Biomedical Entity Representation with Graph-Augmented Multi-Objective Transformer

Andrey Sakhovskiy[1,2,3], Natalia Semenova[1,5], Artur Kadurin[5], Elena Tutubalina[1,2,4,5]

[1]Sber AI, [2]Kazan Federal University, [3]Skoltech
[4]ISP RAS Research Center for Trusted Artificial Intelligence, [5]AIRI

Correspondence: {andrey.sakhovskiy,tutubalinaev}@gmail.com

Abstract

Modern biomedical concept representations are mostly trained on synonymous concept names from a biomedical knowledge base, ignoring the inter-concept interactions and a concept's local neighborhood in a knowledge base graph. In this paper, we introduce Biomedical Entity Representation with a Graph-Augmented Multi-Objective Transformer (BERGAMOT), which adopts the power of pre-trained language models (LMs) and graph neural networks to capture both inter-concept and intra-concept interactions from the multilingual UMLS graph. To obtain fine-grained graph representations, we introduce two additional graph-based objectives: (i) a node-level contrastive objective and (ii) the Deep Graph Infomax (DGI) loss, which maximizes the mutual information between a local subgraph and a high-level graph summary. We apply contrastive loss on textual and graph representations to make them less sensitive to surface forms and enable intermodal knowledge exchange. BERGAMOT achieves state-of-the-art results in zero-shot entity linking without task-specific supervision on 4 of 5 languages of the Mantra corpus and on 8 of 10 languages of the XL-BEL benchmark.

1 Introduction

Biomedical concepts, such as diseases, symptoms, drugs, genes, and proteins, are critical for many biomedical applications, including drug discovery (Wu et al., 2018; Khrabrov et al., 2022; Zitnik et al., 2018), clinical decision making (Sutton et al., 2020; Peiffer-Smadja et al., 2020), and biomedical research (Lee et al., 2016; Tutubalina et al., 2017; Fiorini et al., 2018; Soni and Roberts, 2021; Sakhovskiy et al., 2021; Sakhovskiy and Tutubalina, 2022). The same biomedical concept may have multiple nonstandard names, abbreviations, and misspellings. Medical concept normalization (MCN), also known as medical concept linking, is a task of mapping entity mentions to a large set of medical concept names and their unique identifiers (CUIs) from a knowledge base (KB). The biomedical domain is characterized by extensive KBs such as the Unified Medical Language System (UMLS) (Bodenreider, 2004), which includes over 166 lexicons/thesauri with over 4M concepts and 15M concept names in 27 languages.

The development of meaningful and robust biomedical entity representations continues to be a challenging task for language models (LMs). Recent studies have probed LMs trained on biomedical texts in English and discovered that domain-specific pre-trained language models (PLMs), such as BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019), exhibit high levels of bias and lack synonym awareness (Sung et al., 2021). For Spanish, the second language by number of concept names in UMLS, PLMs pre-trained on clinical data fall short compared to the simplistic sparse baseline in the MCN task (Alekseev et al., 2022).

Textual triples from a KB are commonly used to incorporate knowledge into neural networks with metric learning and contrastive learning framework (Phan et al., 2019; Miftahutdinov et al., 2021; Liu et al., 2021a; Yuan et al., 2022; Zhou et al., 2022). Positive and negative pairs are created using head and tail terms of the same or different concepts, as illustrated in Fig. 1, where *maux de tête* is a French synonym of *headache* but differs from *sharp headache* (headache is a broader concept for sharp headache). Graph-based representations are another way to represent biomedical knowledge with concepts as nodes and relationships as edges. Inspired by semantic matching methods like TransE (Bordes et al., 2013) and DistMult (Yang et al., 2015), Yuan et al. (2022) proposed a method to integrate term-relation-term similarities into backbone LM. However, this approach doesn't fully utilize inter-concept interactions since it learns from individual relation triplets rather than performing an aggregation over the whole con-

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4626–4643
June 16-21, 2024 ©2024 Association for Computational Linguistics

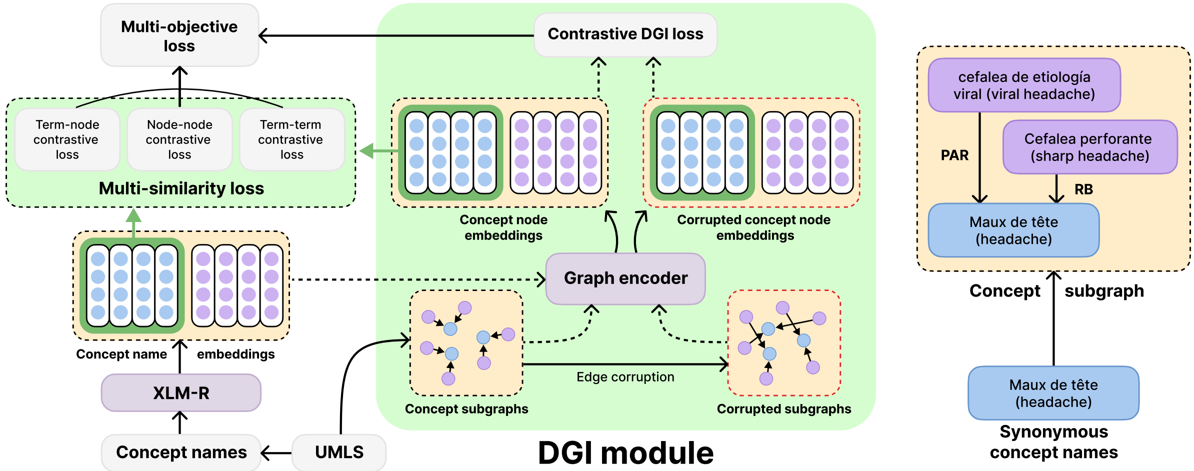

Figure 1: BERGAMOT model's architecture overview. Our model consists of two encoders for text and graph data. Graph encoder uses textual embeddings from BERT as an additional input. The final loss function is a weighted sum of four terms: term-node, node-node, term-term contrastive losses, and local-global mutual information maximization loss on node embeddings. As an example, the local subgraph contains two relation types from UMLS: `PAR` (has parent relationship) and `RB` (has a broader relationship).

cept's local neighborhood in Knowledge Graph (KG) described by UMLS KB.

In this paper, we present **B**iomedical **E**ntity **R**epresentation with **G**raph-**A**ugmented **M**ulti-**O**bjective **T**ransformer (BERGAMOT) which utilizes PLMs and graph neural networks (GNNs) to capture inter-concept and intra-concept interactions from the multilingual UMLS graph. As shown in Fig. 1, the BERGAMOT architecture includes four losses: (i) a textual term-term contrastive loss that learns from positive and negative concept name pairs; (ii) a node-node contrastive loss that learns on concept nodes to distinguish between nodes based on their local subgraphs in larger KG; (iii) DGI loss that lets a GNN distinguish between factually accurate (present in the knowledge graph) concept subgraphs and corrupted ones; (iv) an intermodal contrastive loss that enables mutual information exchange between textual and graph encoders. The source code and pre-trained models are freely available at: `https://github.com/Andoree/BERGAMOT`.

