Biomedical Entity Representation with Graph-Augmented Multi-Objective Transformer

Anonymous ACL submission

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. We apply contrastive loss on textual and graph representations to make them less sensitive to surface forms and enable intermodal knowledge exchange between two uni-modal encoders. BERGAMOT achieves state-of-the-art results in zero-shot entity linking without task-specific supervision on three monolingual datasets and Mantra multilingual benchmark. This work is an abridge version of our recent paper (Sakhovskiy et al., 2024).

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), clinical decision making (Sutton et al., 2020; Peiffer-Smadja et al., 2020), and biomedical research (Lee et al., 2016; Tutubalina et al., 2017; Sakhovskiy et al., 2021). These concepts often have multiple nonstandard names, necessitating medical concept normalization (MCN) to map entity mentions to unique identifiers from knowledge bases like the Unified Medical Language System (UMLS) (Bodenreider, 2004), which captures 4 million concepts. Despite the success of pre-trained language models (PLMs) (Lee et al., 2020; Beltagy et al., 2019; Liu et al., 2021b) for biomedical entity representation, challenges remain, particularly regarding bias and synonym recognition (Sung et al., 2021). Existing research (Phan et al., 2019; Miftahutdinov et al., 2021; Liu et al., 2021a; Zhou et al., 2022) integrates knowledge into PLMs by learning from textual triples from Knowledge Bases (KBs) using metric and contrastive learning frameworks. CODER (Yuan et al., 2022) incorporated term-relation similarity to enirch a PLM with KB knowledge. However, this approach learns from individual relation triplets rather than aggregating the whole concept's local neighborhood in the UMLS Knowledge Graph (KG).

2 BERGAMOT

In UMLS, concepts are provided with both multiple multilingual concept names in up to 27 languages and local KG subgraphs. 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 two textual representations (e_c^u, e_c^v) and two graph representations (g_c^u, g_c^v) produced by a PLM and a graph neural networks (GNNs), respectively. The model aims to learn synonym-robust concept representations by learning and aligning two uni-modal encoders on the multilingual UMLS KG. As shown in Fig. 1, the BERGAMOT architecture includes four losses: (i) a **textual term-term contrastive loss** $\mathcal{L}_{sap}$ that seeks to pull textual embeddings (e_c^u, e_c^v) of concept c's synonymous names closer in terms of cosine similarity; (ii) a **node-node contrastive loss** $\mathcal{L}_{node}$ that pulls graph embeddings (g_c^u, g_c^v) representing the same concept c closer; (iii) **DGI** loss $\mathcal{L}_{dgi}$ that encourages a graph encoder GNN to distinguish if nodes $N(c)$ are actual neighbors of a central node c; (iv) an **intermodal contrastive loss** $\mathcal{L}_{int}$ that aligns cross-modal embeddings pairs (e_c^u, g_c^u) enabling mutual information exchange between a textual and a graph encoders. The resulting training loss is obtained as the sum of these four losses: $\mathcal{L} = \mathcal{L}_{sap} + \mathcal{L}_{node} + \mathcal{L}_{int} + \lambda_{dgi}\mathcal{L}_{dgi}$, where

