Cross-domain German Medical Named Entity Recognition using a Pre-Trained Language Model and Unified Medical Semantic Types

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

Abstract

Information extraction from clinical text has the potential to facilitate clinical research and personalized clinical care, but annotating large amounts of data for each set of target tasks is prohibitive. We present a German medical Named Entity Recognition (NER) system capable of cross-domain knowledge transferring. The system builds on a pre-trained German language model and a token-level binary classifier, employing semantic types sourced from the Unified Medical Language System (UMLS) as entity labels to identify corresponding entity spans within the input text. To enhance the system's performance and robustness, we pretrain it using a medical literature corpus that incorporates UMLS semantic term annotations. We evaluate the system's effectiveness on two German annotated datasets obtained from different clinics in zero- and few-shot settings. The results show that our approach outperforms task-specific Condition Random Fields (CRF) classifiers in terms of accuracy. Our work contributes to developing robust and transparent German medical NER models that can support the extraction of information from various clinical texts.

1 Introduction

Information extraction from the large volume of unstructured text generated in hospitals and clinics 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, assessments, medication and symptoms, offering a better context for clinical decision-making (Chen et al., 2015; Spasic et al., 2020).

In the field of Natural Language Processing (NLP), the problem of automatically and accurately extracting specific terms from text data is approached as a Named Entity Recognition (NER) task. NER methods ranging from rule-based to deep learning methods are the core technologies for automatically identifying medical instances from clinical narratives, such as diseases, diagnosis, drugs, and treatments (Sonntag et al., 2016; Sonntag and Profitlich, 2019; Miotto et al., 2018; Lerner et al., 2020; Wei et al., 2020; Kim and Meystre, 2020; Bose et al., 2021). Building clinical NER systems for non-English languages, e.g. German in our case, is challenging due to data scarcity. Only a few real-world annotated resources in German are publicly available (Starlinger et al., 2017; Kittner et al., 2021). This problem can be overcome by cross-domain transfer learning, where models transfer knowledge learned from data-rich relevant domains to domain-specific target tasks with less or no annotated data (Wang et al., 2019; Xie et al., 2018; Yuan et al., 2020; Plank, 2019; Artetxe et al., 2020; Lauscher et al., 2020).

We propose a simple but effective transfer learning framework based on a German BERT[1] encoder that is given a prompt consisting of a semantic type from UMLS semantic network[2], followed by a separator token and the medical text, e.g. *"[CLS]Clinical Drug[SEP]Zofran 4mg for nausea."*. On top of the encoder is a binary token classifier which predicts a probability for each token to determine whether it belongs to the given semantic type or not. Our approach, denoted as BERT-SNER (code[3]) and depicted in Figure 1, is based on three insights from recent research in transfer learning: i) Pre-trained Language Models (PLMs), e.g. BERT (Devlin et al., 2019), facilitate downstream tasks in specific domains (Lee et al., 2020; Alsentzer et al., 2019; Rasmy et al., 2021). ii) Prompting PLMs is becoming increasingly popular for solving low-resource NER tasks, as it can successfully exploit

[1] https://www.deepset.ai/german-bert

[2] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

[3] https://github.com/sitingGZ/bert-sner.git

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 259–271
July 14, 2023 ©2023 Association for Computational Linguistics

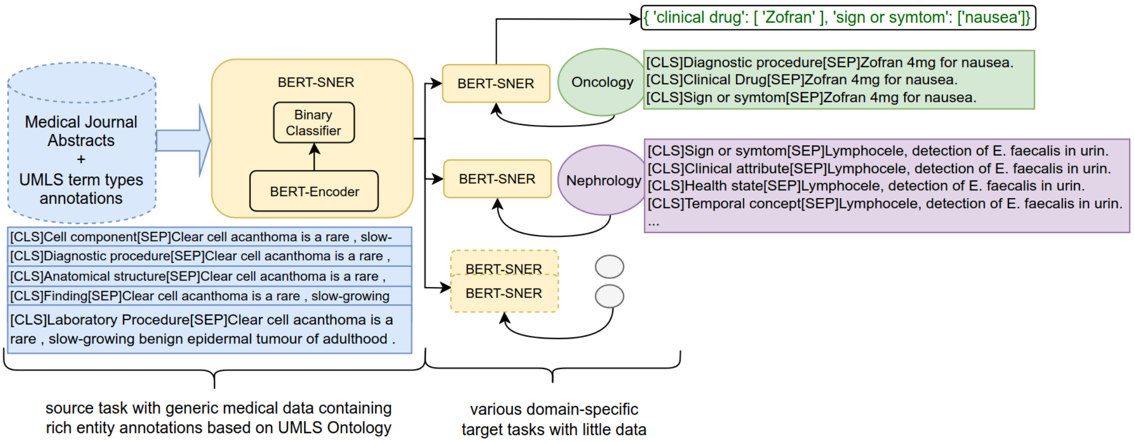

{ 'clinical drug': ['Zofran'], 'sign or symtom': ['nausea']}
Medical Journal Abstracts + UMLS term types annotations
BERT-SNER
Binary Classifier
BERT-Encoder
BERT-SNER
Oncology
[CLS]Diagnostic procedure[SEP]Zofran 4mg for nausea.
[CLS]Clinical Drug[SEP]Zofran 4mg for nausea.
[CLS]Sign or symtom[SEP]Zofran 4mg for nausea.
BERT-SNER
Nephrology
[CLS]Sign or symtom[SEP]Lymphocele, detection of E. faecalis in urin.
[CLS]Clinical attribute[SEP]Lymphocele, detection of E. faecalis in urin.
[CLS]Health state[SEP]Lymphocele, detection of E. faecalis in urin.
[CLS]Temporal concept[SEP]Lymphocele, detection of E. faecalis in urin.
...
BERT-SNER
BERT-SNER
[CLS]Cell component[SEP]Clear cell acanthoma is a rare , slow-
[CLS]Diagnostic procedure[SEP]Clear cell acanthoma is a rare ,
[CLS]Anatomical structure[SEP]Clear cell acanthoma is a rare ,
[CLS]Finding[SEP]Clear cell acanthoma is a rare , slow-growing
[CLS]Laboratory Procedure[SEP]Clear cell acanthoma is a rare , slow-growing benign epidermal tumour of adulthood .
source task with generic medical data containing rich entity annotations based on UMLS Ontology
various domain-specific target tasks with little data