2 Related Work

MCN and entity representations There are several conventional approaches to address an MCN problem. The most common is the entity classification (Niu et al., 2019; Lou et al., 2020) into a small number of target concepts. However, UMLS and similar KBs may include millions of different concepts arranged in a hierarchical structure. Another popular approach is ranking mentions or concepts by the mutual similarity term learnt from positive and negative pairs on some corpora (Mondal et al., 2019). Aside from MCN methods, a plethora of works are focused on features and representations of entity mentions based on syntax, morphology, and synonyms (Aronson, 2001; Van Mulligen et al., 2016; Dermouche et al., 2016). Mondal et al. (2019) chose a straightforward convolutional layer as an encoder. The network was trained with the triplets of a disease mention, as well as positive and negative concept candidates. The supervised BioSyn model (Sung et al., 2020) maximizes the likelihood of synonym appearance among the most similar 20 terms. Morphology was encoded with a character-level TF-IDF representation to obtain a sparse similarity score. The distance between BioBERT (Lee et al., 2020) `CLS` tokens was utilized as a high-level dense similarity. The final similarity score is a weighted sum of both sparse and dense similarities. DILBERT (Miftahutdinov et al., 2021) introduces novel negative sampling strategies for a triplet loss, utilizing the hierarchical structure of the UMLS. Both BioSyn and DILBERT are limited to a single language (English) and a small concept subset within a specific terminology. Entity representation learning may be augmented with external knowledge from domain-specific KBs (Phan et al., 2019; Michalopoulos et al., 2021; Liu et al., 2021a,b; Yuan et al., 2022). (Phan et al., 2019) proposes encoding contextual meaning, conceptual meaning, and the similarity between synonyms during the representation learning process. Two novel training objectives are forcing the similarity between the representation of a named entity and its synonym or target concept, respectively. Despite the promising results, the ranking based on representations from this model performs worse than a plain dictionary-based baseline. UmlsBERT (Michalopoulos et al., 2021) discussed a novel knowledge augmentation strategy to utilize domain-specific knowledge from UMLS during a model's pretraining phase. SapBERT (Liu et al., 2021a) model takes advantage of a self-alignment pretraining (SAP) on the UMLS synonymous pairs. The resulting BERT-like model outperforms its predecessors (BioBERT, SciBERT, UmlsBERT). CODER, using the UMLS graph and a relational loss with SAP, outperforms SapBERT on the MCN task (Yuan et al., 2022).

Graph neural networks in biomedical domain GNNs have gained attention in the last decade, with comprehensive surveys exploring their applications in various fields, including biology and medicine. GNNs are successfully applied to a wide range of graph- and node-level tasks in fields of drug discovery and material design (Wu et al., 2018; Khrabrov et al., 2022; Zitnik et al., 2018), medicine (Ahmedt-Aristizabal et al., 2021; Gligorijević et al., 2021), and Question-Answering (QA) tasks related to knowledge graphs (KG) (Vollmers et al., 2021; Chen et al., 2020). Graph-level biomedical tasks are provided in Open Graph Benchmark (Hu et al., 2020), while KG-related evaluation is not that straightforward. The majority of approaches focused on QA evaluation or conventional KG link prediction. We can divide the considered methods into three groups: LM- or KG-based and joint LM+KG. Despite approximately similar training data (subsets of UMLS) and backbone (SciBERT, BART (Lewis et al., 2020)), the approaches are entirely different due to data preparation procedures and finetuning. Chang et al. (2020) proposed a completely KG-based biomedical benchmark. They trained TransE, ComplEx (Trouillon et al., 2016), DistMult, SimplE (Kazemi and Poole, 2018), and RotatE (Sun et al., 2019) on the SNOMED-CT dataset to compare the results with static Snomed2Vec (Agarwal et al., 2019) and Cui2Vec (Beam et al., 2020) baselines. While static methods did not surpass any KG-trained methods, KG-based models performed remarkably worse in comparison with LM-based ones due to the lack of a text encoder. An alternative approach to benchmarking is provided in the LM+KG architectures' evaluation. QA-GNN (Yasunaga et al., 2021) and GreaseLM (Zhang et al., 2022) achieved state-of-the-art scores on MedQA (Jin et al., 2021). However, they are both trained in a manner of fine-tuning LM with KG-augmentation. This strategy does not allow us to completely adapt models to the MCN task.

3 Background and architecture

3.1 UMLS knowledge graph

Let V denote a set of all concepts present in a knowledge base and R denote a set of possible relation types between concepts from V. Knowledge graphs, such as UMLS, usually store relational information in the form of relation triplets $(h, r, t) \in V \times R \times V$. Let $\mathcal{E}$ denote a set of all unique relation triplets from a given KG. Thus, the UMLS graph can be defined as an oriented edge-labeled graph $G = G(V, \mathcal{E}, R)$ with a set of nodes V, a set of labeled oriented edges $\mathcal{E}$, and a set of possible edge labels R. For each concept $c \in V$, UMLS presents a set of k synonymous concept names $S_c = \{s_1^c, s_2^c, \ldots, s_k^c\}$. For each name from S_c, UMLS stores the label of the language it came from.

3.2 Self-alignment pretraining

A reasonable and straightforward way to learn an informative representation space of biomedical entities is to represent textual knowledge from KG in the form of positive and negative term pairs and optimize some contrastive learning loss function.

In this work, we adopt the self-alignment pretraining (SAP) procedure proposed by Liu et al. (2021a). To enrich the training procedure with harder negative samples, SAP employs online hard mining for valid triplets (Mikolov et al., 2013; Gillick et al., 2019). During SAP, the model is encouraged to produce similar representations for all terms that represent the same concept (share the same CUI). At each pretraining step, we sample a set T that consists of N positive samples $(c, s_i^c, s_j^c) \in V \times S_c \times S_c$. Given T, SAP constructs all possible term triplets (s^p, s^a, s^n) such that $p = a$ and $n \neq a$. s^a is called an anchor term; s^p is a positive term for s^a (i.e., s^p and s^a are synonymous terms representing the same concept $a = p$); s^n is a negative term for s^a (i.e., s^n and s^a represent non-matching concepts). Each triple produces a positive pair (s^a, s^p) and a negative pair

(s^a, s^n). To keep only the most informative triples, we use online hard mining for valid triplets in respect to the following constraint:

$$\|f_{enc}(s^a) - f_{enc}(s^p)\| < \|f_{enc}(s^a) - f_{enc}(s^n)\| + \lambda$$

where f_{enc} is a BERT-based textual encoder, $\|\cdot\|$ is the normalized L_2-norm, and λ is a pre-defined mining margin. Thus, the mining procedure discards all the triplets such that the distance from an anchor to its negative sample is greater than the distance to its positive sample by more than λ. Let $\mathcal{P}$ and $\mathcal{N}$ denote the sets of all positive and negative term pairs, respectively. The SAP procedure utilizes the Multi-Similarity (MS) loss (Wang et al., 2019) to learn from $\mathcal{P}$ and $\mathcal{N}$.

$$\mathcal{L}_{sap} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left(\frac{1}{\alpha} \log\left(1 + \sum_{n \in \mathcal{N}_i} e^{\alpha(S_{in} - \epsilon)}\right) + \right. \\ \left. + \frac{1}{\beta} \log\left(1 + \sum_{p \in \mathcal{P}_i} e^{-\beta(S_{ip} - \epsilon)}\right)\right), \tag{1}$$

where α, β, and ϵ are the parameters of MS-loss. $\mathcal{P}_i$ and $\mathcal{N}_i$ are the sets of positive and negative samples for the anchor concept i. S_{in} and S_{ip} are the cosine similarities of anchor i to negative sample n and positive sample p, respectively.

