Clinical Text Classification to SNOMED CT Codes using Transformers Trained on Linked Open Medical Ontologies

Anton Hristov[1], **Petar Ivanov**[1], **Anna Aksenova**[1], **Tsvetan Asamov**[1],
Pavlin Gyurov[1], **Todor Primov**[1], **Svetla Boytcheva**[1],
[1]*OntotextAD, Bulgaria*
petar.ivanov@ontotext.com, anna.aksenova@ontotext.com,
tsvetan.asamov@ontotext.com, pavlin.gyurov@ontotext.com,
todor.primov@ontotext.com, svetla.boytcheva@ontotext.com

Abstract

We present an approach for medical text coding with SNOMED CT. Our approach uses publicly available linked open data from terminologies and ontologies as training data for the algorithms. We claim that even small training corpora made of short text snippets can be used to train models for the given task. We propose a method based on transformers enhanced with clustering and filtering of the candidates. Further, we adopt a classical machine learning approach - support vector classification (SVC) using the transformer embeddings. The resulting approach proves to be more accurate than the predictions given by Large Language Models. We evaluate on a dataset generated from linked open data for SNOMED codes related to morphology and topography for four use cases. Our transformers-based approach achieves an F1-score of 0.82 for morphology and 0.99 for topography codes. Further, we validate the applicability of our approach in a clinical context using labelled real clinical data that are not used for model training.

1 Introduction

Despite being widely applicable in healthcare, medical insurance and medical research, medical coding remains an under-automated process. This is mainly due to the huge amount of codes in medical ontologies on one hand and the very limited access to medical texts for training natural language processing systems on the other. We are presenting research on the clinical text classification task using SNOMED CT[1] codes as target values. Although the recent advances in Artificial intelligence (AI) show significant improvement in transformer-based models' performance on various Natural Language Processing (NLP) tasks, medical coding remains challenging due to the large number of classes in SNOMED (about 350K). Moreover, such systems need to be precise and reliable, hence they are usually integrated in Hospital information systems or used in Health insurance companies. Thus we propose ML-based approach that is developed on publicly available data. In addition, we compare our system to domain-specific Large Language Models (LLMs).

2 Related Work

As manual annotation in the biomedical domain is insufficient, there's a rise in the adoption of ML approaches that leverage clinical text data for task automation, predictive modelling, and knowledge discovery (Khattak et al., 2019; Mujtaba et al., 2019). However, as free-text clinical notes are unstructured, and contain spelling errors, abbreviations, and domain-specific terminology (Leaman et al., 2015), the problem of correct information extraction from clinical free-text remains a bottleneck to be properly addressed.

The limited scope of available data leads to a limited range of models that can be employed and, consequently, to poorer results. This problem can be partially alleviated by using an English-centric multilingual approach that can leverage larger sets of data available in English for applications intended for other languages (Yarowsky and Ngai, 2001).

The other way of coping with the lack of annotated training data is leveraging Large Language Models (LLMs). As those models are trained on vast amounts of data, they can perform quite well on simple classification tasks in zero-shot setting (Törnberg, 2023).

Despite the fact that LLMs are quite powerful for common-domain NLP tasks, their efficiency in medicine is yet to be explored. Singhal et al. (2022) present impressive results for LLM application in clinical domain, evaluating performance over sev-

[1] https://www.snomed.org/

Proceedings of Recent Advances in Natural Language Processing, pages 519–526
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_057

eral benchmark datasets for question answering and named entity recognition. However, Au Yeung et al. (2023) argue that such models are not ready for application in real clinical practice.

Due to its widespread adoption The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) has been employed in clinical text processing for a range of tasks. Gaudet-Blavignac et al. (2021), however, concluded that the majority of the applied approaches are rule-based.

We present a method for semi-automated annotation through the classification of machine-translated histopathology reports to SNOMED CT codes corresponding to relevant morphology or topography terms. We examine the performance of our model on four data sets composed of diagnostic and/or synoptic reports for Cervical cancer, colon cancer, lung cancer and celiac disease use cases. We address additional problems such as small sample sizes and class distribution imbalance and compare our approach with domain-specific LLMs.