Figure 1: An overview of the transfer learning framework with BERT-SNER. We first train the model using a generic medical corpus with UMLS semantic term types as entity labels and further apply the model to different clinical domain-specific NER tasks with no or limited annotated training data.

generic knowledge learned in the pre-training tasks (Cui et al., 2021; Chen et al., 2021; Wang et al., 2022). iii) The Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) is a useful knowledge source for mining medical terms in both biomedical and clinical documents (Aronson, 2001, 2006; Savova et al., 2010; Perez-Miguel et al., 2018; Kang et al., 2021; Michalopoulos et al., 2020).

The lack of domain-specific annotations is our motivation to develop models that can easily be adapted after pre-training on non-domain-specific annotated data. In our transfer learning experiments, we first derive training data from the open-source MUCHMORE corpus[4] to train BERT-SNER. MUCHMORE consists of German abstracts from 41 medical journals and entities are annotated with 134 UMLS semantic types (Archive 2001[5]). For more details on the annotation process of this corpus, please refer to Volk et al. (2002). After that, we map the entity labels of the two clinical target tasks to UMLS semantic types to be consistent with the annotations in the MUCHMORE corpus and perform zero- and few-shot experiments with 10, 50 and 100 shots for the two clinical target tasks.

The contributions of our work can be summarized as follows: 1) Our approach addresses low-resource German clinical NER tasks effectively. 2) We identify effective ways of transferring open-source medical knowledge for improving the performance of German clinical NER models.

2 Approach

Our approach explores the feasibility of knowledge transfer between different datasets by incorporating UMLS semantic term types to unify the entity labels. Table 1 shows how we construct training data from different domains to train BERT-SNER.

Input	Target
[CLS] Clinical Drug [SEP] Zofran 4mg for nausea	[0, 1, 1, 0, 1, 0, 0, 0]
[CLS] Sign or Symptom [SEP] Zofran 4mg for nausea	[0, 1, 1, 1, 0, 0, 0, 0, 1]
[CLS] Diagnostic Procedure [SEP] Zofran 4mg for nausea	[0, 0, 0, 0, 0, 0, 0, 0]

Table 1: Examples of training data (translated from German to English) using UMLS semantic types as entity labels. For each preceding entity label, if corresponding entity phrases (highlighted in orange) are found in the medical text, the tokens of the entity label and the entity phrases are annotated as class 1. The remaining part of the input is marked as class 0. If no entity phrase can be extracted for a given entity label (here *Diagnostic Procedure*), the entire target sequence contains only class 0 labels.

We compare the resulting NER system to a baseline architecture of BERT encoder combined with a task-specific conditional random fields (CRF) classifier (Wallach, 2004), i.e. BERT-CRF (Chaudhary et al., 2019; Souza et al., 2019; Pang et al., 2019; Liu et al., 2022; Mahendran and McInnes, 2021). In contrast to BERT-CRF models, BERT-SNER does not require the introduction of new task-specific parameters for solving the cross-domain target tasks, which benefits few-shot fine-tuning, while the BERT-CRF models fail if there are less than 100 samples in the target domain available.

[4] https://muchmore.dfki.de/resources1.htm

[5] https://lhncbc.nlm.nih.gov/semanticnetwork/SemanticNetworkArchive.html

Entity Type	Description	Semantic Type
DIAG	A disease, a symptom or a medical observation that can be matched with the German Modification of the International Classification of Diseases.	Sign or Symptom; Disease or Syndrome; Finding
TREAT	A diagnostic procedure, an operation or a systemic cancer treatment that can be found in the Operation.	Diagnostic Procedure; Therapeutic or Preventive Procedure
MED	A pharmaceutical substance or a drug that can be related to the Anatomical Therapeutic Chemical Classification System.	Pharmacologic Substance; Clinical Drug

Table 2: Original entity types and descriptions in BRONCO, and the best-matched selected semantic types from UMLS semantic network.

We use two datasets from different German clinical domains as target tasks: the Berlin-Tübingen-Oncology Corpus BRONCO (Kittner et al., 2021) and Ex4CDS (Roller et al., 2022). BRONCO consists of German discharge summaries for cancer patients annotated with medical entities of interest, such as *Medication (MED)*, *Diagnosis (DIAG)* and *Treatment (TREAT)*. Ex4CDS is a corpus of textual explanations for supporting system predictions of three possible outcomes (rejection, infection, graft failure) after kidney transplantation in the nephrology clinic. It focuses on entities that indicate the patient's *Health State* as well as *Laboratory Measures* after a Process. Table 7 presents the number of training samples and Table 9 presents the most frequent annotated semantic types in Appendix A and D. In order to achieve effective cross-domain transferability, we replace the original entity types of the target tasks with the best-matched UMLS semantic types during training. The matching to semantic types is determined by the ranking of the cosine similarity scores between the hidden representations of the entity types and the semantic types. The English descriptions of entity types are provided with the BRONCO and Ex4CDS datasets, and the hidden representations of type descriptions are obtained from the final hidden states of the encoder output from an English pre-trained language model[6]. The matched semantic types are validated by domain experts. Table 2 and Table 3 show the matched results. All English words of the selected semantic types are manually translated into German in our experiments.