3.3 Graph neural networks

3.3.1 Message passing layers

A common approach to capture the complex relationships between nodes in the graph is to iteratively update the representation of a node v by passing and aggregating messages from its local node neighborhood $N(v)$ using a graph neural network. Gilmer et al. (2017) proposed a general Message Passing Neural Network (MPNN), which applies a composition of message function f_m and f_u to update the node representation $h_v^{(l)}$ at the $(l+1)$-th MPNN layer:

$$h_v^{(l+1)} = f_u(h_v^{(l)}, \sum_{(r,u) \in N(v)} f_m(h_v^{(l)}, h_u^{(l)}, e_r))$$

where $N(v) = \{(r,u) | (v,r,u) \in \mathcal{E}\}$ is the set of pairs describing the local neighborhood of node v; e_r are the edge features. Each pair (r,u) indicates the presence of a directed edge of type r from node v to node u in graph G. To avoid excessive computational complexity caused by significant variation in the number of neighbors across different nodes, we use a uniformly drawn fixed-size subset of neighbors instead of the entire node neighborhood as proposed by Hamilton et al. (2017). The primary distinguishing factor among various GNN models is the selection of f_m and f_u functions in the MPNN computational block. In GraphSAGE (Hamilton et al., 2017), a common and rather simple implementation of MPNN framework, an element-wise operator (e.g., max- or mean-pooling) is used as an f_m to aggregate the vectors of neighbor nodes $N(v)$ into a single vector. The aggregated representation is further concatenated with the original representation and passed to a linear layer W^{l+1} with a non-linear activation function σ. In this work, we use the GraphSAGE implementation with mean-pooling aggregation:

$$h_v^{(l+1)} = \sigma(W^l \cdot [h_v^{(l)} \parallel MEAN(N(v))])$$

where $MEAN$ is the mean-pooling operator, $[\cdot \parallel \cdot]$ is the concatenation of two vectors. Since the parameter matrix W^l is the same for each $r \in R$ GraphSAGE operator prevents a thorough use of edge types and features. Regarding the UMLS graph, it means that GraphSAGE can only capture the textual node features and knowledge graph geometrical structure, though it doesn't consider relation-specific information.

Schlichtkrull et al. (2018) proposed Relational Graph Convolutional Network (R-GCN) architecture that performs relation-aware message passing by introducing a relation-specific parameter matrix W_r^l for each relation $r \in R$ at the neighborhood aggregation step of Message Passing block:

$$h_v^{(l+1)} = \sigma\left(\sum_{(r,u) \in N(v)} \frac{1}{|N(v)|}(W_r^l h_u^{(l)} + W_o^l h_v^{(l)})\right)$$

σ is a non-linear activation function and W_o^l is a self-loop parameter matrix.

Despite being able to perform relation-aware message passing, R-GCN shares the limitation of GraphSAGE as it does not allow learning the relative importance of neighboring nodes. Graph attention network (GAT) (Veličković et al., 2018; Brody et al., 2022) addresses the limitation by introducing the self-attention over neighboring nodes and learning the aggregated neighborhood representation as the weighted sum of neighboring nodes representations. Given two node representations $h_u^{(l-1)}$ and $h_v^{(l-1)}$, the l-th GAT layer computes the relevance of node u for the target node v as the normalized attention score $\alpha_{uv}^{(l)}$:

$$e_{uv}^{(l)} = a^T \cdot LeakyReLU(W^l \cdot [h_u^{(l-1)} \parallel h_v^{(l-1)}])$$

$$\alpha_{uv}^{(l)} = \frac{exp(e_{uv}^{(l)})}{\sum_{(r,w)\in N(v)} exp(e_{wv}^{(l)})},$$

where a is a learnable weight vector. With the attention scores obtained, the aggregated neighborhood representation is computed as a weighted sum of neighboring nodes embeddings:

$$h_v^{(l+1)} = \sigma\left(\sum_{(r,u)\in N(v)} \alpha_{uv} \cdot W^l h_u^{(l)} + W_o^l h_v^{(l)}\right)$$

3.3.2 Deep Graph Infomax framework

To enrich an LM and a GNN with graph structure knowledge, we utilize the Deep Graph InfoMax framework (Veličković et al., 2019). DGI adopts an encoder model GNN to maximize the mutual information between global graph structure and its local subgraphs. This method is grounded on minimizing a noise-contrastive loss function

$$\mathcal{L}_{dgi} = \frac{1}{N+M}\sum_{i=1}^{N} GNN\left[\log D(\vec{h_i}, \vec{s})\right] + \\ + \sum_{j=1}^{M} GNN\left[\log\left(1 - D(\vec{\tilde{h}_j}, \vec{s})\right)\right],$$

where $\vec{s}$ is a graph summary obtained as a mean embedding of nodes in a graph; D is a learnable discriminator with sigmoid activation which scores summary-graph pairs. The loss aims at distinguishing between input node representations h and negative samples $\tilde{h}$. The procedure starts with collecting initial and corrupted node features of graph G, after that we obtain representations via GNN, respectively, for each type of feature vectors. By leveraging local mutual information maximization and graph convolutional architectures, we make the features of local subgraphs to be consistent with the global properties of a larger graph.

3.4 BERGAMOT

We introduce a novel biomedical entity representation learning model BERGAMOT which leverages graph and textual encoders to infuse inter-concept interactions from a biomedical KG into LM. In BERGAMOT, we adopt and extend the pretraining procedure described in Sec. 3.2. Before moving on to the description of the pretraining procedure, let us first describe the structure of a training batch B which consists of (i) textual positive concept name pairs and (ii) local concept subgraphs.

Textual pairs We begin by sampling a set T, consisting of N random positive concept name pairs for concepts that have at least two distinct concept names. Let (u, v) denote a pair of concept names that represent the same concept c (e.g., "Maux de tête" and "headache" as shown in Fig 1). We encode both textual names with a BERT-based LM resulting in two vector representations $e_c^u = LM(u)$ and $e_c^v = LM(v)$, respectively.

Concept subgraphs A positive textual pair (u, v) in T corresponds to a single UMLS concept c. As c is also a node in the UMLS KG, we sample the graph G_c that is centered around c and includes a set of its neighboring nodes $N(c)$. Next, we have two ways to initialize the central node c from G_c with either $LM(v)$ or $LM(u)$ resulting in two graphs G_c^v and G_c^u with identical structure but different initial node features. For instance, we can initialize a central node with a BERT embedding of either "headache" or "Maux de tête" (see Fig. 1). Non-central concept nodes are initialized with LM embeddings of their random concept names independently for both graphs. By applying a L-layer graph neural network we obtain two graph representations of the concept c: $g_c^u = GNN(G_c^u)$ and $g_c^v = GNN(G_c^v)$.

BERGAMOT's design is inspired by two major goals. First, we want to encode concept names and graph structure into a single shared embedding space while preserving an essential property for entity linking: various (both textual and graph) representations of the same concept must have similar embeddings. Second, since obtaining a relevant KG subgraph is not feasible at the inference stage, we want to infuse rich structural knowledge from the UMLS graph into a language model. To reach both goals, our model simultaneously learns 4 contrastive objectives on concept representations as shown in Fig. 1.