3 Data

Our data collection and preparation approach is:

- curate a large set of medical terminology for pre-training of BERT models
- identify a subset of SNOMED CT codes related to a particular use-case (e.g. lung cancer)
- map to well-known medical ontologies and classifications to obtain additional descriptions (samples) for each code in the subset
- map of other (legacy, proprietary, etc) ontologies / classifications found in the validation data to SNOMED CT codes
- machine translation of validation data (descriptions in histopathology reports) from source languages to English

3.1 Data Sources

Pre-Training Data Our base model, previously described in Hristov et al. (2021), was trained on 600 thousand linked biomedical concepts. The corpus is based on MONDO[2], links to concepts from other common medical ontologies (ICD-9, ICD-10, ICD-O-3, MESH, ORDO, UMLS), and is further enriched with relevant input from Wikidata[3].

Broad Fine-Tuning Data We first fine-tune our transformer models using the SapBert scheme for self-alignment, described in Liu et al. (2021), for 1 epoch using a subset of the English UMLS 2022AA dataset. In contrast to Liu et al. (2021), who use up to 50 positive pairs for each UMLS Concept Unique Identifier (CUI), we employed subsets with up to 5, 10 and 50 names for each CUI. A positive pair is composed of two names (labels) corresponding to the particular CUI. More details on the data statistics could be found in Appendix.

We found no extra improvement in performance with the larger UMLS subsets, hence we used the smallest subset (up to 5 names for each CUI).

Narrow Fine-Tuning Data The task in the present study is to identify morphology and terminology concepts that are relevant to or found in a particular clinical text (e.g. a histopathology report). As such, we further fine-tune our model with additional data, more specifically pertaining to morphology and topography SNOMED CT codes of various anatomical structures for which validation data is available to us and are related to the four use-cases: cervical cancer, colon cancer, lung cancer and celiac disease.

Following the approach described in Hristov et al. (2021), for each SNOMED CT code in our subset we add alternative names (textual descriptions) in English from other medical ontologies, terminologies and vocabularies, among them the International Classification of Diseases, 10th revision (ICD-10)[4], the International Classification of Diseases, 9th revision (ICD-9)[5], the Systematized Nomenclature of Medicine, International Version (SNMI)[6], the National Cancer Institute Thesaurus (NCIT)[7], the Mondo Disease Ontology (MONDO)[8], and the Unified Medical Language System (UMLS)[9].

This set, composed of SNOMED CT codes and multiple names for each code, is the input to a BERT model that generates the embeddings corresponding to each name.

[2] https://mondo.monarchinitiative.org/
[3] https://www.wikidata.org/
[4] https://icd.who.int/browse10/2019/en
[5] https://apps.who.int/iris/handle/10665/39473
[6] https://bioportal.bioontology.org/ontologies/SNMI
[7] https://ncithesaurus.nci.nih.gov/
[8] https://mondo.monarchinitiative.org/
[9] https://www.nlm.nih.gov/research/umls/index.html

	Morphology		Topography	
Use case	**Classes**	**Samples**	**Classes**	**Samples**
cervical cancer	59	413	6	46
lung cancer	36	244	6	47
celiac disease	8	43	1	7
colon cancer	99	687	46	337
total	121	808	56	404

Table 1: Number of classes and samples of morphology and topography codes for each fine-tuning dataset. Note that some classes (and their respective samples) pertain to more than one use-case.

3.2 Data Integration

As described in 3.1 we limit the scope of SNOMED CT codes considered, to those related to Cervical cancer, colon cancer, lung cancer and celiac disease morphological or topographical features.

Our initial approach was to split the task in two and predict the relevant morphology codes separately from the topography codes. The histopathology reports in our validation data each contain 121 morphology and 57 topography codes, so our aim was to ensure that both types of codes are effectively predicted by our model. We observed that the resulting performance was not consistent along the two tasks (morphology and topography) and the four validation sets.

Our second approach was to fine-tune our models on the whole subset of selected SNOMED CT codes (morphology and topography codes). This approach has the benefit of using one common fine-tuning dataset, requiring fine-tuning of the model only once before applying it to any of the four validation sets.

The simplicity, however, comes at a cost - more obscure codes (classes) are less likely to be predicted, due to two main factors, the first being the imbalance in the number of samples for different codes, while the second is the difference in variability across the names for different codes. Intuitively, a greater variability in the samples for a given class is likely to result in a larger area of the embedding space being spanned by the samples for that class, while less variability would result in smaller area, but with higher probability for assigning that class within that area.