Entity Type	Description	Semantic Type
Condition	A pathological medical condition of a patient can describe for instance a symptom or a disease.	Sign or Symptom; Disease or Syndrome; Finding
DiagLab	Particular diagnostic procedures have been carried out.	Laboratory Procedure; Diagnostic Procedure
LabValues	Mentions of lab values.	Clinical Attribute
HealthState	A positive condition of the patient.	Health State*
Measure	Mostly numeric values, often in the context of medications or lab values, but can also be a description if a value changes, e.g. raises.	Quantitative Concept
Medication	A medication.	Pharmacologic Substance
Process	Describes particular process, such as blood pressure, or heart rate, often related to vital parameters.	Physiologic Function
TimeInfo	Describes temporal information, such as 2 weeks ago or January.	Temporal Concept

Table 3: Entity types, descriptions in Ex4CDS and the matched semantic types (*except for *HealthState*, where no proper semantic type is found and retained the natural words from the original entity type).

3 Results and Discussion

The binary classifier of BERT-SNER predicts a probability for each token in the input sequence affected by the preceding semantic type and the sentence that follows. The classification result for each token is determined by setting a threshold. If the predicted probability is less than the threshold, the token is assigned to class 0, otherwise to class 1. The lower the threshold, the higher the false positive prediction rate, and conversely, a high threshold may result in a lower recall rate. We determine the threshold value for each entity label by finding the optimal precision-recall trade-off on the validation set of both target tasks based on the calculation results using the $sklearn.metrics.precision\_recall\_curve$ function. Figure 2 presents the range of thresholds in different shot settings. In the 10-shot case in both target tasks, the predicted probabilities of each token are smaller and the thresholds for individual entity types as a result are set lower. Figures in Appendix C show more details about the range of thresholds and Precision-Recall curves in different few-shot settings. In the case where a token is assigned multiple semantic types as several classi-

[6] microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext

	BRONCO				Ex4CDS			
shots	0	10	50	100	0	10	50	100
BERT-SNER	0.56±0.011	0.43±0.023	0.63±0.018	0.70±0.014	0.31±0.015	0.41±0.011	0.66±0.024	0.72±0.013
BERT-CRF	-	-	-	0.34±0.014	-	-	-	0.24±0.012
BERT-SNER(*)	-	0.26±0.020	0.33±0.013	0.36±0.024	-	0.18±0.012	0.27±0.024	0.38±0.016

Table 4: Macro-averaged F-scores of few-shot results on two target datasets. BERT-CRF is initialized with a 135-class classifier for the source task including the 134 semantic types adding an *OUT* (outside of the entity span) class, and is first pre-trained on MUCHMORE. Then, the encoder of BERT-CRF is further fine-tuned with domain-specific classifiers for BRONCO and Ex4CDS when switching domains and datasets. BERT-SNER(*) is our proposed NER framework without pre-training on MUCHMORE, i.e. trained only on data of each target task. '-' indicates a classification failure with an F-score < 0.1. '±' indicates the variance in scores caused by 2 different seeds, 3 times of random sampling and selection of semantic types in cases with multiple best-fit semantic types for individual entity types in each target task.

	BRONCO			
	0	10	50	100
MED	0.54±0.03	0.21±0.05	0.71±0.01	0.81±0.03
TREAT	0.31±0.01	0.22±0.03	0.39±0.03	0.43±0.03
DIAG	0.48 ±0.02	0.42±0.02	0.45±0.03	0.56±0.03

Table 5: F-scores of individual entity type for BRONCO test data and the BERT-SNER model with optimal thresholds in different settings.

	Ex4CDS			
	0	10	50	100
Condition	0.30±0.03	0.50±0.03	0.67±0.01	0.72±0.03
DiagLab	0.43±0.04	0.65±0.01	0.73±0.05	0.81±0.02
LabValues	0.20 ±0.03	0.64±0.02	0.78±0.03	0.88±0.01
HealthState	0.31±0.04	0.40±0.02	0.86±0.02	0.90±0.02
Measure	0.20±0.02	0.24±0.03	0.62±0.01	0.66±0.03
Medication	0.14 ±0.02	0.22±0.02	0.22±0.03	0.22±0.01
Process	0.19±0.02	0.24±0.01	0.78±0.01	0.83±0.03
TimeInfo	0.16 ±0.02	0.16±0.02	0.41±0.01	0.60±0.02

Table 6: F-scores of individual entity type for Ex4CDS test data and the BERT-SNER model with optimal thresholds in different few-shot settings.

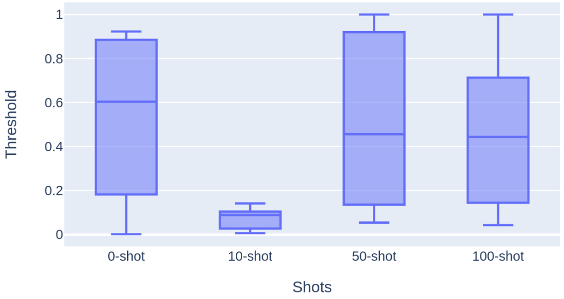


Figure 2: Ranges of thresholds by finding the best precision-recall trade-off on validation datasets. In the case of 10-shot, the prediction scores for each token in both target tasks are low, and therefore the thresholds are found lower compared to the other settings.

fication probabilities exceed the threshold, we rank the semantic types assigned to the token by their probabilities and retain the first type as the final classification result. Figure 3 in Appendix B provides an interpretation of the token-level prediction using BERT-SNER for an input sentence preceded by various semantic types.