Term-term contrastive loss $\mathcal{L}_{sap}$ To let a BERT-based model learn semantic and lexical similarity of various concept names, we apply MS-loss on textual pairs. This objective seeks to pull textual embeddings (e_c^u, e_c^v) of concept c's synonymous names closer in terms of cosine similarity, ignoring inter-concept relations. We use in-batch hard

Model	English		Spanish		Dutch		French		German		avg	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
Full test set												
Supervised SOTA (Alekseev et al., 2022)	94.25	97.12	**85.54**	92.17	85.83	87.40	84.23	**94.14**	89.55	95.02	87.88	93.17
mSapBERT	94.03	96.90	83.73	92.17	84.25	87.40	82.43	93.69	88.06	95.52	86.5	93.14
mCODER	93.14	96.68	83.73	90.96	85.04	**90.55**	**89.19**	92.79	86.57	94.53	87.54	93.1
GraphSAGE-BERGAMOT	93.81	96.90	81.93	90.96	84.25	89.76	84.68	93.24	88.06	94.03	86.55	92.98
RGCN-BERGAMOT	93.36	96.46	83.73	92.17	84.25	89.76	85.59	92.79	87.56	96.02	86.9	93.44
GAT-BERGAMOT	**94.69**	**97.57**	**85.54**	**94.58**	**86.61**	89.76	84.68	**94.14**	**91.54**	**97.01**	**88.61**	**94.61**
Filtered test set												
Supervised SOTA (Alekseev et al., 2022)	80.95	91.27	**75.32**	87.01	78.46	80.00	66.67	**86.87**	80.37	90.65	76.35	87.16
mSapBERT	80.16	90.48	71.43	87.01	75.38	80.0	62.63	85.86	77.57	91.59	73.43	86.99
mCODER	76.98	89.68	71.43	81.82	76.92	**86.15**	**77.78**	83.84	74.77	89.72	75.58	86.24
GraphSAGE-BERGAMOT	79.37	90.48	67.53	84.42	75.38	84.62	67.68	84.85	77.57	88.79	73.51	86.63
RGCN-BERGAMOT	77.78	88.89	71.43	85.71	75.38	84.62	69.70	83.84	76.64	92.52	74.19	87.12
GAT-BERGAMOT	**82.54**	**92.86**	**75.32**	**90.91**	**80.0**	84.62	67.68	**86.87**	**84.11**	**94.39**	**77.93**	**89.93**

Table 1: Multilingual evaluation results in terms of acc@1 and acc@5 on the English, Spanish, Dutch, French, and German subsets of Mantra corpus. The best results are highlighted in bold.

Model	Base model	Graph	Spanish		Dutch		French		German		avg	
			@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
mSapBERT	mSapBERT	—	71.43	87.01	75.38	80.0	62.63	85.86	77.57	91.59	71.75	86.12
SapBERT text only	mSapBERT	—	67.53	88.31	76.92	83.08	**71.72**	**88.89**	**84.11**	92.52	75.07	88.2
BERGAMOT	mSapBERT	Monolingual	71.43	83.12	80.0	83.08	69.70	85.86	80.37	92.52	75.38	86.15
	XLMR	Multilingual	**75.32**	**90.91**	**80.0**	**84.62**	67.68	86.87	**84.11**	**94.39**	**76.78**	**89.2**

Table 2: Evaluation results of GAT-BERGAMOT models pre-trained on monolingual and multilingual graphs in terms of Acc@1 and Acc@5 on filtered test sets of the Mantra corpus. The best results are highlighted in bold. "SapBERT text only" is the mSapBERT model additionally trained on monolingual positive term pairs with the textual loss only. For each model except the mSapBERT baseline, we trained its two variations using the XLMR and mSapBERT checkpoints for the initialization. The best checkpoint is reported.

negative samples to push textual representations of non-matching concepts far from each other.

Node-node contrastive loss $\mathcal{L}_{node}$ Similarly to term-term loss, we apply MS-loss to pull graph embeddings (g_c^u, g_c^v) representing the same concept c closer pushing away the representations of non-matching ones. Unlike textual encoder, GNN is aware of relations between concepts as it sees neighboring nodes $N(c)$ rather than a single concept c only.

Term-node contrastive loss $\mathcal{L}_{int}$ In batch B, a central concept node c has four representations: e_c^u e_c^v, g_c^u, and g_c^v. While $\mathcal{L}_{sap}$ and $\mathcal{L}_{node}$ optimize unimodal textual and graph models separately, our ultimate goal is to enhance a language model with graph structure knowledge accumulated in graph embeddings. To enable mutual information sharing between a text and a graph encoder we introduce a third contrastive loss which learns from two-modal positive pairs (e_c^u, g_c^u) and (e_c^v, g_c^v). Intuitively, we expect our LM to memorize in-domain knowledge from KG. Similarly to term-term and node-node objectives, we adopt MS-loss and two-modal in-batch hard negative samples. Since $\mathcal{L}_{int}$ makes LM and graph encoders exchange their knowledge, we encourage an LM to be more aware of inter-concept relations.

DGI loss As an additional pretraining objective for a graph encoder, we employ the DGI framework and calculate the DGI loss $\mathcal{L}_{dgi}$. We use a union of all local concept subgraphs from the batch B to form a global batch G_B. Readout function $R(G_B)$ calculates a mean embedding over N central nodes of concept subgraphs. To obtain a corrupted graph $\tilde{G_B}$, we randomly shuffle central nodes across all positive paired samples. The choice of the corruption function encourages our model to distinguish if nodes $N(c)$ are actual neighbors of a central node c. The intuition behind adding the DGI loss is to learn more informative node features. Since DGI captures graph structural information by learning on true and corrupted graphs, we expected it to contribute to LM's structural awareness of concept relations.

For $\mathcal{L}_{sap}, \mathcal{L}_{node}$, and $\mathcal{L}_{int}$ we employ MS-loss (see Eq. 1) for maximum comparability with multilingual versions of SapBERT (Liu et al., 2021b) and CODER (Yuan et al., 2022) as both models utilize this contrastive objective in $\mathcal{L}_{sap}$. Also, MS-loss was experimentally shown to be the most effective contrastive objective to learn from UMLS synonyms (Liu et al., 2021a) among 7 variants. The final training objective used to pre-train the BERGAMOT model is defined as the weighted sum of four loss functions:

$$\mathcal{L} = \mathcal{L}_{sap} + \mathcal{L}_{node} + \mathcal{L}_{int} + \lambda_{dgi}\mathcal{L}_{dgi},$$

where λ_{dgi} is the pre-selected weight of DGI loss. After pretraining on UMLS we discard the graph encoder as in general case there is no graph available during inference. Thus, the result of the pretraining procedure is a BERT-based model enriched with knowledge from the UMLS graph.

4 Experimental evaluation

To train BERGAMOT, we use the UMLS 2020AB release which contains approximately 4.4 million concepts and 15.9 million unique concept names from 215 source vocabularies. The full statistics on pre-training multilingual and monolingual data as well as the number of concept names for languages are shown in Appendix A. We remove all duplicated edges (i.e., edges with matching source and target concepts, and relation type). To ensure each batch includes a sufficient amount of positive samples, we pre-generate synonymous concept name pairs following Liu et al. (2021a). Since random sampling could result in the underrepresentation of languages other than English, we explicitly add (i) monolingual non-English and (ii) cross-lingual positive concept name pairs to each batch. However, the quadratic growth in the number of possible language pairs with respect to the number of languages limits our ability to enrich the training dataset with multilingual data. We discard all cross-lingual pairs which consist of two non-English terms thus forcing all languages to benefit from extensive English knowledge.

To explore how the BERGAMOT's performance is affected by a monolingual low-resource setting, we additionally subsample a monolingual subgraph and generate a monolingual set of positive pairs for Spanish, Dutch, French, and German. We create each dataset by removing all terms that came from a non-target language and all nodes that have no concept name in a target language. For statistics on positive textual pairs count, number of edges and nodes, please see Tab. 8 in Appx B.

