UMLS-KGI-BERT: Data-Centric Knowledge Integration in Transformers for Biomedical Entity Recognition

Aidan Mannion[1,2], **Thierry Chevalier**[3], **Didier Schwab**[1], **Lorraine Goeuriot**[1]

[1]Univ. Grenoble Alpes, CNRS, LIG, Grenoble, France

[2]EPOS SAS, Issy-les-Moulineaux, France

[3]UFR de Médecine Univ. Grenoble Alpes, La Tronche, France

Abstract

Pre-trained transformer language models (LMs) have in recent years become the dominant paradigm in applied NLP. These models have achieved state-of-the-art performance on tasks such as information extraction, question answering, sentiment analysis, document classification and many others. In the biomedical domain, significant progress has been made in adapting this paradigm to NLP tasks that require the integration of domain-specific knowledge as well as statistical modelling of language. In particular, research in this area has focused on the question of how best to construct LMs that take into account not only the patterns of token distribution in medical text, but also the wealth of structured information contained in terminology resources such as the UMLS. This work contributes a data-centric paradigm for enriching the language representations of biomedical transformer-encoder LMs by extracting text sequences from the UMLS. This allows for graph-based learning objectives to be combined with masked-language pre-training. Preliminary results from experiments in the extension of pre-trained LMs as well as training from scratch show that this framework improves downstream performance on multiple biomedical and clinical Named Entity Recognition (NER) tasks. All pre-trained models, data processing pipelines and evaluation scripts will be made publicly available.

1 Introduction

In recent times, transformer language models (Vaswani et al., 2017) have become the most popular and effective sequence modelling framework in almost all areas of applied Natural Language Processing. Unsupervised pre-training on large quantities of text allows transformers to capture rich semantic and syntactic patterns that can be transferred to many specialised language processing objectives. As such, transformer models that use the transfer learning paradigm whereby the model is trained in an unsupervised manner on a large text corpus and then fine-tuned on a downstream supervised-learning task have achieved state-of-the-art results across a wide range of general and domain-specific applications.

The proliferation of textual data in the biomedical domain (Electronic Health Records (EHRs), clinical documents, pharmaceutical specifications, etc) has precipitated the broad adoption of deep learning & NLP techniques for information extraction and processing (Li et al., 2021; Tiwari et al., 2020; Dubois et al., 2017). Moreover, it has been shown that language models are capable of encoding clinical knowledge to a certain extent (Singhal et al., 2022). Biomedical and clinical NLP, however, is widely recognised to present particular challenges that do not apply to the same extent in other domains, in particular the need to incorporate structured domain knowledge into text encodings (Chang et al., 2020). In order for neural language modelling to be reliable in a discipline as highly specialised as medicine, there is a more acute need for models to learn directly from domain-specific terminologies, as opposed to relying solely on corpus-based learning. Thus, a significant amount of research effort in the medical NLP community has been directed towards the question of how best to inject information from knowledge graphs (KGs) into LMs (He et al., 2022; Naseem et al., 2022; Li et al., 2020). However, a generalisable, widely-accepted approach to this technique that can be easily transferred across different problem settings, models and training corpora has yet to emerge. In addition, research into knowledge graph integration in NLP in the biomedical domain has tended to focus on English-language corpora; the utility and transferability of these techniques for other languages, for which less textual resources are available, as well as for multilingual models, remains therefore an under-explored area.

This paper aims to contribute to the resolution of

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 312–322

these issues by proposing a general framework for training BERT encoders (Devlin et al., 2019) using the UMLS (Unified Medical Language System, Bodenreider (2004)) alongside free-text corpora.

The main contributions of this work are as follows:

- We propose a data-centric method for formulating the KG-based learning objectives of triple classification and entity/link prediction in the language modelling paradigm, and implement a framework for training transformers using the UMLS knowledge base in parallel with masked-language pre-training.
- Pre-training on the UMLS alongside the European Clinical Case Corpus (Minard et al., 2021; Magnini et al., 2020), we show that this method brings improvements to pre-trained models across a range of biomedical entity recognition tasks in three different languages, as well as functioning as a competitive pre-training strategy that requires much less training data in comparison to state-of-the-art transformer models. We release the monolingual and multilingual model weights trained in this way, UMLS-KGI-BERT, as open-source resources for the clinical NLP research community.
- Based on this work, we release the Python library `bertify_umls`, built mainly on the `transformers` and `pandas` libraries, which allows researchers to create custom text datasets and effectively use the UMLS knowledge base as a training corpus for BERT-style LMs.