Table 4 presents macro-averaged F-scores for BERT-SNER and baseline BERT-CRF on the two target datasets for different numbers of shots for fine-tuning the models. BERT-SNER first trained on MUCHMORE performs much better than the BERT-CRF models trained with the same resource in few-shot settings. Even without additional source data, BERT-SNER(*) shows comparable or better performance than BERT-CRF in both clinical domains. Applying the CRF classifier of the source task directly to the target tasks in the BERT-CRF framework shows worse performance than resetting the classifier with a specific label set on different target tasks.

Tables 5 and 6 present F-scores per individual entity types. When comparing the results between zero- and few-shot, we find some semantic types that can not be generalized well to the target domains, such as *(TREAT -> Diagnostic Procedure)* in BRONCO vs *(DiagLab -> Diagnostic)* in Ex4CDS, and *(MED -> Clinical Drug)* in BRONCO vs *(Medication -> Clinical Drug)* in Ex4CDS. In contrast, domain-specific entity types, *HealthState*, *LabValues* and *Process*, which are unseen or infrequent in the source task, can benefit the most from the increasing number of shots in the self-domain. These results suggest that in future work, there is a need to investigate more deeply the semantic differences of domain-specific entities matched to the same unified semantic type when experimenting with the BERT-SNER system for more diverse clinical domains. In addition, we need to examine more the impact of the amount of training data from the MUCHMORE corpus on individual entity types in target tasks.

4 Related Work

Our work focuses on solving low-resource NER tasks in the clinical domain leveraging additional resources from related domains, and in non-English languages. A common solution is to perform downstream tasks for non-English languages, especially typologically close to English through cross-lingual transfer from large-scale pre-trained multilingual BERT models (Lauscher et al., 2020; Souza et al., 2019; Jørgensen et al., 2021; Hakala and Pyysalo, 2019; Souza et al., 2019) or English language models (Artetxe et al., 2020; Plank, 2019). Frei and Kramer (2022) and Schäfer et al. (2022) attempt to use synthesised data through translation from English resources (Henry et al., 2019) to train a German medical NER model. Most of the previous works in this field have focused on a single task and it's unclear if these task-specific approaches can easily be extended to other clinical datasets with different label sets.

Sequence-to-Sequence (Seq2Seq) PLMs with prompt-based methods in another line have been shown to be useful for solving low-resource NER problems (Han et al., 2021; Gao et al., 2020; Cui et al., 2021; Yan et al., 2021; Chen et al., 2021; Wang et al., 2022). Other previous work of this line (Cui et al., 2021; Chen et al., 2021; Wang et al., 2022) utilized NER data from a resource-rich domain to fine-tune the Seq2Seq models on NER tasks before applying them to low-resource NER tasks. Although no new parameters are introduced to the pre-trained Seq2Seq language model when formulating the NER tasks in a generative framework, these methods require much effort for finding the optimal prompts and framework to transform an input sequence of tokens (words or characters) into an output sequence of entity labels. Unlike Seq2Seq NER models, our BERT-SNER model uses semantic types as prompts in front of the input directly, and the binary classifier is more efficient in terms of computational requirements, inference time and post-processing needs.

5 Conclusion

Our results suggest that transferring knowledge from publicly available medical resources with BERT-SNER is more effective than with BERT-CRF in low-resource scenarios. The overall benefit of the BERT-SNER in real-world use cases is that it can be used as an initial model to effectively develop domain-specific models in a variety of clinical applications, as it requires much less fine-tuning data than training a NER model from scratch. In future work, we will explore transfer learning more to generalize BERT-SNER to more different clinical NER tasks in low-resource situations. To apply BERT-SNER to new clinical applications without annotated samples, we will use active learning strategies such as Least Confidence oracle (Settles and Craven, 2008) to query the most informative samples to obtain annotations for fine-tuning.

Limitations

Due to strict data protection regulations and a high annotation workload in the clinical domain, obtaining more diverse target tasks to validate our approach is a challenge. In this work, we focused on only two use cases in German clinical applications and need to extend our experiments to English or other non-English languages in the field. In addition, we need to conduct more experiments in future work in order to achieve a better balance between the amount of training data required for the source and target tasks.

Acknowledgements

This work is funded by the pAItient project (BMG, 2520DAT0P2).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. page 17. American Medical Informatics Association.

Alan R Aronson. 2006. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, 1:26.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319.

Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime 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, Hong Kong, China. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Lei Li, Xin Xie, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Lightner: A lightweight generative framework with prompt-guided attention for low-resource ner. *arXiv preprint arXiv:2109.00720*.

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

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Johann Frei and Frank Kramer. 2022. GERNERMED: An open German medical NER model. *Software Impacts*, 11:100212.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 56–61.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained Models: Past, Present and Future. *AI Open*, 2:225–250.

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.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3404–3418.

Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. 2021. UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823.

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.

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.

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*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. 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*.

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.

Darshini Mahendran and Bridget T McInnes. 2021. Extracting adverse drug events from clinical notes. *AMIA Summits on Translational Science Proceedings*, 2021:420.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen H. Chen, and Alexander Wong. 2020. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *CoRR*, abs/2010.10391.

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.

Na Pang, Li Qian, Weimin Lyu, and Jin-Dong Yang. 2019. Transfer learning for scientific data chain extraction in small chemical corpus with BERT-CRF model. *arXiv preprint arXiv:1905.05615*.