For evaluation on the MCN task, we use two cross-lingual benchmarks:

- a medical-crossing benchmark (Alekseev et al., 2022) based on Mantra corpus (Kors et al., 2015) of text units such as scientific abstract titles, drug labels, patent claims mapped to the UML concepts.
- XL-BEL (Liu et al., 2021b), with Wikipedia entities linked to the UMLS.

Mantra corpus covers mentions in English, French, German, Spanish, and Dutch while XL-BEL covers 10 languages. Both benchmarks allow zero-shot evaluation only, i.e., there are no training sets. Additionally, we employ the French Quaero corpus (Névéol et al., 2014) and two datasets in Spanish: (i) CodiEsp-Diagnostico (Miranda-Escalada et al., 2020b) and (ii) CANTEMIST (Miranda-Escalada et al., 2020a). Alekseev et al. (2022) introduced a novel *filtering* procedure which drops test set mentions that are identical to a term from the dictionary. For dataset descriptions and statistics, refer to Appendix B.

To explore the ability of our model to solve diverse non-MCN NLP tasks, we additionally evaluate our model on biomedical Question Answering (QA) and Textual entailment (TE) tasks. We conducted QA experiments on two datasets: (i) PubMedQA (Jin et al., 2019) and BioASQ (Nentidis et al., 2019). The goal of TE task is to determine a logical relationship between two pieces of text: a premise and a hypothesis. For TE experiments, we utilize two corpora: (i) MedNLI (Shivade et al., 2019) and (ii) SciTail (Khot et al., 2018). For

Model	QUAERO-E		QUAERO-M		CodiEsp-D		CANTEMIST	
	@1	@5	@1	@5	@1	@5	@1	@5
mSapBERT	32.43	41.64	39.42	51.6	45.98	61.96	52.82	61.44
mCODER	33.59	40.80	40.30	50.26	35.52	49.14	48.59	58.84
GraphSAGE-BERGAMOT	35.30	41.60	40.94	51.24	46.45	59.55	51.93	**61.54**
RGCN-BERGAMOT	33.59	39.55	40.83	50.26	46.3	62.1	52.33	60.43
GAT-BERGAMOT	**35.39**†	**43.92**	**42.94**†	**53.88**	**48.74**†	**63.61**	**57.41**†	61.38

Table 3: Evaluation results in terms of acc@1 and acc@5 on filtered test sets of the French QUAERO corpus (EMEA and MEDLINE subsets) and the Spanish CodiEsp Diagnostico and CANTEMIST corpora. The best results are highlighted in bold. † denotes statistical significance of GAT-BERGAMOT over both mSapBERT and CODER (Wilcoxon, $\rho < 0.01$).

Model	en		es		de		fi		ru		tr		ko		zh		ja		th		avg	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
mSapBERT	**.787**	**.817**	.578	.634	.331	.371	.188	.22	.36	.422	**.423**	.469	.181	.221	.188	.235	.235	.278	.207	.274	.348	.394
mCODER	.765	.796	**.582**	.626	.331	**.384**	.203	.244	.344	.386	.403	.441	.022	.024	.184	.234	.238	.276	.029	.032	.31	.344
BERGAMOT	.786	.810	**.582**	**.644**	**.332**	.382	**.229**	**.261**	**.373**	**.447**	.419	**.482**	**.185**	**.227**	**.189**	**.244**	**.254**	**.3**	**.215**	**.275**	**.356**	**.407**

Table 4: Evaluation of BERGAMOT model with GAT graph encoder on multilingual XL-BEL benchmark. The benchmark includes entities in English (en), Spanish (es), German (de), Finnish (fi), Russian (ru), Turkish (tr), Korean (ko), Chinese (zh), Japanese (jp), Thai (th) languages. The best results are highlighted in bold.

Model	QA		Entailment	
	PMQA	BioASQ	MedNLI	ST
SapBERT	**63.1**	74.3	82.8	90.2
CODER	**63.1**	73.3	82.4	**90.9**
BERGAMOT	62.3	**76.4**	**83.1**	90.3

Table 5: Evaluation of GAT-BERGAMOT on: (i) Question Answering PubMedQA (PMQA) and BioASQ datasets; (ii) Textual Entailment MedNLI and SciTail (ST) datasets in terms of accuracy.

Model set-up	Mantra		XL-BEL	
	@1	@5	@1	@5
BERGAMOT	**77.93**	**89.93**	35.6	40.7
$\mathcal{L}_{sap}$ only	73.43	86.99	34.8	39.4
-$\mathcal{L}_{dgi}$	76.25	88.52	34.1	39
-$\mathcal{L}_{sap}$	72.64	83.09	30.6	35
-$\mathcal{L}_{node}$	74.55	88.38	**36.1**	**41.2**
-$\mathcal{L}_{int}$	77.41	88.86	34.9	39.5

Table 6: Ablation results of GAT-BERGAMOT. We report the mean acc@1 and acc@5 over all languages present in the Mantra and XL-BEL corpora. $\mathcal{L}_{sap}$ only set-up refers to mSapBERT which trains with a single text-based training objective.

details on QA and TE datasets, see Appendix D. For TE, we adopted the Next Sentence Prediction (NSP) data format from vanilla BERT (Devlin et al., 2019): two sentences for the entailment task are separated with a special separator token, and the model is trained on a classification task.

4.1 Experimental Setup

For evaluation, we employ a ranking approach over embeddings of mentions and potential concepts. After applying an average pooling layer over a BERT-based encoder, the inference task is then reduced to finding the closest concept name representation to the entity mention representation in a joint embedding space. We use the Euclidean distance as the metric. Nearest concept names are chosen as top-k concepts for entities. We evaluate the models in the information retrieval scenario, where the goal is to find top-k concepts for every entity mention in a dictionary of concept names and their identifiers. Following previous works on entity linking (Suominen et al., 2013; Pradhan et al., 2014; Wright et al., 2019; Phan et al., 2019; Liu et al., 2021a), we use the top-k accuracy as the evaluation metric: $\text{Acc@k} = 1$ if the correct UMLS concept unique identifier is retrieved at the rank $\leq k$, otherwise $\text{Acc@k} = 0$. We note that BERGAMOT's graph encoder is discarded during inference and only a BERT-based encoder is used for ranking.

4.2 Results

Medical Concept Normalization Tab. 1 shows the acc@1 and acc@5 metrics for datasets in five languages. BERGAMOT outperformed all models on four sets except the French dataset on both full and filtered test sets. The best results are achieved by GAT-BERGAMOT which consistently ourperforms mSapBERT on all languages proving the effectiveness of three additional training objectives that rely on graph embeddings. Poor performance of GraphSAGE- and RGCN- versions of our model indicate the effectiveness of relative neighbor importance learnt in GAT via the attention mechanism. Interestingly, GAT-BERGAMOT achieves state-of-the-art results on both resource-rich (English and Spanish) and low-resource (German and Dutch) languages. On English, it outperforms a supervised model that is fine-tuned on English data. We must note that due to the small dataset size, GAT-BERGAMOT's improvement over mSapBERT and mCODER is statistically significant (Wilcoxon, $\rho < 0.05$) for the German part only. We provide examples of models' predictions in Appx. G.