2 Related Work

2.1 Pre-trained LMs for Medical Applications

In general, the standard methodology for adapting neural text encoders to the biomedical domain has been to take a model that has been pre-trained on general-domain text corpora and continue this unsupervised pre-training on a medical corpus (Alrowili and Shanker, 2021; Lee et al., 2020; Alsentzer et al., 2019). However, recent work has suggested that, given enough training data, it is preferable to pre-train these models on large domain-specific corpora only, without starting from a general-domain checkpoint (Gu et al., 2021; Rasmy et al., 2021). In this work we explore both approaches, extending existing biomedical and general-domain models as well as training BERT models from scratch on our own generated datasets.

2.2 Knowledge-enhanced LMs

Techniques for the incorporation of knowledge graph structure into BERT models can, broadly speaking, be divided into three categories, each focusing on one of the three fundamental components of a machine learning system, i.e. 1) the training data, 2) the model architecture and 3) the objective function to be optimised. The first type of approach prioritises the augmentation of BERT's input data with information extracted from a knowledge graph. This extra information can be numerical, e.g. pre-computed graph embeddings (Jeong et al., 2019) or textual, e.g. KG triples linked to input sentences (Liu et al., 2019).

The second type of approach focuses on adapting the architecture of BERT so that its language representations become fused with knowledge graph embeddings (KGEs) (Wang et al., 2021; Peters et al., 2019; Zhang et al., 2019). Knowledge graph fusion techniques such as these have been shown to be beneficial on certain English-language medical NLP tasks (Meng et al., 2021; Roy and Pan, 2021).

Thirdly, the self-supervised pre-training objective of BERT models can be augmented using the kind of knowledge graph reasoning tasks used to build KGE models. This approach is more commonly used for knowledge graph completion (Kim et al., 2020; Yao et al., 2019) but has also been shown to be an effective strategy in the biomedical NLP domain (Hao et al., 2020).

As previously mentioned, given that the medical domain is particularly exacting in terms of requirements for the use of structured facts, the exploration of ways in which ontological knowledge can be integrated into automated text processing is a very active area of research (Khosla et al., 2020; Mondal et al., 2019). In particular, there have been multiple successful efforts to integrate the UMLS knowledge graph into BERT models, notably UmlsBERT (Michalopoulos et al., 2021), which proposes a data-augmentation technique allowing for concept and semantic type information to be linked to input text, and SapBERT (Liu et al., 2021b,a), which introduced a self-alignment strategy for learning from UMLS synonym pairs via a multi-similarity (MS) loss function to force related concepts closer to one another in BERT's repre-

