (German *n2c2* contains 6878 annotated German sentences)	Abstracts obtained from PubMed publications of 39 subjects, 7823 in English and 7808 in German, (6374 of them are En-De parallel aligned)	200 deidentified German discharge summaries of cancer patients
Relevant Annotations	Drug, Strength, Duration, Route, (ADE) Diseases Form, Dosages, Frequency	Disorders, Anatomy, Procedures, Chemicals	Diagnosis, Treatments, Medications
Utilization	transfer learning	transfer learning	active learning

Table 1: Training and test data in our framework. *BRONCO* is the target data in our active learning setting involving human annotators. We pre-train the base NER model with *n2c2* and *muchmore*.

Token	he	had	not	had	any	di	##ar	##r	##hea	other	than	1	episode
Tags	O	O	O	O	O	B-Disorder	x	x	x	O	O	O	O

Table 2: One example of BIO tagging on subtokens. Only the first subtoken of the entity word "diarrhea" is assigned the entity label and the remaining parts of the word is excluded in the loss computation during training.

3. **Information Density (ID)** (Settles and Craven, 2008). To address the problem of sampling outliers, the informativeness of data point $\mathbf{x}$ should be weighted by its similarity to other samples in the original dataset.

$$\phi^{ID}(\mathbf{x}) = \phi^{LTP}(\mathbf{x}) \times (\frac{1}{U} \sum_{1}^{U} sim(\mathbf{x}, \mathbf{x^{(u)}}))^{\beta} \quad (4)$$

The cross-similarity matrix $sim(\mathbf{x}, \mathbf{x^{(u)}}))$ among all instances in the dataset can be first computed once and later looked up for each sample in the active learning process. The base query function here $\phi^{LTP}(\mathbf{x})$ is replaceable. u is the set of samples in the dataset and with the size of U. β term is a parameter for controlling the importance of the similarity information added to the base informativeness for the given sample (Settles and Craven, 2008).