In the next series of experiments, we explored how BERGAMOT benefits from both monolingual and multilingual pre-training set-ups. For each of four non-English parts of the Mantra corpus, we trained a monolingual SapBERT and a monolingual GAT-BERGAMOT model. Tab. 2 presents the performance of models pre-trained on multilingual and non-English monolingual graphs and concept names. Based on the results, we can notice that BERGAMOT benefits from the multilingual bi-modal data the most. It appears that training a single multilingual graph-augmented BERGAMOT eliminates a need for further pretraining of monolingual LMs (either SapBERT or BERGAMOT) on UMLS. On average, monolingual BERGAMOT models perform on par with monolingual SapBERT models.

Tab. 3 presents the evaluation results on filtered test sets of the French QUAERO corpus and the Spanish CodiEsp-D and CANTEMIST corpora. GAT-BERGAMOT consistently outperforms two other BERGAMOT models as well as mSapBERT and mCODER baselines pushing the existing state-of-the-art on these corpora. Notably, the GAT-BERGAMOT's improvement over both baselines is statistically significant (Wilcoxon, $\rho < 0.01$). While mCODER outperforms GAT-BERGAMOT on the French part of Mantra, the latter shows a statistically significant improvement on both subsets of QUAERO corpus.

We further investigated the BERGAMOT's performance on the XL-BEL benchmark. Tab. 4 presents the evaluation results. GAT-BERGAMOT outperforms mSapBERT on 8 of 10 languages with an average improvement of 0.8% acc@1. The largest acc@1 growth compared to mSapBERT is observed for low-resource languages ($+4.1\%$ for fi, $+1.9\%$ for ja, $+1.3\%$ for ru). Since mCODER does not support Korean and Thai, it is not directly comparable to GAT-BERGAMOT on full XL-BEL. However, our model outperforms mCODER on 8 remaining languages.

The evaluation results of GAT-BERGAMOT on question answering and textual entailment tasks are presented in Table 5. Despite an introduction of additional entity-oriented graph-based losses, GAT-BERGAMOT did not lose the capability to solve tasks that require text understanding. It performs on par or better than text-only SapBERT. An improvement of 2.1% over SapBERT on PubMedQA indicates a potential of using graph-induced pretraining objectives for tasks involving domain-specific knowledge, such as question answering.

Ablation study To explore the effectiveness of each training objectives, we conducted an ablation analysis by training a GAT modification with each of four individual training objectives removed. Tab. 6 shows the change in performance on XL-BEL and the filtered version of Mantra. Despite losing 1.5% and 1.68% acc@1 on Mantra and XL-BEL, respectively, a model with no $\mathcal{L}_{dgi}$ still shows a decent performance outperforming SapBERT on Mantra. With DGI loss removed, BERGAMOT still outperforms both SapBERT ($+2.82\%$) and CODER ($+0.67\%$). Removal of explicit modality interaction introduced through $\mathcal{L}_{int}$ results in a slight drop of 0.52% and 0.7% acc@1 on Mantra and XL-BEL, respectively. Interestingly, removal of $\mathcal{L}_{node}$ leads to an average 3.38% acc@1 drop on Mantra and an average improvement of 0.5% on XL-BEL.

5 Conclusion and Future Work

We presented BERGAMOT, a graph-augmented architecture with backbone LM designed to learn inter-concept and intra-concept interactions from the multilingual knowledge graph. BERGAMOT outperforms existing language models pre-trained on knowledge triples from UMLS on two concept normalization benchmarks with a diverse set of languages. Since BERGAMOT is currently an encoder model and it is not able to generate texts, an important future research would be to extend BERGAMOT to include language generation capabilities and advance KG-enhanced language generation.

Acknowledgments

The work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of dated November 2, 2021 No. 70-2021-00142. We specially thank Zulfat Miftahutdinov for useful discussions and feedback pointing out issues in a previous iteration of the paper. We thank anonymous reviewers for valuable feedback.

6 Limitations

Large domain-specific graphs. The graph neural networks in BERGAMOT employ a large biomedical knowledge graph, the Unified Medical Language System (UMLS), which contains over 4 million concepts and 15 million concept names. It is important to note that the use of knowledge graphs for different domains with a smaller number of nodes and edges may affect the performance. The size and complexity of the knowledge graph can have a significant impact on the ability of the model to learn and make accurate predictions. Additionally, it is worth noting that while the study focused on the biomedical domain, BERGAMOT could potentially be trained on general-domain knowledge graphs, such as DBPedia or Wikidata.

Only entity-related tasks for evaluation. In this paper, experiments were done on cross-lingual entity linking to see how well BERGAMOT captures knowledge of a multilingual knowledge graph in the biomedical domain. However, to fully assess the ability of BERGAMOT to facilitate automatic knowledge base construction in different languages, it may be beneficial to include additional evaluation tasks such as link prediction or candidate-free taxonomy enrichment. Additionally, probing the knowledge contained in BERGAMOT across multiple languages could be evaluated using the multilingual Language Model Analysis (LAMA) benchmark.

Mantra and XL-BEL. We acknowledge that our choice of Mantra and XL-BEL datasets for MCN is non-exhaustive: we focus on standard benchmarks from previous work, yet both datasets have certain limitations. For example, the Mantra dataset is manually annotated but relatively small, which can make it difficult to measure statistical significance. Similarly, the XL-BEL dataset is based on Wikipedia texts without manual annotation by medical experts, which may limit its relevance to real-world medical applications. Furthermore, both datasets only include entity mentions and terminology in a target language (e.g., French), while BERGAMOT can encode concept names in as many languages as possible for inference to better exploit language connections.

Ethics Statement

One limitation of using external knowledge sources, such as the Unified Medical Language System (UMLS), is that these sources may not be complete for all languages, which can affect the performance of language models and their ability to infer medical concepts from text. Additionally, significant changes to UMLS may require re-training of the language model. BERGAMOT, like any language model, may be subject to representation biases and potentially misleading results, which is a critical concern in the healthcare domain.

All pre-trained models, UMLS, Mantra, XL-BEL, BioASQ, PubMedQA, MedNLI, and SciTail datasets used in this work are publicly available for research purposes.

We honor and support the ACL Code of Ethics.

References

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne Tamang, and Robert Rallo. 2019. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics. *CoRR*, abs/1907.08650.

David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. 2021. Graph-based deep learning for medical diagnosis and analysis: Past, present and future. *Sensors*, 21(14):4758.

Anton Alekseev, Zulfat Miftahutdinov, Elena Tutubalina, Artem Shelmanov, Vladimir Ivanov, Vladimir Kokh, Alexander Nesterov, Manvel Avetisian, Andrei Chertok, and Sergey Nikolenko. 2022. Medical crossing: a cross-lingual evaluation of clinical entity linking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4212–4220, Marseille, France. European Language Resources Association.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In *AMIA 2001, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3-7, 2001*. AMIA.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin M. Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical concept embeddings learned from massive sources of multimodal medical data. pages 295–306.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Shaked Brody, Uri Alon, and Eran Yahav. 2022. How attentive are graph attention networks? In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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.

Yongrui Chen, Huiying Li, Yuncheng Hua, and Guilin Qi. 2020. Formal query building with query structure prediction for complex question answering over knowledge base. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3751–3758. ijcai.org.

Medical CodeBooks. 2016. Icd-10-cm complete code set 2016. 1(1).

Mohamed Dermouche, Vincent Looten, Rémi Flicoteaux, Sylvie Chevret, Julien Velcin, and Namik Taright. 2016. ECSTRA-INSERM @ CLEF ehealth2016-task 2: ICD10 code extraction from death certificates. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, volume 1609 of *CEUR Workshop Proceedings*, pages 61–68. CEUR-WS.org.