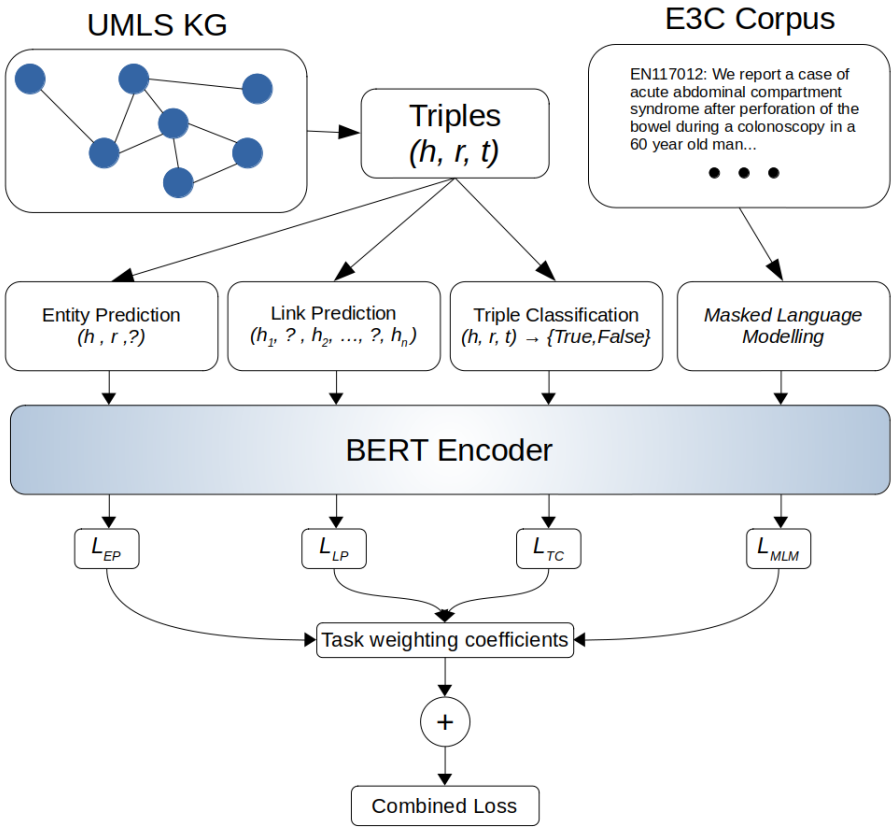

Figure 1: Overview of the UMLS-KGI pre-training process.

sentation space. Yuan et al. (2022) build on this strategy by applying MS loss to relation triples. In contrast, in this work we show that information from the UMLS can be incorporated into BERT models in a simpler way, using only cross-entropy classification loss, while also balancing this training process with standard masked-language BERT pre-training.

Recent general overviews of the landscape of AI research have highlighted the importance of data-centric approaches to building models (Zha et al., 2023; Hamid, 2022; Jakubik et al., 2022) and in light of these trends this work focuses on types 1) and 3) of knowledge base integration described above, i.e. on improving the performance of standard model architectures by constructing high-quality datasets that can be integrated into the self-supervised language modelling paradigm by modifying the BERT objective function. The motivation for this kind of approach is also to provide a pre-training framework that is more widely transferable and does not rely on any particular transformer-encoder architecture.

3 Methodology

In this work, we experiment with training BERT language models with three knowledge graph reasoning tasks derived from the UMLS, in addition to the standard masked-language modelling objective: entity prediction, link prediction and triple classification.

3.1 Dataset Construction

Formally, we consider the UMLS KG in the standard fashion, as a directed graph $G = (C, E, R)$ where C is the set of all medical concepts in the KG, E the set of all edges or relations that link these concepts to one another, and R the set of possible relation types, i.e. the labels r for each $e \in E$. The training sequences are thus generated from the KG dataset of ordered triples (h, r, t) where $(h, r) \in C \times C$ and $r \in R$. As a compendium of multiple different sources of taxonomic biomedical information, the UMLS metathesaurus contains multiple levels of granularity at which meaning representation can be analysed. We consider three such levels of granularity in our work:

- Terms - string descriptors for conceptual enti-

Table 1: Pre-training corpora sizes used in the experiments.

	Triple Classification	Entity Prediction	Paths	E3C corpus (num. documents)	Total Training Examples	Memory Footprint
French	200K	100K	64,208	25,740	389,948	604MB
Spanish	200K	100K	100K	1,876	401,876	162MB
English	200K	100K	100K	9,779	409,779	174MB
Total	600K	300K	264,208	37,395	1,201,603	940MB

ties

- CUIs (Concept Unique Identifiers) - the basic unit of meaning representation for the nodes in the knowledge graph, i.e. the elements of the set C.
- Semantic Groups - these are groupings of concepts that can be considered to define the type of entity a concept represents; e.g. anatomical structure, chemical, disorder etc.