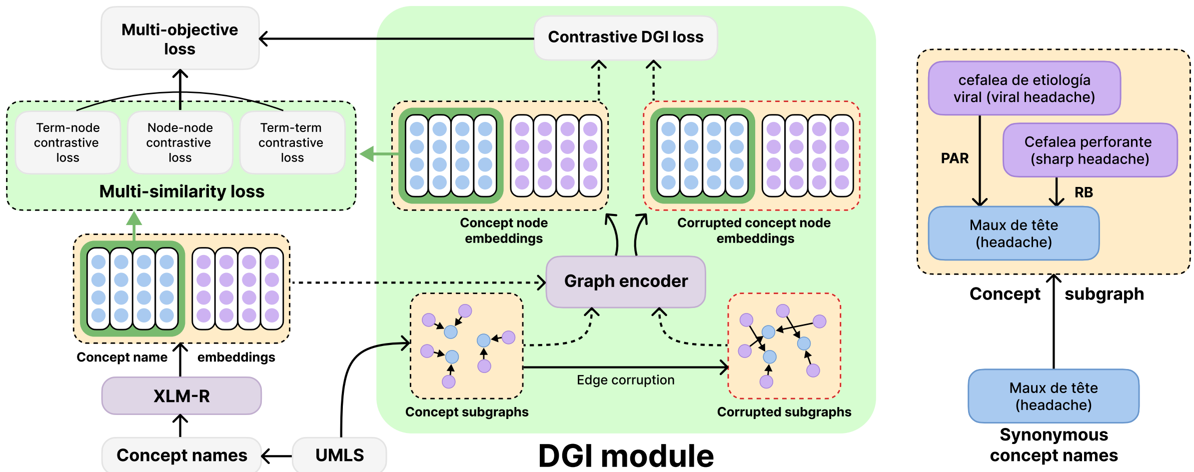

Figure 1: BERGAMOT model's architecture overview. Our model consists of two encoders for text and graph data. Graph encoder uses textual embeddings from BERT as an additional input. The final loss function is a weighted sum of four terms: term-node, node-node, term-term contrastive losses, and local-global 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).

Model	**Mantra**		**QUAERO-E**		**QUAERO-M**		**CodiEsp-D**		**CANTEMIST**	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
mSapBERT	73.43	78.12	32.43	41.64	39.42	51.60	45.98	61.96	52.82	61.44
mCODER	75.58	80.25	33.59	40.80	40.30	50.26	35.52	49.14	48.59	58.84
GraphSAGE-BERGAMOT	73.51	79.00	35.30	41.60	40.94	51.24	46.45	59.55	51.93	**61.54**
RGCN-BERGAMOT	74.19	80.10	33.59	39.55	40.83	50.26	46.30	62.10	52.33	60.43
GAT-BERGAMOT	**77.93**	**83.15**	**35.39**	**43.92**	**42.94**	**53.88**	**48.74**	**63.61**	**57.41**	61.38

Table 1: Multilingual evaluation results in terms of acc@1 and acc@5 on the Mantra benchmark, two subsets of the French QUAERO corpus, and the Spanish CodiEsp-D and CANTEMIST corpora.

λ_{dgi} is the weight of the DGI objective.

3 Experiments

BERGAMOT is trained on the UMLS 2020AB release whith 4.4 million concepts and 15.9 million unique concept names. We experiment with three graph encoders: (i) GraphSAGE (Hamilton et al., 2017), (ii) RGCN (Schlichtkrull et al., 2018), Graph attention network (GAT) (Veličković et al., 2018; Brody et al., 2022) and adopt current state-of-the-art multilingual SapBERT (Liu et al., 2021b) and CODER (Yuan et al., 2022) models as baselines. For evaluation on the entity linking task, we adopt the medical-crossing benchmark (Kors et al., 2015; Alekseev et al., 2022),the French Quaero corpus (Névéol et al., 2014) Spanish CodiEsp-Diagnostico (Miranda-Escalada et al., 2020b) and CANTEMIST (Miranda-Escalada et al., 2020a) corpora with set set filtration Alekseev et al. (2022). We employ a ranking approach over embeddings of mentions and potential concepts with top-k retrieval accuracy as the evaluation metric: Acc@k $= 1$ if the correct UMLS concept is retrieved at rank $\leq k$, otherwise Acc@k $= 0$.

Tab. 1 shows the acc@1 and acc@5 metrics for Mantra benchmark as well as the the French QUAERO corpus and the Spanish CodiEsp-D and CANTEMIST. The best results are achieved by GAT-BERGAMOT which consistently outperforms mSapBERT as well as other two BERGAMOT implementations on all languages proving the effectiveness of three additional training objectives that rely on graph embeddings.

4 Conclusion

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 multiple multilingual concept normalization datasets.

References

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2024. Biomedical entity representation with graph-augmented multi-objective transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4626–4643, Mexico City, Mexico. Association for Computational Linguistics.

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.

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.

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.

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.

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.

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.

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.

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.