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.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, et al. 2018. Best match: new relevance search for pubmed. *PLoS biology*, 16(8):e2005343.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, Ramnik J Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. 2021. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.*, 12(1):3168.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pages 22118–22133. Curran Associates, Inc.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4289–4300.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5189–5197. AAAI Press.

Kuzma Khrabrov, Ilya Shenbin, Alexander Ryabov, Artem Tsypin, Alexander Telepov, Anton Alekseev, Alexander Grishin, Pavel Strashnov, Petr Zhilyaev, Sergey Nikolenko, and Artur Kadurin. 2022. nablaDFT: Large-Scale conformational energy and hamiltonian prediction benchmark and dataset. *Phys. Chem. Chem. Phys.*, 24(42):25853–25863.

Jan A. Kors, Simon Clematide, Saber A. Akhondi, Erik M. van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra GSC. *J. Am. Medical Informatics Assoc.*, 22(5):948–956.

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. *Bioinform.*, 36(4):1234–1240.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574, Online. Association for Computational Linguistics.

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, María Teresa Martín-Valdivia, and Luis Alfonso Ureña López. 2020. Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020*, volume 2664 of *CEUR Workshop Proceedings*, pages 324–334. CEUR-WS.org.

Yinxia Lou, Tao Qian, Fei Li, Junxiang Zhou, Donghong Ji, and Ming Cheng. 2020. Investigating of disease name normalization using neural network and pre-training. *IEEE Access*, 8:85729–85739.

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.

Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, and Elena Tutubalina. 2021. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics*, 37(21):3856–3864.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 2020a. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020*, volume 2664 of *CEUR Workshop Proceedings*, pages 303–323. CEUR-WS.org.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020b. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF ehealth 2020. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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, Minneapolis, USA*, pages 95–100.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the seventh edition of the bioasq challenge. In *Machine Learning and Knowledge Discovery in Databases - International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II*, volume 1168 of *Communications in Computer and Information Science*, pages 553–568. Springer.

John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. 2008. Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for? *Queue*, 6(2):40–53.

NIH. 2021. Nih umls statistics.

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2019. Multi-task character-level attentional networks for medical concept normalization. *Neural Process. Lett.*, 49(3):1239–1256.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P Georgiou, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. 2020. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5):584–595.

Minh C. Phan, Aixin Sun, and Yi Tay. 2019. Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285, Florence, Italy. Association for Computational Linguistics.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland. Association for Computational Linguistics.

Kirk Roberts, Dina Demner-Fushman, and Joseph M Tonning. 2017. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *TAC*.

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. 2021. KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 39–43, Mexico City, Mexico. Association for Computational Linguistics.

Andrey Sakhovskiy and Elena Tutubalina. 2022. Multimodal model with text and drug embeddings for adverse drug reaction classification. *Journal of Biomedical Informatics*, 135:104182.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Chaitanya Shivade et al. 2019. Mednli-a natural language inference dataset for the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics*, pages 1586–1596.

Sarvesh Soni and Kirk Roberts. 2021. An evaluation of two commercial deep learning-based information retrieval systems for COVID-19 literature. *J. Am. Medical Informatics Assoc.*, 28(1):132–137.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy Webber Chapman, Guergana K. Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martínez, and Guido Zuccon. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*, volume 8138 of *Lecture Notes in Computer Science*, pages 212–231. Springer.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA. PMLR.

EV Tutubalina, Z Sh Miftahutdinov, RI Nugmanov, TI Madzhidov, SI Nikolenko, IS Alimova, and AE Tropsha. 2017. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. *Russian Chemical Bulletin*, 66:2180–2189.

E Van Mulligen, Zubair Afzal, Saber A Akhondi, Dang Vo, and Jan A Kors. 2016. Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts. CLEF.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep Graph Infomax. In *International Conference on Learning Representations*.

Daniel Vollmers, Rricha Jalota, Diego Moussallem, Hardik Topiwala, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. 2021. Knowledge graph question answering using graph-pattern isomorphism. In *Further with Knowledge Graphs - Proceedings of the 17th International Conference on Semantic Systems, SEMANTiCS 2017, Amsterdam, The Netherlands, September 6-9, 2021*, volume 53 of *Studies on the Semantic Web*, pages 103–117. IOS Press.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5022–5030. Computer Vision Foundation / IEEE.

World Health Organization. 2013. International classification of diseases for oncology (icd-o).

Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. 2019. Normco: Deep disease normalization for biomedical knowledge base construction. In *Automated Knowledge Base Construction*.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

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.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. 2022. Prix-LM: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5412–5424, Dublin, Ireland. Association for Computational Linguistics.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466.

A Statistics on UMLS graph

Tab. 7 presents statistics on the number of concept names for each language.

The statistics on pre-training graph and number of positive concept name pairs is shown in Tab. 8. The most resource-rich languages in the biomedical domain are English and Spanish, while German and Dutch can be considered as low-resource languages. In particular, the German UMLS is only 1.48% of the UMLS in vocabulary and 3.72% in source counts (NIH).

B Concept Normalization data

Mantra GSC (Kors et al., 2015) is a collection of biomedical text units such as drug labels and patent claims manually cross-labeled by several annotators in five different languages: English, French, German, Spanish, and Dutch. The Mantra terminology is a subset of UMLS with concepts from MeSH, SNOMED-CT, and MedDRA extracted from the UMLS 2020 AA release. Since the Mantra corpus is too small for fine-tuning, all entity mentions are used as test data.

The *CodiEsp* dataset was presented at Clinical Case Coding in Spanish Shared Task at the CLEF 2020 evaluation forum (Miranda-Escalada et al., 2020b). It contains clinical records with entities mapped to the ICD-10 vocabulary (CodeBooks, 2016); we use the CodiEsp Diagnosis (CodiEsp-D) subset and the dictionary provided in *CodiEsp*.

CANTEMIST (CANcer TExt MIning Shared Task on IberLEF 2020 (Miranda-Escalada et al., 2020a) is a manually annotated text corpus of tumor morphology mentions in Spanish mapped to the latest Spanish version of the oncological ontology, which is a part of ICD-O (World Health Organization, 2013); we use the dictionary from (López-Úbeda et al., 2020).

Table 7: UMLS statistics on the number of concept names.

Language	# concept names	percentage
English	11,280,428	70.78%
Spanish	1,589,581	9.97%
French	431,527	2.71%
Portuguese	423,826	2.66%
Japanese	332,099	2.08%
Dutch	293,817	1.84%
Russian	293,031	1.84%
Italian	251,912	1.58%
German	235,736	1.48%
Czech	198,115	1.24%
Korean	147,217	0.92%
Hungarian	109,271	0.69%
Chinese	81,916	0.51%
Norwegian	63,797	0.4%
Polish	51,778	0.32%
Turkish	51,597	0.32%
Estonian	31,183	0.2%
Swedish	30,439	0.19%
Finnish	25,489	0.16%
Croatian	10,035	0.06%
Greek	2,286	0.01%
Latvian	1405	0.01%
Danish	723	0.1%
Basque	695	<0.1%
Hebrew	485	<0.1%

UMLS dataset	# Positive term pairs	# graph nodes	# graph edges
Full UMLS (Multilingual)	30.6M	4.36M	38.81M
Spanish	2.22M	0.51M	11.38M
French	0.44M	0.155M	5.08M
Dutch	0.2M	0.162M	4.62M
German	0.17M	0.116M	4.54M

Table 8: Statistics of the UMLS sets of positive term pairs, nodes, and edges.