Each concept (CUI) can be associated with multiple terms and multiple semantic groups. Thus, given that the entities h and t that make up the knowledge graph triples are represented as CUIs, in order to represent them as input text sequences for BERT models, we use the "preferred term" strings associated with the concepts h and t, except in the case of synonym relations where we randomly select another of the terms associated with the concept in question to associate with t. We also introduce a set of special tokens to represent the relation types R, of which there are seven (parent, child, synonymy, allowed qualifier, qualified by, broader, narrower). Concretely, the tokenization function for BERT models forms text classification sequences from triples in the following way;

$$\text{Tokenize}(h, r, t) = \texttt{[CLS]} w_1^h \cdots w_m^h \texttt{[REL]} w_1^t \cdots w_n^t \texttt{[SEP]} \quad (1)$$

where the w_i represent the token sequences corresponding to the strings h and t, [CLS] and [SEP] are BERT's standard classification and sequence-separation tokens as defined by Devlin et al. (2019), and [REL] is one of the relation tokens. For link prediction, we construct a dataset of variable-length paths through the KG by iteratively selecting a list of triples $(h_1, r_1, t_1), \ldots, (h_n, r_n, t_n)$ where $h_{i+1} = t_i$ to form a path $p = (h_1, r_1, h_2, \ldots, r_n, t_n)$.

Entity Prediction The entity classification task can be trivially integrated into the masked-language objective of BERT, by masking the tokens associated with the concept t.

Link Prediction We formulate link prediction as a narrow masked-language task by masking the relation tokens in the path dataset with another *hidden relation* token, for which the model is trained to fill in one of six relation types - as the triple classification and entity prediction tasks already have the partial goal of improving the model's capability to associate synonymous terms with each other, we exclude synonym relations from the path dataset.

Triple Classification Following the work of Hao et al. (2020), the triple classification objective is formulated as a binary classification problem where the model is tasked with classifying triples as true or false. In order to generate training examples of false triples, we use two different negative sampling strategies. Firstly, to provide directly contrastive examples for existing relations, we sample triples (h, r, t) where h and t belong to different semantic groups and construct corresponding false triples with the same relation type and semantic group categories, i.e. $(\hat{h}, r, \hat{t}) \notin G$ where $\hat{h}$ and $\hat{t}$ are of the same semantic group as h and t respectively. Secondly, to provide contrastive examples for relation types, we sample triples for which h and t are of the same semantic group, and form the negative training example by changing the relation type r. To ensure balance, the triple classification datasets used in this work are made up of 50% positive examples (real triples from the KG), 25% examples generated by the first negative sampling method and the rest by the second.

We perform stratified sampling on the base knowledge graph according to semantic groups, i.e. we ensure that the proportional representation of each semantic group in the knowledge-base triples for each language is maintained in the training datasets.

Mixed Objective Function In order to train BERT models using the UMLS-based reason-

ing tasks described above alongside the masked-language objective, each training example is augmented with an indicator label that tells the model which loss function to apply to the sequence in question. The overall loss function is then calculated as

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \alpha_1 \mathcal{L}_{\text{EP}} + \alpha_2 \mathcal{L}_{\text{LP}} + \alpha_3 \mathcal{L}_{\text{TC}} \quad (2)$$

where the α_i are scalar task-weighting coefficients and $\mathcal{L}_{\text{MLM}}$, $\mathcal{L}_{\text{EP}}$, $\mathcal{L}_{\text{LP}}$, and $\mathcal{L}_{\text{TC}}$ correspond to the loss values for masked language modelling, entity prediction, link prediction and triple classification respectively. We use the standard cross-entropy classification loss for all tasks.

4 Experiments

For the evaluation of the approach described in the previous section, we restrict our attention in this paper to NER tasks. Where possible, we use the datasets and training-evaluation-test splits that are publicly available via the Huggingface datasets library[1].

4.1 KG-integrated pre-training

Pre-training corpora As a resource for masked-language pre-training, we utilise the European Clinical Case Corpus (E3C) version 2.0.0[2], a freely-available multilingual corpus of clinical narratives. We evaluate our method in three different languages; English, French and Spanish. These languages were chosen as they are the three most well-represented languages in the metathesaurus for which we have access to pre-trained clinical BERT models for comparison. The sizes of the combined UMLS-E3C datasets used are shown in Table 1.