Naiara Perez-Miguel, Montse Cuadros, and German Rigau. 2018. Biomedical term normalization of ehrs with umls. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

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.

Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke, and Bilgin Osmanodja. 2022. An annotated corpus of textual explanations for clinical decision support. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, page 2317–2326.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply ner to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62.

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.

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.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv e-prints*, pages arXiv–1909.

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.

Martin Volk, Bärbel Ripplinger, Špela Vintar, Paul Buitelaar, Diana Raileanu, and Bogdan Sacaleanu. 2002. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1-3):97–112.

Hanna M Wallach. 2004. Conditional Random Fields: An introduction. *Technical Reports (CIS)*, page 22.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint arXiv:2203.03903*.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of Multilingual BERT: An empirical study. *arXiv e-prints*, pages arXiv–1912.

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.

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

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A Unified Generative Framework for Various NER Subtasks. 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 5808–5822.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive Refinement of Cross-Lingual Word Embeddings. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996.

A Data statistic

	train	valid	test
MUCHMORE	10000	4000	-
BRONCO	100	100	100
Ex4CDS	100	100	100

Table 7: Number of sentences in datasets used for training, where MUCHMORE is a training source annotated with UMLS semantic types. Training data from BRONCO and Ex4CDS are limited to maximum 100 samples in each subset in few-shot experiments.

shots	10	50	100	test
MED	4	8	17	17
TREAT	3	17	40	57
DIAG	5	31	62	101
Condition	20	95	189	163
DiagLab	3	8	17	11
LabValues	8	28	60	78
HealthState	7	35	65	69
Measure	7	33	65	97
Medication	2	17	23	8
Process	7	25	44	60
TimeInfo	15	63	102	48

Table 8: Average number of annotated tokens of individual entity types from both target tasks in different few-shot samplings and test sets.

B Interpretation of the Model Outcome

In our proposed NER framework, each input sentence is iterated once with a semantic type by the BERT-SNER model. The matched semantic types based on the entity types of each task are shown in Table 2 and 3. Given an example *"stabile Funktion, keine Protenurie noch nie NTX-Versagen"* (In English: stable function, no proteinuria not ever NTX failure) from Ex4CDS, it is tokenized as ['stabile', 'Funktion,', 'keine', 'Protenurie,', 'noch', 'nie', 'NTX-Versagen'] and contains the following token-level entity annotation: ['HealthState', 'Process', 'O', 'LabValues', 'O', 'O', 'Condition'] from the original entity type set.

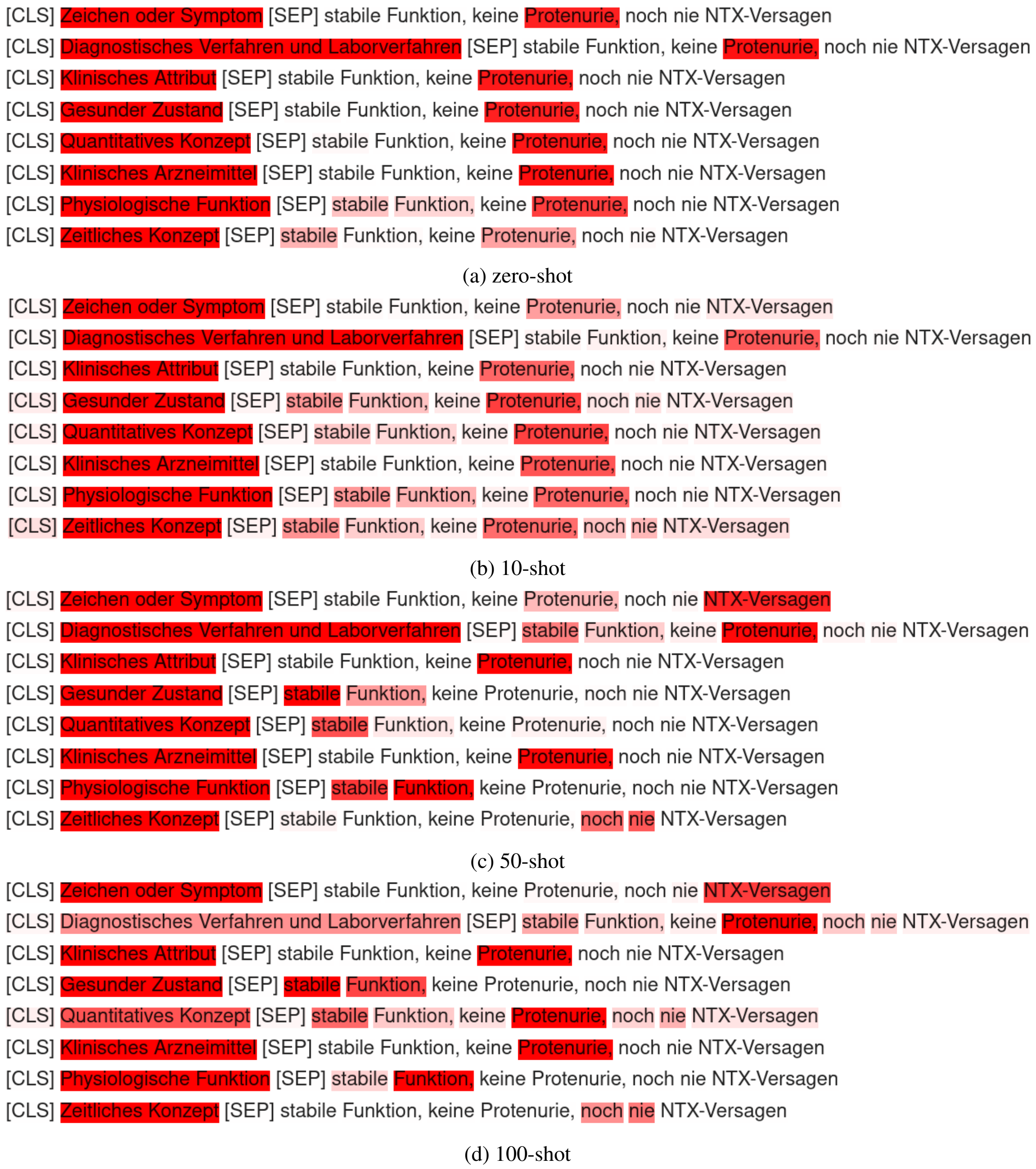
Predictions of the BERT-SNER model are made by a binary classifier, which are probabilities in the range of (0, 1). The scores predicted for the tokens of the semantic types are depending on the text input. The predicted probabilities for each token in an input sentence are affected by the semantic type in front. Figure 3 illustrates that the salience variation of each token in the same input sentence is influenced by the preceding semantic type. As a result, the final probability of each token of the input sentence is multiplied by the probability score of the first token of the given semantic type. We need to rank the scores across the applied semantic types and set a threshold to determine the final entity class for each token. In the following section, we show how to find the optimal threshold ranges to allocate the classification to each token in different few-shot settings based on the final probability scores.

