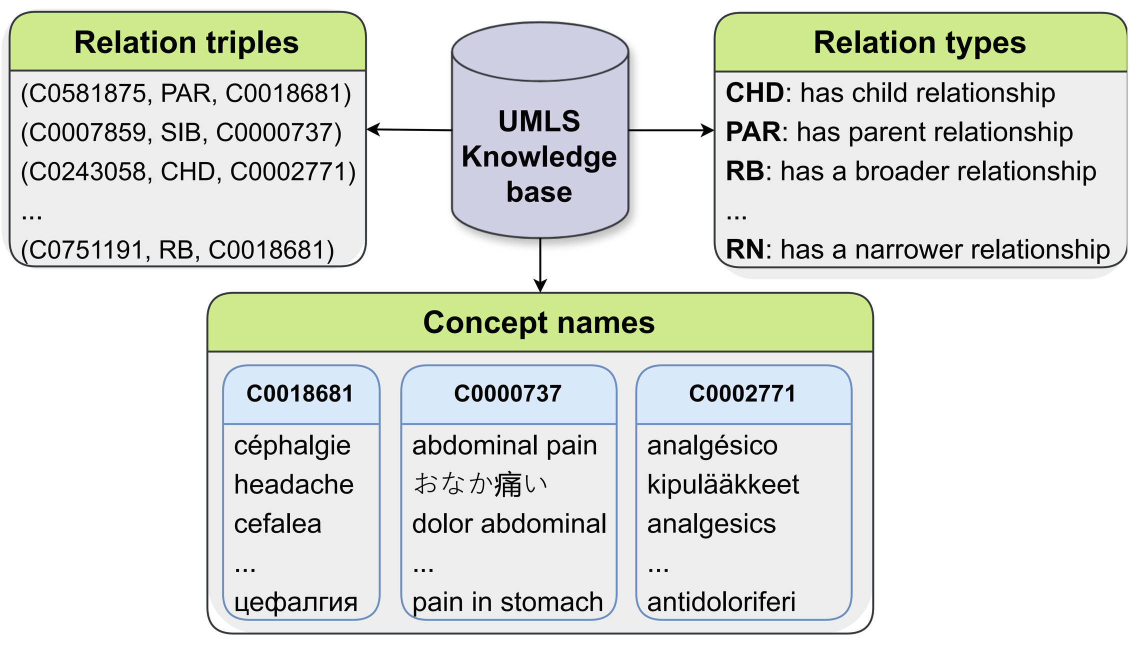
Biomedical Entity Representation with Graph-Augmented Multi-Objective Transformer

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, Elena Tutubalina
{andrey.sakhovskiy, tutubalinaev}@gmail.com

Overview

- **Problem**: multilingual biomedical Language Models (LMs) do not fully utilize domain-specific *Unified Medical Language System* Knowledge Graph (UMLS KG);
- **Idea:** Simultaneously learn textual, graph-based, and intermodal objectives to enrich LM with domain knowledge from KG.
- **Result**: graph-enriched **BERGAMOT** LM achieves SoTA results on multiple zero-shot biomedical entity linking benchmarks and datasets.

UMLS Knowledge Graph



- 4.36M biomedical concepts V;
- 12 relation types ($\mathcal{R}$);
- Knowledge triples $(v, r, u) \in V \times \mathcal{R} \times V$.

Graph encoders

GraphSAGE (mean neighbors pooling):