The *QUAERO* French Medical Corpus (Névéol et al., 2014) is collection of French entities from two categories: (i) information on marketed drugs from the European Medicines Agency (EMEA, 12,761 entities) and (ii) titles of research articles indexed in the MEDLINE database (8,781 entities). Test set filtering reduces the number of entities to 5,533 and 3,534 for EMEA and MEDLINE parts, respectively. We use the French concept names from the UMLS as a dictionary.

XL-BEL (Liu et al., 2021b) is an automatically annotated dataset with concept mentions from Wikipedia articles. Each mention is mapped to a Wikipedia article using a hyperlink and assigned a CUI based on the article's metadata. Each language is represented with 1,000 mentions resulting in 10,000 mentions in total.

As shown in Table 9, there are 3 evaluation types (Alekseev et al., 2022):

- *Full*: compute metrics on the test set as provided in the dataset itself;
- *Filtered*: remove from the set all entities that are already present in the dictionary (exact match);
- *Filtered*$_{0.2}$: remove from the set all entities where the character-based Levenshtein distance to the nearest neighbor in the dictionary is under 0.2.

The *filtered*$_{0.2}$ set contains two times fewer entities compared to the *filtered* set. Hence, we chose the *full* and *filtered* sets for experiments. All sets, dictionaries, and evaluation scripts are available at `https://github.com/AIRI-Institute/medical_crossing`.

Table 9: Statistics of multilingual corpora user for the evaluation.

Dataset	**# in full set**	**Avg. len in chars**	**% with numerals**	**Filtered set**
Entity mentions				
Mantra (de)	201	17.62	0.50	107
Mantra (en)	452	16.42	1.11	126
Mantra (es)	166	19.67	2.41	65
Mantra (fr)	222	17.64	0.45	99
Mantra (nl)	127	16.06	0.00	65
CANTEMIST	10031	18.73	6.92	3268
CodiEsp-D	10874	15.84	1.05	3449
Concepts				
Mantra (de)	169	-	-	97
Mantra (en)	373	-	-	119
Mantra (es)	147	-	-	69
Mantra (fr)	185	-	-	83
Mantra (nl)	117	-	-	62
CANTEMIST	657	-	-	364
CodiEsp-D	2206	-	-	1142

C Hyperparameter settings

A list of hyperparameter values used to train BERGAMOT models is presented in Table 10. We adopted the parameters of the MS-loss from SapBERT for fair comparison. For the DGI weight λ_{dgi}, we found the best value from the list (0.01, 0.1, 1.0).

Table 10: BERGAMOT's hyperparameter values

Hyperparameter name	**Value**
Graph encoder hidden size	768
Number of neighbors	3
Number of graph encoder layers	3
GAT's number of attention heads	2
Weight λ_{dgi}	0.1
Hard pairs miner's margin	0.2
α in MS loss	2
β in MS loss	50
ϵ in MS loss	0.5
Learning rate	$2 \cdot 10^{-5}$
Multilingual models' batch size	256
Monolingual models' batch size	128
Multilingual models' # of epochs	1
Monolingual models' # of epochs	2

Tab. 10 lists BERGAMOT's hyperparameter values used throughout our experiments.

D Question Asnwering and Textual Entailment data

BioASQ BioASQ (Biomedical Question Answering) is a widely recognized dataset in the biomedical domain, specifically designed for evaluating question answering systems. Following (Gu et al., 2022), we restrict the dataset to yes/no questions. We use the official train/dev/test split where each contains 670/75/140 questions respectively.

PubMedQA Similar to BioASQ, the PubMedQA dataset as well presents questions with limited number of answers. In contrast to the previous dataset, the answers to the questions in PubMedQA are selected from yes, no, or maybe. We use the original train/dev/test split with 450, 50, and 500 questions, respectively.

MedNLI MedNLI (Medical Natural Language Inference) is a specialized dataset designed to facilitate research in natural language inference within the medical and healthcare domain. It consists of pairs of sentences, where each pair comprises a premise and a hypothesis. The premise represents a clinical or biomedical context, while the hypothesis is a medical statement or claim that may or may not logically follow from the premise. Each sentence pair is annotated with one of three labels:

"entailment," indicating that the hypothesis can be logically inferred from the premise; "contradiction," suggesting that the hypothesis contradicts the information in the premise; and "neutral," signifying that there is no logical relationship between the two sentences. The dataset comprises a total of 12,627 sentence pairs in the training set and 1,422 sentence pairs in the testing set.

SciTail The SciTail dataset is similar to the MedNLI dataset was designed for the task of natural language inference. Except that it covers a broader scientific domain. The train part of the corpora contains 24900 sentence pairs and the test part of the corpora contains 2126.

E Models for Comparison

- XLM-R (Roberts et al., 2017) – checkpoint xlm-roberta-base (the Hugging Face Hub)
- SapBERT (Liu et al., 2021a,b) – SapBERT-UMLS-2020AB-all-lang-from-XLMR (the Hugging Face Hub)
- CODER (Yuan et al., 2022) – coder_all (the Hugging Face Hub)
- Supervised SOTA (Alekseev et al., 2022) –the results of *SapBERT+mcn-fz4* are directly obtained from the authors' paper. Three versions were fine-tuned using `SapBERT-UMLS-2020AB-all-lang-from-XLMR`: *SapBERT+mcn-fz4*, *SapBERT+mcn-fz10*, and *SapBERT+mcn*. The first one shows slightly better results in their paper.

F Hardware and software specification

Table 11: Hardware specification of the machine used to conduct our experiments

Device	Specification
CPU	8x Intel Xeon Gold 6152 2.1-3.7 Ghz (2 cores each)
GPU	4x NVIDIA Tesla V100 32GB
RAM	768 Gb
Disk space	5 Tb

To implement, train, and evaluate our models, we used the version 1.11.0 of PyTorch (Paszke et al., 2019) with CUDA 11.3 (Nickolls et al., 2008) support. To implement graph neural networks, we used PyTorch Geometric (Fey and Lenssen, 2019) version 2.0.4. The training of each multilingual BERGAMOT model took up to 20 hours on the machine with the hardware specification described in Table 11.

G Examples of predictions

We provide a few examples of SapBERT's and BERGAMOT's predictions on the English Mantra in Table 12. The results show that SapBERT outperforms BERGAMOT when a true concept in a vocabulary has an extensive textual name. When a golden concept has a short name only, text-only SapBERT fails to produce good entity representations, which is not the case for BERGAMOT. Thus, the provided examples let us suggest that BERGAMOT does not only capture semantic and lexical similarity between entity and concept strings but also has additional intuition of what is a concept from the UMLS knowledge base.

Mention	**SapBERT prediction**	**BERGAMOT prediction**	**True concept**	**Winner**
mixed collagen disease	collagen dis	connective tissue dis mixed	connective tissue dis mixed	BERGAMOT
uveitic	ectatic qualifier value	uveitis disorder	uveitis disorder	BERGAMOT
double uterus	uterus didelphus disorder	doubling of uterus nos disorder	doubling of uterus nos disorder	BERGAMOT
hygiene	ability to perform personal hygiene activity observable entity	personal hygiene finding procedure	ability to perform personal hygiene activity observable entity	SapBERT
functional bowel disorders	functional disorder of intestine disorder	x psychogenic ibs	functional disorder of intestine disorder	SapBERT

Table 12: Sevaral SapBERT's and BERGAMOT's predictions on the English Mantra.