Our third and last approach, was to separate our fine-tuning data into 8 subsets corresponding to the two types of codes (morphology and topography) for each of the four use-cases. The resulting subsets are described in Table 1.

3.3 Data Augmentation

A common issue with training models on imbalanced datasets is poor modeling of the decision boundary for minority classes due to the limited number of samples. A solution comes in the form of oversampling the minority classes.

Rather than simple duplication of samples from minority classes, we employ synthetic generation of such samples using the popular Synthetic Minority Oversampling Technique (SMOTE), first described in Chawla et al. (2002). While the authors suggest combining the approach with a priori undersampling of the majority class(es), our dataset did not contain classes with a sufficient number of samples to benefit from such an approach.

SMOTE on its own works by selecting two samples from the minority class which are relatively close to each other (one is among the 5 nearest neighbours of the other) and generating a new sample along the direct line between those two samples in the feature space.

We apply SMOTE to the embeddings generated by our BERT model corresponding to samples from the minority classes. These synthetic data points are then added to the rest of the fine-tuning data and used to train a multiclass Support Vector Classifier (SVC) (see Subsection 4.4).

4 Method

Following the data preparation is the model training and application. Our proposed approach is composed of the following steps:

- start with BERT or other transformer model, ideally one that has been pre-trained on (bio)medical data
- fine-tune the selected model on a broad set of medical concepts (e.g. UMLS terminology) *(depending on the selected model, this step might be optional)*
- further fine-tune the model on a dataset made of samples more specific to the task (e.g. relevant SNOMED CT codes and corresponding names)

  (optional: perform data augmentation to improve the quality of the dataset (e.g. oversampling of minority classes))
- use BERT, a multiclass SVC or another classifier for predicting the SNOMED CT codes corresponding to each validation sample

We illustrate the proposed approach in Figure 1

4.1 Pre-Training BERT Model on Biomedical Data

We employ a BioBERT model Lee et al. (2020) trained on a biomedical corpus of 600 thousand linked concepts that we have previously described in Hristov et al. (2021).

Hereafter, we will refer to the resulting model as our pre-trained BERT.

4.2 Self-Alignment Pre-Training for BERT

Next, we take our pre-trained BERT and employ the sapBERT pre-training scheme that self-aligns the representation space of biomedical entities (Liu et al., 2021). We apply this pre-training scheme using a subset of UMLS 2022AA dataset (see broad fine-tuning data in Subsection 3.1). We use the [CLS] token rather than first-token, mean-pooling or NOSPEC (see Vulić et al. (2020)) as the representation of the input. The model was trained on a single NVIDIA RTX A1000 Laptop GPU.

Hereafter, we will refer to the resulting model as our self-aligned BERT.

4.3 Transfer Learning

Transfer learning is the process of repurposing a model trained on some task or dataset to another task or dataset. One reason for adopting such an approach is that already learnt generic features can be re-used for another task that is less rich in available training data (Bengio, 2012; Marini et al., 2021).

As mentioned in 4.2 we use our pre-trained BERT as a base model for our self-alignment pre-training. Our pre-trained BERT itself is based on bioBERT and is further trained on a large corpus of linked data based primarily on MONDO.

After just one epoch of self-alignment pre-training with a smaller subset of the UMLS dataset (as discussed in 3.1 we only use 5 names per UMLS CUI as opposed to 50), our self-aligned BERT model performs as good or better (see Section 5) than the base sapBert model (called SapBERT-PubMedBERT[10]) published along with Liu et al. (2021).

4.4 Multiclass Classification

As described in 3.1 the task for our model is to assign relevant morphology and topography SNOMED CT classes to (bio)medical texts pertaining to 4 use-cases (see Table 1). For all but one case (small intestine topography) we have multiple classes (up to 99 as in colon morphology and 121 for all use-cases morphology).

Furthermore, the number of samples per class varies widely between classes. Some classes have as little as 2 samples, while others have up to 21. To ensure that each class is represented in our test set and the data distribution is preserved, we select 25% of the datapoints to the test set and at least one object for the minor classes.