C Precision-Recall Trade-off and Finding the Optimal Thresholds

The thresholds are used to determine the final classification result of a binary classifier. If the probability values are less than the threshold, assigned to class 0, while values greater than or equal to the threshold are assigned to class 1. In order to find the optimal threshold ranges in different few-shot settings, we explore the Prediction-Recall Curves and the correlations between the thresholds and F-scores according to entity types and trained shots. We can find similar phenomena in both Ex4CDS and BRONCO data, as shown in Figures 4-11.

D Most frequent annotated UMLS semantic types

134 semantic types from UMLS semantic network ontology in 2001 are annotated in MUCHMORE corpora. However, the number of annotations of each semantic type is extremely imbalanced ranging from less than 10 terms to at most 8202. We show the most frequent annotated semantic types in Table 9.



[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(a) zero-shot

[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(b) 10-shot

[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(c) 50-shot

[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(d) 100-shot

Figure 3: Predicted outcomes of zero-shot or few-shot fine-tuning for the example sentence from Ex4CDS dataset corresponding to various preceding semantic types. These eight semantic types (translated into German words) are used to replace the eight entity types during fine-tuning and inference in the target task with BERT-SNER. The color intensity indicates the value of the prediction score; the darker the color, the higher the value.

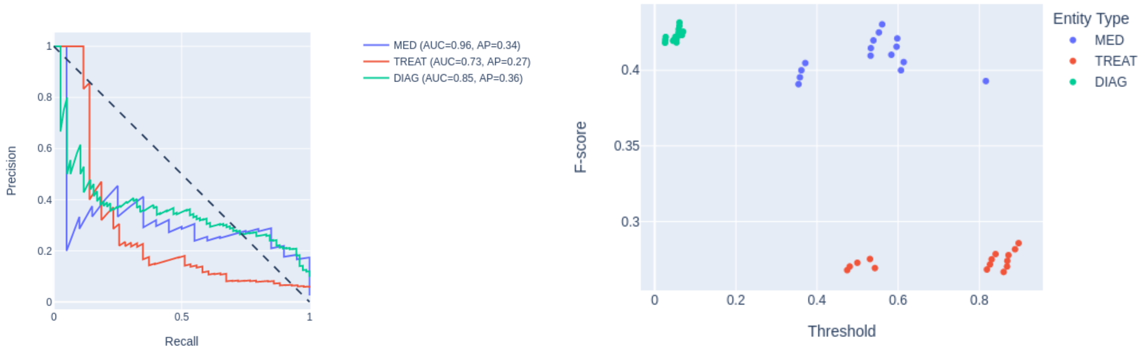


Figure 4: Zero-shot with BRONCO data. Domain-shift presents in types *TREAT* and *DIAG*. The optimal thresholds of each entity types lie in different ranges.

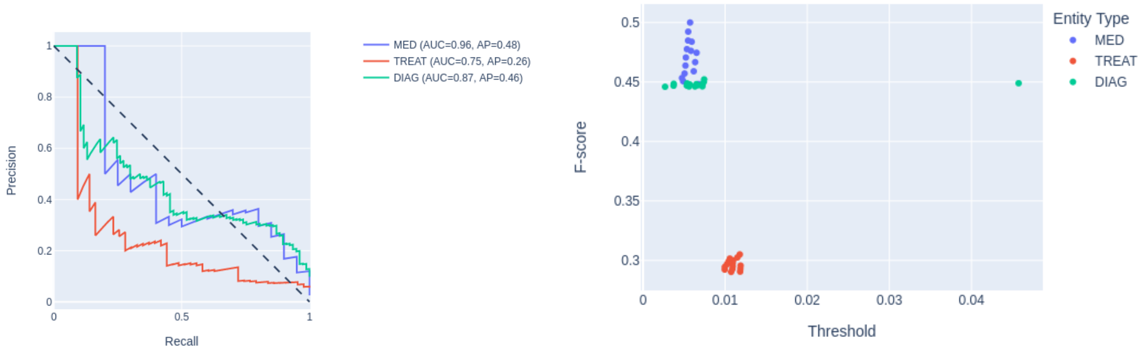


Figure 5: 10-shot with BRONCO data. The optimal thresholds for best F-scores are lowered as the BERT-SNER model has been fine-tuned with 10 samples from the target task compared to zero-shot.