For each language, we compare the performance of 1) a transformer model trained from scratch on each monolingual dataset (KGI-BERT$_{EN,FR,ES}$) against 2) a multilingual version of the same model trained on all three datasets (KGI-BERT$_m$), 3) a pre-trained monolingual biomedical model and 4) the same pre-trained model with supplementary training on the corresponding monolingual UMLS-E3C dataset.

The UMLS-KGI models were trained for 64 epochs on each dataset, using the PyTorch implementation of the weighted ADAM optimizer (Loshchilov and Hutter, 2019) with default parameters. We use a maximal sequence length of 256 for the masked-language modelling sequences, an effective batch size of 1500 and a triangular learning rate schedule peaking at 7.5×10^{-4}. To take into account the varying sizes of the components of the pre-training dataset we set the values of the coefficients of the loss function such that they are inversely proportional to the number of documents available:

$$\alpha_i = \frac{\sum_{j=0, j\neq i}^{3} n_j}{2 \sum_{k=0}^{3} n_k}$$

where the n_k correspond to the number of documents in the training set for each UMLS-based task. In this way, the E3C masked-language loss has the same weighting as the UMLS-based task losses.

Pre-trained models For supplementary training, we make use of what are, to the best of our knowledge, the overall best-performing biomedical BERT models of their size (pre-trained using masked-language tasks only) for each language, according to baseline experiments on the NER tasks.

For French, we use DrBERT (Labrak et al., 2023), for Spanish the RoBERTa-based biomedical model released by Carrino et al. (2021), which we refer to as BioRoBERTa-ES, and for English PubMedBERT (Gu et al., 2021). For training from scratch, we use the DistilBERT model configuration (Sanh et al., 2019) with 12 encoder layers and 12 attention heads.

4.2 Evaluation corpora

We evaluate these models on nine different clinical entity recognition tasks; four in French, two in Spanish and three in English. In order to ensure a fair comparison between models and evaluate more directly the knowledge transfer capabilities of the pre-trained models, we restrict ourselves to a *one-shot* setting for all tasks, i.e. the model is given a single pass over the training data before being evaluated on the test set. For all fine-tuning runs, we use an effective batch size of 4 (we found that very frequent optimizer updates give better results in for few-shot learning), learning rate 2×10^{-5} and weight decay of 0.01.

CAS/ESSAIS CAS (Grabar et al., 2018) and ESSAIS (Dalloux et al., 2021) are corpora of clinical cases in French for which a subset is annotated with part-of-speech tags as well as semantic biomedical annotations (UMLS concepts, negation, and

[1] https://huggingface.co/datasets

[2] https://live.european-language-grid.eu/catalogue/corpus/7618

Table 2: Results on the French-language NER tasks. **Bold:** best result, underlined: next best.

	CAS-POS			CAS-SG			QUAERO-MEDLINE			ESSAI-POS		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
DrBERT-4GB	90.94	91.59	90.84	65.86	64.89	62.20	68.65	69.38	66.66	94.83	95.08	94.69
+ UMLS-KGI	**93.15**	**93.22**	**92.84**	70.82	**69.98**	67.14	71.59	72.37	69.90	94.92	94.76	94.59
KGI-BERT$_{FR}$	88.55	88.40	87.82	**71.57**	66.90	65.79	71.78	**72.93**	70.75	**95.46**	**95.40**	**95.18**
KGI-BERT$_{m}$	90.87	90.58	90.16	71.14	69.81	**67.28**	**72.04**	72.89	**70.96**	94.88	94.84	94.55

Table 3: Results on the English-language NER tasks.

	NCBI-Disease			BioRED-NER			JNLPBA04		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
PubMedBERT	93.81	94.26	93.53	**84.76**	85.33	83.35	81.57	82.59	81.13
+ UMLS-KGI	**94.65**	**95.11**	**94.46**	84.28	**85.92**	**83.64**	**85.75**	**86.04**	**85.15**
KGI-BERT$_{EN}$	89.33	89.43	88.99	82.98	85.89	82.99	81.82	82.90	81.47
KGI-BERT$_{m}$	89.40	90.04	89.16	82.67	84.63	81.97	81.24	82.47	82.02

uncertainty). We evaluate our models on the two corresponding medical POS-tagging tasks, CAS-POS and ESSAI-POS, as well as formulating a semantic-group token classification task using the CAS corpus annotations (CAS-SG).