In addition, as self-aligned training requires at least 2 train samples per class, employing the SMOTE approach to add samples to minority classes ensures that even for the classes with least representation we have at least 3 samples (2 for training and 1 for testing).

Our final solution is comprised of two types of approaches to multiclass classification. Both of them use as input the embeddings of the samples generated by our self-aligned BERT model (described in 4.2).

Multiclass Classification using Self-Aligned BERT We fine-tune separate models for morphology and topography use cases with different hyperparameters (see Appendix).

We compared the performance of BERT models with multiclass SVCs trained with a variety of kernels and on subsets of the whole trained data, as described below.

Multiclass Classification using Support Vector Classifier We employ a one-vs-rest approach to multiclass classification using Support Vector Classifier (SVC). We choose this over a one-vs-one approach due to the high number of classes and low number of samples for many of the classes.

As mentioned above, the input used to train and evaluate the SVC was embeddings of the samples, rather than the raw, unprocessed samples.

We trained the SVC with linear, polynomial and RBF kernels separately for each task (morphology and topography), as well as for each combination of use-case (cervical cancer, colon cancer, lung cancer and celiac disease) and task.

4.5 Large Language Models Fine-Tuning and Prompting

Large Language Models achieve state-of-the-art results on many of the current NLP tasks, therefore

[10] https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext

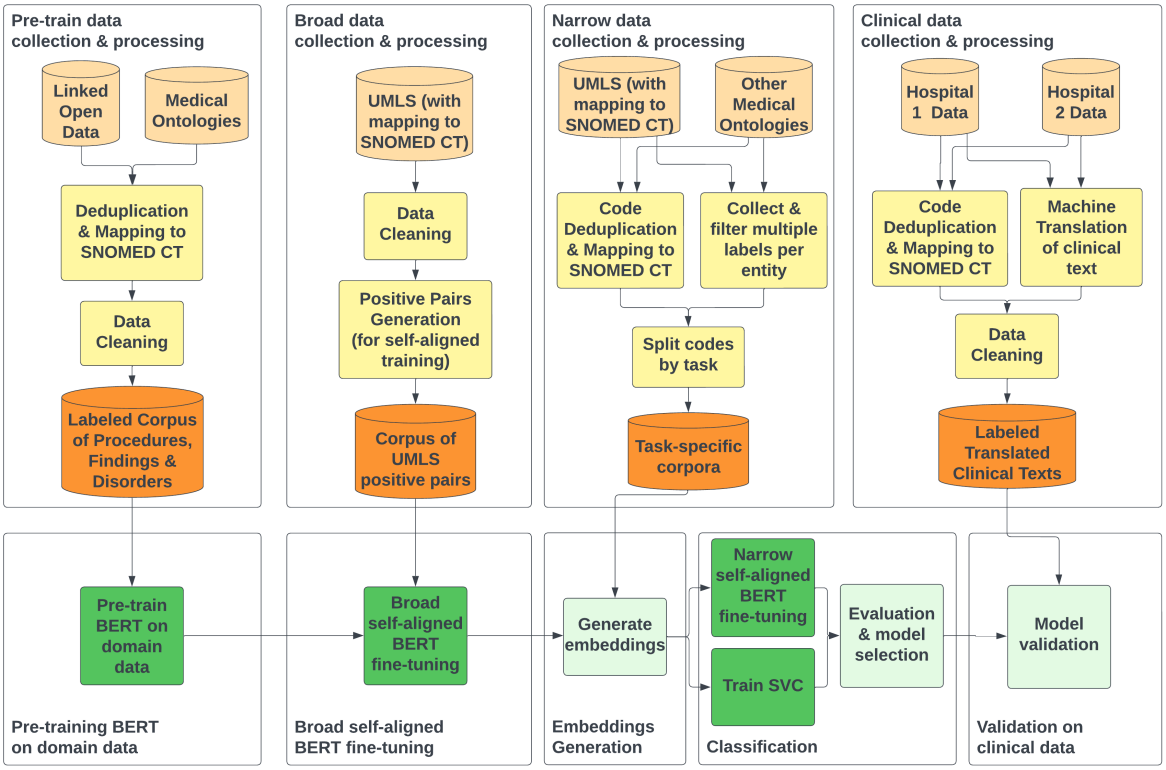

Figure 1: Steps and corresponding datasets towards building our model.

we decided to compare our methods against those. We focused on the two most widely used LLM architectures, namely GPT (Radford et al., 2019) and T5 (Raffel et al., 2020).