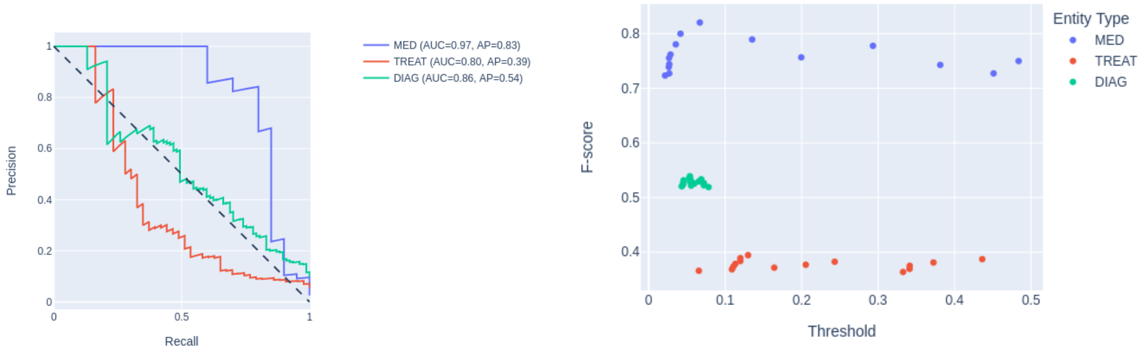


Figure 6: 50-shot with BRONCO data. The AUC scores are improved after fine-tuning with 50 samples from the target task. The optimal thresholds for best F-scores are increased compared to 10-shot fine-tuning.

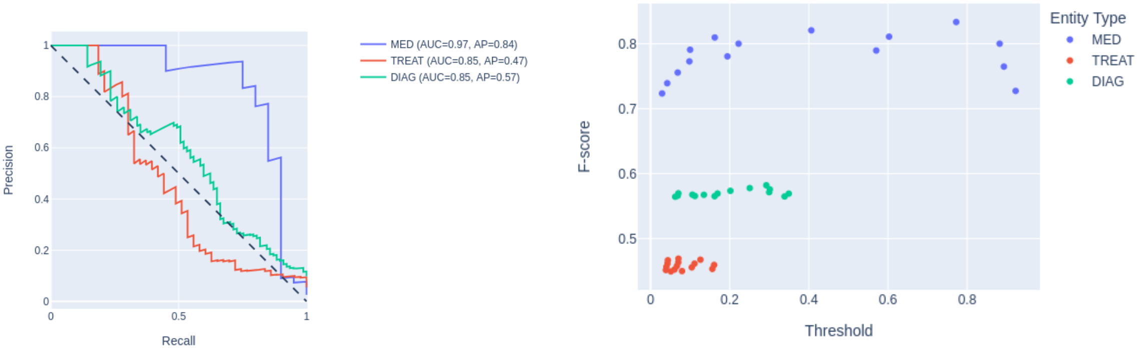


Figure 7: 100-shot with BRONCO data. The optimal thresholds for obtaining the best F-scores are increased for *MED* and *DIAG* types as the BERT-SNER model has been fine-tuned with 100 samples from the target task. From the results of F-scores and AUC scores, we find that identifying the entities of type *TREAT* in BRONCO task is a challenge for BERT-SNER.

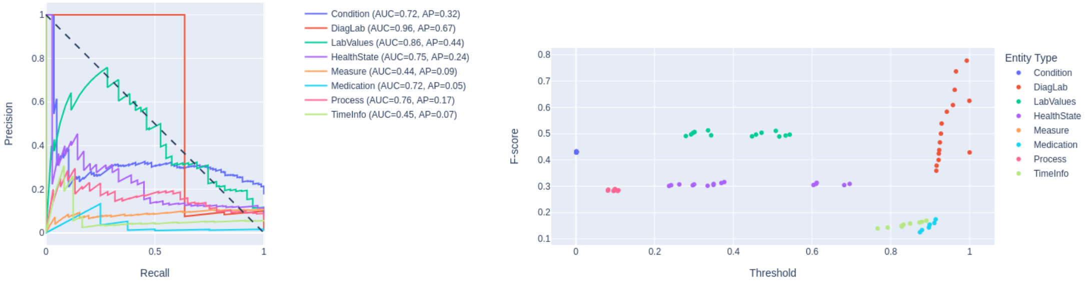


Figure 8: Zero-shot with Ex4CDS data.

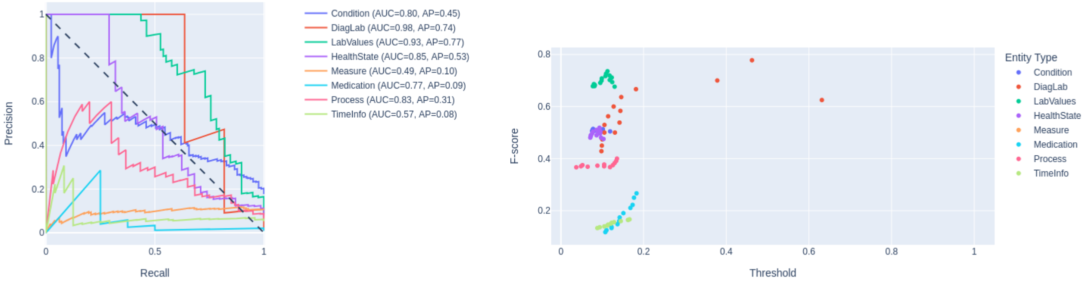


Figure 9: 10-shot with Ex4CDS data.