QUAERO The QUAERO French Medical Corpus (Névéol et al., 2014) is a corpus of biomedical documents from EMEA and Medline annotated with UMLS concepts to facilitate entity recognition and document classification tasks. The NER evaluation task we make use of here, QUAERO-MEDLINE, involves semantic group identification in the Medline documents.

PharmaCoNER (Gonzalez-Agirre et al., 2019) Designed for the automated recognition of pharmacological substances, compounds and proteins in Spanish-language clinical documents, this is a manually annotated subset of the Spanish Clinical Case Corpus (SPACCC (Intxaurrondo, 2018)).

MEDDOCAN Similarly to PharmaCoNER, the MEDDOCAN corpus (Marimon et al., 2019) is an annotated subset of SPACCC, in this case with semantic entity types relevant to clinical document anonymisation, i.e. words and expressions constituting Personal Health Information (PHI).

NCBI-Disease (Doğan et al., 2014) The NCBI disease corpus is made up of PubMed abstracts with annotated disease mentions. In this work, we restrict our attention to token classification at the mention level.

BioRED (Luo et al., 2022) This corpus is designed for biomedical relation extraction and entity recognition; we focus on the latter in this work. This task can be considered a more semantically general version of the NCBI disease recognition task, in that the BioRED corpus consists of PubMed abstracts annotated with a diverse range of entity types including genes, proteins and chemicals.

JNLPBA04 NER Dataset (Collier and Kim, 2004) Developed in the context of a biomedical entity recognition shared task, this corpus consists of Medline documents annotated with mentions of DNA, RNA, proteins, cell types and cell lines.

We report the macro-averaged precision, recall and F1-score for each task. Results for the French, English and Spanish tasks can be seen in Tables 2, 3, and 4 respectively. We find that the best-performing models are in general the pre-trained checkpoints for which training has been extended via knowledge graph integration. This is unsurprising given that these are the models that have undergone the most domain-specific pre-training among all variants. It is important to highlight, moreover, the fact that the KGI-BERT variants are competitive with the pre-trained baselines for many tasks, despite being trained on less data. The largest improvements brought about by the UMLS-KGI training strategy can be seen in the French and Spanish tasks, suggesting that this technique will be more beneficial for lower-resource languages for which there is more room for improvement with respect to existing models.

The number of documents and target label classes for each evaluation task is show in Table 5.

Table 4: Results on the Spanish-language NER tasks. **Bold:** best result, underlined: next best.

	PharmaCoNER			MEDDOCAN		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
BioRoberta-ES	81.11	81.99	80.41	91.41	93.15	91.84
+ UMLS-KGI	83.52	84.30	83.90	**93.65**	**95.32**	**91.99**
KGI-BERT$_{ES}$	79.95	80.14	78.11	92.28	92.93	92.17
KGI-BERT$_{m}$	**85.05**	**85.95**	**85.49**	92.32	92.65	91.98

Table 5: Number of documents and target classes in the NER evaluation datasets

Dataset	**Train**	**Dev**	**Test**	**N. Classes**
CAS-POS	2,652	569	569	31
CAS-SG	167	54	54	15
QUAERO-MEDLINE	788	790	787	11
ESSAI-POS	5,072	1,088	1,087	34
NCBI-Disease	5,433	924	941	3
BioRED-NER	387	98	97	7
JNLPBA04	16,619	1,927	3,856	11
PharmaCoNER	500	250	250	5
MEDDOCAN	500	250	250	22

4.3 Ablation Experiments

In order to measure the relative effect of the three KG-derived pre-training tasks on downstream performance, we perform ablation experiments with the continually pre-trained models. This involved comparing the downstream performance on the NER tasks of different versions of the UMLS-extended models, each with one of the three KG-based pre-training tasks excluded from the pre-training process. For ablation, we use identical experimental settings to those described previously, except with 32 pre-training epochs rather than 64.

In general, the ablation results, for which the macro F1 scores are shown in Table 6, suggest that the majority of the benefits in terms of NER performance are brought about by the link prediction task, although there are not enough statistically significant differences among the results to fully justify this conclusion.