We focused on fine-tuning open-source BioGPT model (Luo et al., 2022) [11] and two versions of T5 adapted to biomedical domain [12].

We have performed BioGPT fine-tuning in the format of prefix-tuning by introducing additional token *[SNOMED]*, which should prompt the model to generate SNOMED codes after the input text.

Example of input data for BioGPT fine-tuning:

Transverse colon [SNOMED] 42400003

The selected T5-based models were fine-tuned in a manner similar to BioGPT. However, both of them failed to generate comprehensive codes afterwards, therefore we do not report the results for these models.

As an additional experiment we tried zero-shot prompting for ChatGPT and MedAlpaca [13]. ChatGPT refused to generate codes, stating that this question should be addressed by a healthcare professional. MedAlpaca managed to predict items similar to SNOMED CT codes, but guessed none of them. In some cases, the codes were followed by further text descriptions. For some of the examples, UMLS-like codes were predicted.

Overall, LLMs are not yet ready to solve medical coding tasks with a limited amount of data.

	Morphology			Topography		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
BioGPT	0.21	0.23	0.20	0.20	0.23	0.19
SVC	0.75	0.74	0.72	0.97	0.94	0.94
BERT	**0.84**	**0.83**	**0.82**	**0.99**	**0.99**	**0.99**

Table 2: Precision (P), Recall (R) and F1-score (F1) of LLMs and our approaches (SVC and BERT) on labels corresponding to SNOMED CT codes.

5 Experiments and Results

As described in Section 4.4 we split our dataset into train and test sets. As shown in Table 1 there's a significant imbalance between the number of classes and samples for the various use-cases. We did pre-

[11] https://huggingface.co/microsoft/biogpt

[12] https://huggingface.co/flexudy/t5-base-multi-sentence-doctor, https://huggingface.co/ozcangundes/T5-base-for-BioQA

[13] https://huggingface.co/medalpaca/medalpaca-7b

		Morphology		Topography	
Hospital	**Use-case**	**BioGPT**	**Our approach**	**BioGPT**	**Our approach**
Hospital 1	cervical cancer	0.01	**0.10** (BERT)	0.00	**0.29** (BERT)
	lung cancer	0.02	**0.48** (SVC)	0.00	**0.46** (BERT)
	celiac disease	0.00	**1.00** (SVC)	0.00	**1.00** (BERT)
	colon cancer	0.01	**0.61** (BERT)	0.03	**0.09** (BERT)
Hospital 2	colon cancer	0.09	**0.10** (BERT)	0.00	**0.34** (BERT)

Table 3: F1 score of LLM and our approach (results for best model shown) on clinical data

liminary tests by training our models using codes and samples for all use-cases and tasks which resulted in poor performance on all models for the use-cases with few classes (cervical cancer topography, lung cancer topography, celiac disease morphology and topography).

Consequently, we split our dataset in two - one part containing morphology codes only and the other topography codes only. For the BERT model, 1 epoch fine-tuning was enough to achieve near perfect results, while RBF kernel was the best performing choice for SVC. The results of our models are compared to BioGPT in Table 2.

5.1 Validation on Real Clinical Data

The models were validated on real clinical data. We were granted access to proprietary data pertaining to our use-cases by two hospitals. Hospital 1 provided us with histopathology reports in Italian that were labeled with morphology and topography codes for all four use-cases. Hospital 2 provided us with histopathology reports in Dutch labeled with morphology and topography codes for the colon use-case. We used UMLS thesaurus in combination with additional mapping resources to map the hospital labels to SNOMED CT labels and used Machine Translation to obtain an English version of the original reports (as our models are trained with samples in English).

Unlike our earlier dataset, the clinical data consisted of longer text spans, usually 1-5 (or more) sentences heavily containing medical jargon and abbreviations. Nonetheless, the performance of our approach remained high on this type of data for the majority of use-cases.

The models were compared based on F1 score (Table 3). In all 10 cases our approach outperforms BioGPT. Self-aligned BERT models are consistently better than SVC on all topography use-cases, while SVC is better at classification of lung cancer and celiac disease morphology. Notably, our approach achieved perfect scores on the 2 use-cases with the least number of training samples - celiac disease morphology and topography.