$$h'_v = \sigma(W^l \cdot [h_v \parallel MEAN(N(v))])$$

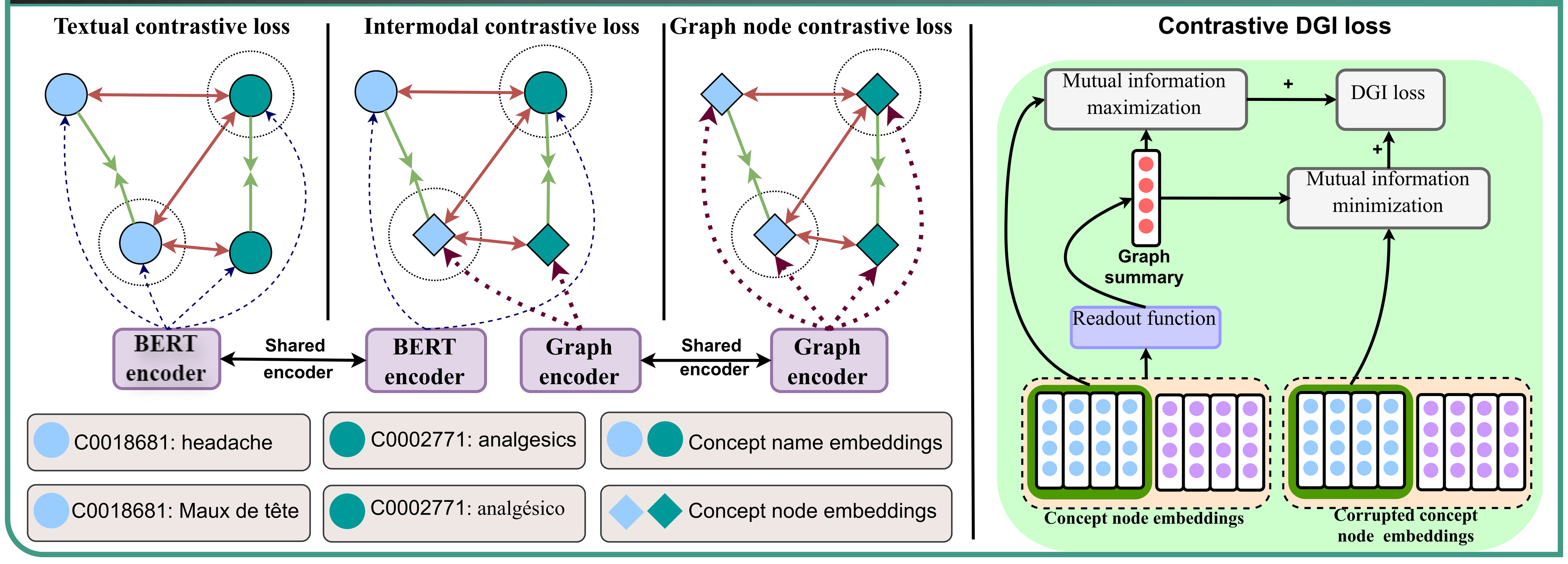
R-GCN (relation r-specific transformations):

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

GAT (Attention-based neighbors weighing):