It is clear also that certain tasks tend to add unhelpful noise to the model with respect to some tasks, in particular the ESSAI-POS task in French and the MEDDOCAN task in Spanish. This may be due to the nature of these entity recognition tasks being more linked to general semantic patterns (i.e. parts-of-speech and identifying information) such that the addition of biomedical knowledge to the models does not improve their representation of the relevant concepts.

5 Conclusions and Future Work

This paper introduces UMLS-KGI, a framework for training BERT models using knowledge graphs requiring highly minimal adjustments to the standard language modelling paradigm. We show the potential of this method to increase the performance of BERT models on various NER tasks. The results presented in this paper suggest that for clinical NER tasks, high-quality small-scale datasets derived from structured information, alongside alongside relatively small clinical text corpora, can be as effective as large-scale corpora for pre-training BERT models. We make our models and data-processing pipelines freely available online.

Future work in this direction will involve the incorporation of more diverse graph-based reasoning tasks in the pre-training strategy with more fine-grained representation of relation types, as well as intrinsic evaluation of the UMLS-KGI-BERT language representations via embedding visualisation and interpretability studies.

Limitations

The work presented in this paper is subject to a number of limitations which will be addressed in future work. Firstly, we evaluate UMLS-KGI-BERT on a very narrow range of tasks limited to token classification - a broader range of information extraction and reasoning tasks would be necessary for a more complete picture of the utility of our pre-training methods. In addition, we only train models for mid-to-high-resource languages; to properly validate the applicability of this approach, in particular the lessening of the need to rely on large training corpora, it will be necessary to train and evaluate such models in more low-resource settings.

Table 6: Macro-F1 scores for the ablation experiments.

		Dataset			
Base Model	**KG Tasks**	**CAS-POS**	**CAS-SG**	**QUAERO-MEDLINE**	**ESSAI-POS**
DrBERT-4GB	-	90.84	62.20	66.66	**94.69**
	EP+LP	91.59	64.85	66.08	94.62
	EP+TC	90.86	62.11	66.75	94.88
	TC+LP	92.01	65.98	66.89	94.41
	all	**92.04**	**66.22**	**67.15**	94.50
		NCBI-Disease	**BioRED-NER**	**JNLPBA04**	
PubMedBERT	-	93.53	83.35	81.13	
	EP+LP	93.24	82.40	81.25	
	EP+TC	93.37	83.09	82.66	
	TC+LP	**94.13**	83.38	84.30	
	all	94.11	**83.45**	**84.36**	
		PharmaCoNER	**MEDDOCAN**		
BioRoberta-ES	-	81.11	91.84		
	EP+LP	81.12	**91.86**		
	EP+TC	82.40	91.80		
	TC+LP	83.22	91.71		
	all	**83.46**	91.77		

References

Sultan Alrowili and Vijay Shanker. 2021. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

O. Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *CoRR*, abs/2109.03570. ArXiv preprint 2109.03570.

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Andrew Taylor. 2020. Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 167–176, Online. Association for Computational Linguistics.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Sébastien Dubois, Nathanael Romano, Kenneth Jung, Nigam Haresh Shah, and David C. Kale. 2017. The effectiveness of transfer learning in electronic health records data. In *International Conference on Learning Representations*.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity

recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French Corpus with Clinical Cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Oussama H. Hamid. 2022. From model-centric to data-centric AI: A paradigm shift or rather a complementary approach? In *2022 8th International Conference on Information Technology Trends (ITT)*, pages 196–199.

Boran Hao, Henghui Zhu, and Ioannis Paschalidis. 2020. Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yong He, Cheng Wang, Shun Zhang, Nan Li, Zhaorong Li, and Zhenyu Zeng. 2022. KG-MTT-BERT: Knowledge graph enhanced BERT for multi-type medical text classification. ArXiv preprint 2210.03970.

Ander Intxaurrondo. 2018. SPACCC. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2022. Data-centric artificial intelligence. ArXiv preprint 2212.11854.

Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Lucy Park, and Sungchul Choi. 2019. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, pages 1–16.