6 Conclusion

We have demonstrated an approach for extracting SNOMED CT concepts from clinical texts in multiple languages. Employing a combination of Machine Translation, Linked Open Data (both general resources, as Wikidata, and narrower, as specific medical ontologies), Transformers and more, we are able to leverage the rich resources available in English for classification of texts in languages with limited corpora available.

While we apply our approach to the clinical field, more specifically histopathology texts, we believe the same approach can be tailored to another task or another discipline with similar success, as long as both pre-trained domain-specific models (or, alternatively, enough data and computational resources for pre-training) and linked open domain-specific ontologies and terminologies are available (or could be rather easily developed).

Our model is pre-trained and fine-tuned on open data only. As such, it can be further tailored towards a specific task where richer proprietary data is also available to fine-tune the model.

One drawback of our approach is employing Machine Translation tools that are not domain-specific and cannot be fine-tuned. While not included in the present study, we expect that using relevant parallel corpora in the narrow fine-tuning step (or following it) could allow for sufficient transfer of embedded knowledge from the context-rich English corpora to the context-poor other language and allow for classification directly on the untranslated text. Obtaining such parallel corpora, however, is likely to be an even bigger obstacle. Comparison of the two approaches, where such corpora is available, would be an interesting direction for future study.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292 (ExaMode, http://www.examode.eu/).

References

Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. Ai chatbots not yet ready for clinical use. *Frontiers in digital health*, 5:60.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Christophe Gaudet-Blavignac, Vasiliki Foufi, Mina Bjelogrlic, and Christian Lovis. 2021. Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: Systematic scoping review. *Journal of Medical Internet Research*, 23(1):e24594.

Anton Hristov, Aleksandar Tahchiev, Hristo Papazov, Nikola Tulechki, Todor Primov, and Svetla Boytcheva. 2021. Application of deep learning methods to SNOMED CT encoding of clinical texts: From data collection to extreme multi-label text-based classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 557–565, Held Online. INCOMA Ltd.

Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100:100057.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37.

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.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Niccolò Marini, Sebastian Otálora, Henning Müller, and Manfredo Atzori. 2021. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. *Medical Image Analysis*, 73:102165.

Ghulam Mujtaba, Liyana Shuib, Norisma Idris, Wai Lam Hoo, Ram Gopal Raj, Kamran Khowaja, Khairunisa Shaikh, and Henry Friday Nweke. 2019. Clinical text classification research trends: systematic literature review and open issues. *Expert systems with applications*, 116:494–520.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Appendix

Hyperparameters

1. **SapBert Broad Fine-Tuning**

   We found no extra improvement from additional training after 3 and 5 epochs, hence we used the model trained for just 1 epoch on the smallest subset of UMLS.

2. **BERT-Based Multiclass Classification**

   We fine-tune our self-aligned BERT model with the train samples for all morphology or all topography codes for 1, 5 and 10 epochs. This gives us a total of 6 fine-tuned models for classification - 3 for morphology and 3 for topography codes classification. In addition to those six models, we trained a separate model for 1, 5 and 10 epochs for the colon topography task only using the samples for the 46 classes corresponding to this task.

3. **BioGPT-Based Multiclass Classification**

   As the input data was limited, we tried fine-tuning the model on a small number of epochs (1, 3, 5) and we report the result for 3 epochs as it appeared to be the best. The learning rate was set to 1e-5. No other special parameters was set as we used this method for basic evaluation against the main proposed approach. The model was trained on single NVIDIA RTX A5000 GPU. As the predictions of generative models largely depend on inference settings and candidate generation, we report the parameters related to inference too. The fine-tuned model was set to return top-5 best predictions with top-5 beam search candidates, and generation temperature set to 0.7.

UMLS Subsets

The following table presents the UMLS subsets characteristics.

UMLS Subset	Size (GB)	Positive Pairs
5 names per CUI	0.497	5,309,569
10 names per CUI	0.676	7,317,660
50 names per CUI	1.025	11,570,155

Table 4: Subsets of UMLS 2022AA (number of names per UMLS CUI) with the corresponding dataset size and total number of resulting positive pairs.