$$h'_v = \alpha_{v,v} W h_v + \sum_{(r,u)\in N(v)} \alpha_{v,u} W h_u$$

Training objectives



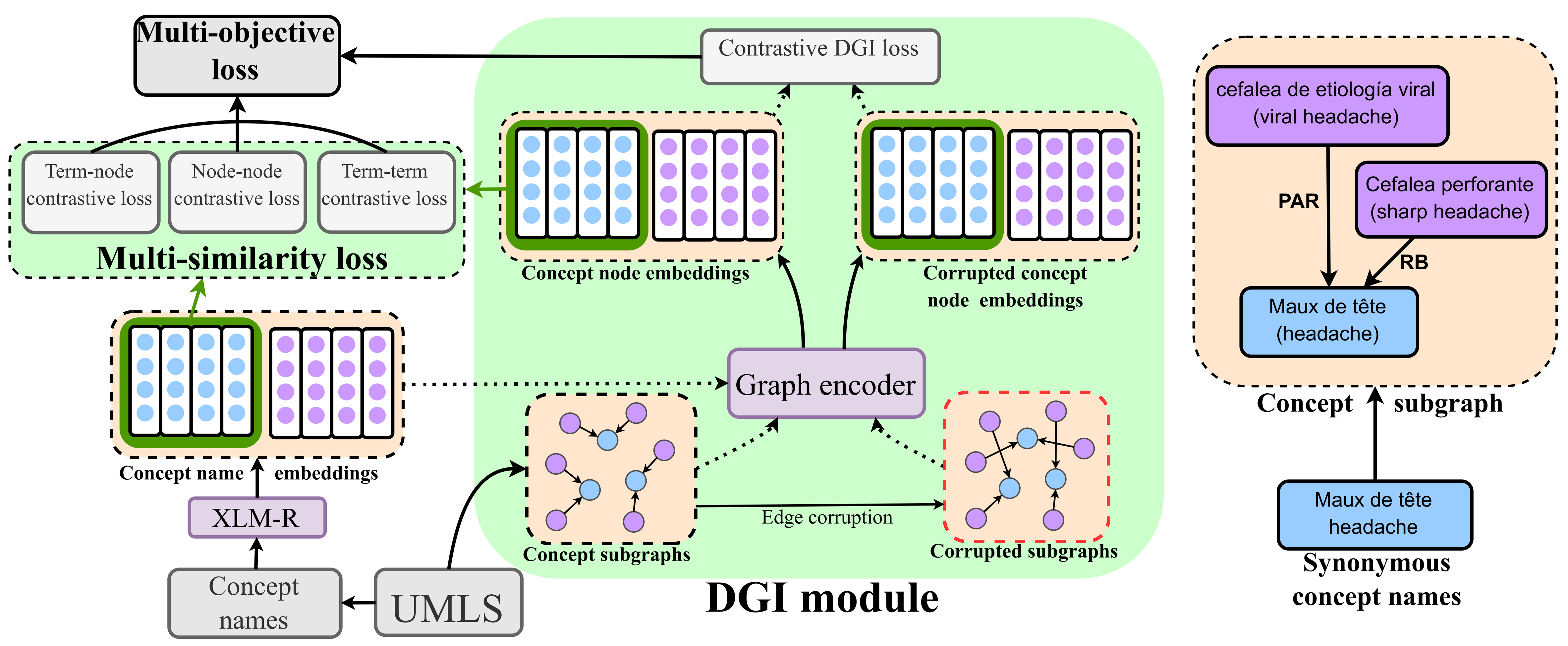
Objectives motivation

- $\mathcal{L}_{sap}$ **(textual)**: clusterizes synonymous concept names;
- $\mathcal{L}_{node}$ **(node level)**: clusterizes concept node representations;
- $\mathcal{L}_{dgi}$ **(subgraph level)**: enriches node embeddings with structural information;
- $\mathcal{L}_{int}$ **(Intermodal)**: aligns two modalities to enrich LM with graph information.

Multi-task training objective:

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

BERGAMOT architecture



Pretraining data

- 30.6M cross-lingual synonym pairs;
- UMLS graph: 4.4M nodes, 38M. edges;
- Batch-level neighbors shuffling corruption;

Evaluation & baselines

- **Entity linking**
  - Mantra GSC and XL-BEL benchmarks;
  - Spanish CodiEsp & CANTEMIST corpora;
  - French QUAERO corpus.
- **Question Answering**: PubMedQA & BioASQ datasets;
- **Textual entailment**: MedNLI & SciTail (ST) datasets;
- **Domain-specific BERT-based baselines**:
  - **SapBERT** [1]: pre-trained on multilingual UMLS concept names only;
  - **CODER** [2]: pre-trained on UMLS concept names and relation triples.

Zero-shot entity linking

Model	QUAERO-E		QUAERO-M		CodiEsp-D		CANTEMIST		Mantra		XL-BEL	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
SapBERT	32.43	41.64	39.42	51.6	45.98	61.96	52.82	61.44	73.43	86.99	34.8	39.4
CODER	33.59	40.80	40.30	50.26	35.52	49.14	48.59	58.84	75.58	86.24	31.0	34.4
GraphSAGE-BERGAMOT	35.30	41.60	40.94	51.24	46.45	59.55	51.93	**61.54**	73.51	86.63	–	–
RGCN-BERGAMOT	33.59	39.55	40.83	50.26	46.3	62.1	52.33	60.43	74.19	87.12	–	–
GAT-BERGAMOT	**35.39**†	**43.92**	**42.94**†	**53.88**	**48.74**†	**63.61**	**57.41**†	61.38	**77.93**	**89.93**	35.6	40.7

Non-linking results

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

Conclusion

- **GAT-BERGAMOT** achieves SOTA biomedical entity linking on two multilingual two Spanish and one French corpora with decent performance on non-linking tasks;
- Our best GAT-BERGAMOT is publicly available[a];
- Biomedical LMs can benefit from knowledge graph modality learnt via external graph neural network.

[a] huggingface.co/andorei/BERGAMOT-multilingual-GAT

References

[1] Fangyu Liu et al. 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.F
[2] Zheng Yuan et al. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. Journal of Biomedical Informatics, 126:103983.
[3] Petar Velickovic et al. 2019. Deep graph infomax. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.