Sopan Khosla, Shikhar Vashishth, Jill Fain Lehman, and Carolyn Rose. 2020. MedFilter: Improving Extraction of Task-relevant Utterances through Integration of Discourse Structure and Ontological Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7781–7797, Online. Association for Computational Linguistics.

Bosung Kim, Taesuk Hong, Youngjoong Ko, and Jungyun Seo. 2020. Multi-Task Learning for Knowledge Graph Completion with Pre-trained Language Models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1737–1743, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. ArXiv preprint 2304.00958.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. Ho So, and J. Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 2020*.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. 2021. Neural natural language processing for unstructured data in electronic health records: a review. ArXiv preprint 2107.02975.

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):7155. Number: 1 Publisher: Nature Publishing Group.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. 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.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*, pages 565–574.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019. K-BERT: Enabling Language Representation with Knowledge Graph. In *AAAI Conference on Artificial Intelligence*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. ArXiv:1711.05101.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022. Biored: A comprehensive biomedical relation extraction dataset. *CoRR*, abs/2204.04263.

B. Magnini, B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and

evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Zaiqiao Meng, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4672–4681, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Anne-Lyse Minard, Roberto Zanoli, Begoña Altuna, Manuela Speranza, Bernardo Magnini, and Alberto Lavelli. 2021. European clinical case corpus. Bruno Kessler Foundation.

Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical entity linking using triplet network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 95–100, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid, and Bharathi Raja Chakravarthi. 2022. Incorporating Medical Knowledge to Transformer-based Language Models for Medical Dialogue Generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 110–115, Dublin, Ireland. Association for Computational Linguistics.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, pages 24–30.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

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):86.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. ArXiv:2212.13138 [cs].

Premanand Tiwari, Kathryn L. Colborn, Derek E. Smith, Fuyong Xing, Debashis Ghosh, and Michael A. Rosenberg. 2020. Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Network Open*, 3(1). Publisher: American Medical Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194. Place: Cambridge, MA Publisher: MIT Press.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. ArXiv preprint 1909.03193.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983. Place: United States.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative

Table 7: Size of the UMLS dataset from which the KG-based pre-training corpus was sampled.

Language	Terms	CUIs	Relations
English	3,912,195	2,245,468	17,121,829
Spanish	303,978	118,061	437,578
French	202,963	171,060	669,006
Total	4,419,136	2,534,589	18,228,413

Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A Dataset Statistics

A.1 UMLS Knowledge Graph

We use the 2022AB release of the UMLS knowledge graph, which contains 8,751,471 concepts defined by 3,711,072 unique identifiers (CUIs), and 25,369,590 relations. Restricting our attention to semantic types related to human biology and medicine, we end up with the base dataset outlined in Table 7.

B Supplementary Experimental Details

Pre-trained Checkpoints We use the following pre-trained model weights downloaded from the HuggingFace model hub as baseline models;

- DrBERT: `Dr-BERT/DrBERT-4GB`
- PubMedBERT: `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`
- BioRoBERTa-ES: `PlanTL-GOB-ES/roberta-base-biomedical-clinical-es`

Model Hyperparameters The hyperparameter settings used for the pre-training on the UMLS-based dataset are shown in Table 8. The pre-training process used a linear learning rate schedule with warmup, where the learning rate increases from zero over the warmup period until it reaches the specified before decaying linearly over the rest of the training steps. In the interest of minimising the energy consumption of our experiments, we carried out very minimal hyperparameter search, leaving most parameters at their default values. The experiments were run using Python 3.8.15, with PyTorch version 2.0.0 and CUDA 11.8, along with the `transformers` library version 4.27.4.

Table 8: Hyperparameter settings for pre-training the UMLS-KGI models.

Parameter	Value
Sequence Length	256
Learning rate	0.00075
Learning rate warmup steps	10,770
Batch size	15
Gradient accumulation steps	100
MLM probability	0.15

Hardware specifications The pre-training experiments were run on four Nvidia Tesla V100 GPUs with 32GB of RAM, while the fine-tuning experiments were run on an RTX 2080 Ti with 11GB of RAM.