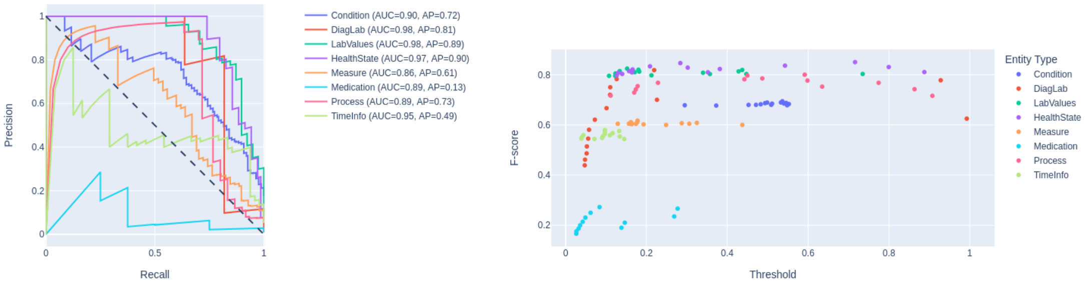


Figure 10: 50-shot with Ex4CDS data.

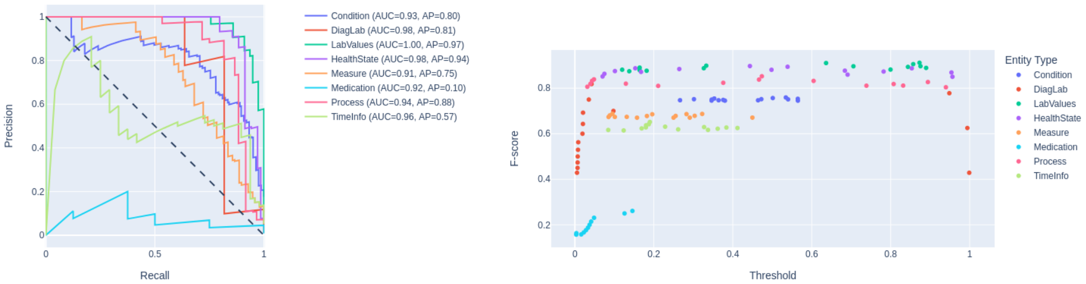


Figure 11: 100-shot with Ex4CDS data. The most challenging type for BERT-SNER in Ex4CDS is *Medication*.

ID	Type Name	Description	Amount
T101	Patient or Disabled Group	An individual or individuals classified according to a disability, disease, condition or treatment.	8202
T047	Disease or Syndrome	A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder.	7636
T023	Body Part, Organ, or Organ Component	A collection of cells and tissues which are localized to a specific area or combine and carry out one or more specialized functions of an organism. This ranges from gross structures to small components of complex organs. These structures are relatively localized in comparison to tissues.	7070
T169	Functional Concept	A concept which is of interest because it pertains to the carrying out of a process or activity.	5569
T061	Therapeutic or Preventive Procedure	A procedure, method, or technique designed to prevent a disease or a disorder, or to improve physical function, or used in the process of treating a disease or injury.	5542
T046	Pathologic Function	A disordered process, activity, or state of the organism as a whole, of a body system or systems, or of multiple organs or tissues. Included here are normal responses to a negative stimulus as well as pathololigic conditions or states that are less specific than a disease. Pathologic functions frequently have systemic effects.	3974
T191	Neoplastic Process	A new and abnormal growth of tissue in which the growth is uncontrolled and progressive. The growths may be malignant or benign.	3806
T170	Intellectual Product	A conceptual entity resulting from human endeavor. Concepts assigned to this type generally refer to information created by humans for some purpose.	3266
T081	Quantitative Concept	A concept which involves the dimensions, quantity or capacity of something using some unit of measure, or which involves the quantitative comparison of entities.	3049
T033	Finding	That which is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient. The history of the presence of a disease is a 'Finding' and is distinguished from the disease itself.	2621
T060	Diagnostic Procedure	A procedure, method, or technique used to determine the nature or identity of a disease or disorder. This excludes procedures which are primarily carried out on specimens in a laboratory.	2621
T184	Sign or Symptom	An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation.	2547
T024	Tissue	An aggregation of similarly specialized cells and the associated intercellular substance. Tissues are relatively non-localized in comparison to body parts, organs or organ components.	2533
T121	Pharmacologic Substance	A substance used in the treatment or prevention of pathologic disorders. This includes substances that occur naturally in the body and are administered therapeutically.	2403
T037	Injury or Poisoning	A traumatic wound, injury, or poisoning caused by an external agent or force.	2080
T029	Body Location or Region	An area, subdivision, or region of the body demarcated for the purpose of topographical description.	1865
T040	Organism Function	A physiologic function of the organism as a whole, of multiple organ systems, or of multiple organs or tissues.	1540
T041	Mental Process	A physiologic function involving the mind or cognitive processing.	1429
T078	Idea or Concept	An abstract concept, such as a social, religious or philosophical concept.	1309
T032	Organism Attribute	A property of the organism or its major parts.	1281
T073	Manufactured Object	A physical object made by human beings.	1226
T091	Biomedical Occupation or Discipline	A vocation, academic discipline, or field of study related to biomedicine.	1213
T123	Biologically Active Substance	A generally endogenous substance produced or required by an organism, of primary interest because of its role in the biologic functioning of the organism that produces it.	1187
T100	Age Group	An individual or individuals classified according to their age.	1149
T062	Research Activity	An activity carried out as part of research or experimentation.	1148
T079	Temporal Concept	A concept which pertains to time or duration.	1124

Table 9: Most frequent UMLS semantic types annotated in the MUCHMORE data. The numbers in the third column are the amount of annotated terms of the semantic type.