Graph-Augmented Typed-Marker Models for Nested Biomedical Relation Extraction*

BioNNE-R Shared Task

Lakshan Karunathilake[1,2,*], Shixiong Zhao[1,2] and Hideaki Takeda[2,1]

[1]The Graduate University for Advanced Studies, SOKENDAI, Shonan Village, Hayama, Kanagawa 240-0193, Japan
[2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Abstract

Extracting relations between nested biomedical entities is an open problem that complicates knowledge extraction: a single span may sit inside, overlap with, or exactly coincide with another, and the relation distribution is steeply long-tailed, leaving rare classes such as ALTERNATIVE_NAME out of reach of trivial baselines. Standard marker-based transformers must infer this nesting and overlap geometry implicitly from token positions, in direct competition with the semantic signal they are meant to learn. We argue that nesting and adjacency are properties of the document, not of the input window, and should be modelled as such. We present a graph-augmented framework for the BioNNE-R shared task at CLEF 2026, spanning 14 relation types across English, Russian, and a bilingual track. A typed-marker pair classifier — built on PubMedBERT, RuBERT, and multilingual BERT — serves as a strong fully fine-tuned baseline. Atop frozen contextual entity embeddings, we construct a heterogeneous document graph whose typed edges encode containment, overlap, sequential adjacency, and same-sentence proximity, with optional UMLS concept-identity links. Three R-GCN layers, an R-GAT variant, and a structurally enriched pair head perform relation classification under inference-time schema masking. On the official test set, the graph model is the top-ranked system in the Russian and bilingual tracks (test macro-F1 of 0.5283 and 0.5015) and ranked second on English (0.5016), with the typed-marker baseline consistently placing one rank below the graph in every track (2nd in Russian and bilingual, 3rd in English). Development-set results corroborate the same ordering, with the graph model reaching macro-F1 of 0.5321 (English), 0.5558 (Russian), and 0.5517 (Bilingual). Explicit nesting edges alone account for up to $+1.5$ nested-F1 points on development, while a direct UMLS pair embedding offers a targeted gradient path for the hardest co-reference relation. The results indicate that handing structural priors to the model explicitly, rather than asking the encoder to rediscover them, is a reliable lever for nested biomedical relation extraction.

Keywords

Biomedical relation extraction, Nested entities, Graph neural networks, Typed markers, UMLS, Multilingual NLP, BioNNE-R

1. Introduction

Biomedical relation extraction (RE) is fundamental to knowledge-base construction, literature mining, and clinical decision support. The BioNNE-R shared task [1], run as part of the BioASQ lab at CLEF 2026 [2], poses a particularly challenging variant: 14 relation types, 8 entity types, nested entity spans (one mention fully or partially inside another), and a multilingual setting spanning English and Russian.

Three properties make this task difficult. First, nested entities require the model to correctly orient relations between a containing and a contained span, which standard windowed-context encoders handle only implicitly. Second, the relation distribution is highly imbalanced: SUBCLASS_OF accounts for roughly a quarter of all gold relations (rising to over 30% on Russian training data) while ALTERNATIVE_NAME, APPLIED_TO, TREATED_USING, and USED_IN each appear in fewer than 3% of training instances, leading to near-zero F1 for rare classes in naive baselines. Third, the cross-lingual setting demands either multilingual encoders or separate monolingual models.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*System description paper for the BioNNE-R multilingual biomedical relation extraction shared task at CLEF 2026.

*Corresponding author.
lakshan@nii.ac.jp (L. Karunathilake); shixiong@nii.ac.jp (S. Zhao); takeda@nii.ac.jp (H. Takeda)

Our approach follows a progressive architecture strategy. We begin with a fully fine-tuned typed-marker pair classifier (`bionne_re`) that encodes structural properties — entity type, nesting status, token distance — as special tokens inside the BERT context window. We then introduce a document-level entity graph model (`bionne_re_graph`) that makes nesting and adjacency relations explicit as typed graph edges, propagating contextual information via R-GCN and R-GAT layers over frozen BERT entity embeddings, and optionally augmented with UMLS-grounded concept-identity features.

The remainder of the paper is organized as follows. Section 2 surveys related work in biomedical relation extraction, typed-marker models, graph-based RE, and UMLS-grounded approaches, and Section 3 describes the task and data. Sections 4 and 5 present each system, Section 6 reports the main results, and Section 7 consolidates the ablation studies covering graph edge typology, UMLS integration, schema masking, and negative sampling. Section 8 provides per-relation and per-language error analysis followed by key findings, and Section 9 concludes. The implementation code for both systems is publicly available.[1]

2. Related Work

Biomedical relation extraction. Relation extraction over biomedical text has been studied for more than a decade, with benchmarks such as BioREL [3] and BioRelEx [4] driving progress on protein–protein and chemical–disease interactions, and the BLUE suite [5] consolidating evaluation across multiple biomedical NLP tasks. Domain-adapted transformer encoders — most prominently PubMedBERT [6] — have become the de facto starting point for English biomedical RE, consistently outperforming general-domain BERT on tasks involving specialised terminology. For Russian biomedical text, the analogous role is played by RuBERT [7] and its domain-adapted variants, while multilingual BERT [8] remains the common fallback in cross-lingual settings. The BioNNE-R shared task builds on the NEREL-BIO corpus introduced by Loukachevitch et al. [9], which extended the Russian NEREL dataset with biomedical nested annotations, and extends the lineage further by combining nested entity annotation with a multilingual (English + Russian) evaluation, exposing limitations of pipelines designed for flat, monolingual settings.

Typed entity markers. The injection of entity-marker tokens around head and tail spans was formalized by Zhong and Chen [10] in the PURE framework, which decoupled entity recognition from relation classification and demonstrated that simple marker-based pair representations could match more elaborate architectures on ACE-style benchmarks. Subsequent work extended the formulation to typed markers (e.g., `[HEAD:DISO]`), encoding entity type directly into the special token vocabulary so that the encoder sees type information at every layer without a separate type embedding. We adopt this typed-marker formulation as our single-encoder baseline; its weakness on nested span configurations — where marker positions become ambiguous — motivates the graph extension presented in this paper.

Graph-based document-level RE. A parallel line of work moves beyond pair-local context by constructing document-level graphs over entity mentions and propagating information through graph neural networks. EoG [11] introduced edge-oriented GNNs for intra-document relations; subsequent work [12, 13] explored mention-level and latent-graph alternatives. The relational backbones we adopt — R-GCN with basis decomposition [14] for typed multi-relational propagation, and GAT [15] for learned per-edge attention — are the standard building blocks of this family. Our contribution is not the use of these layers in isolation but the choice of edge typology: rather than treating all within-document mentions as homogeneous nodes, we encode nesting, overlap, sequential adjacency, and same-sentence proximity as distinct edge types, so that the relational layers can specialist to the structural geometry that BioNNE-R analysis identifies as the dominant signal.

[1] https://github.com/LakshanKarunathilake/BioASQ-challenge/tree/main/Experiments

Table 1
BioNNE-R corpus statistics by language and split. "Unique Ent." is the number of distinct entity mentions after collapsing annotations that share the same character span and entity type; these de-duplicated mentions form the nodes of the document graph in System 2.

Language	Split	Docs	Entities	Unique Ent.	Relations
English	Train	55	3,861	2,208	3,193
English	Dev	50	3,478	1,992	2,967
Russian	Train	717	37,879	19,647	23,773
Russian	Dev	50	3,210	2,170	2,521

UMLS entity linking and co-reference. Bringing knowledge-base structure into neural RE models has a long history; in the biomedical domain UMLS [16] provides a canonical concept inventory that has been exploited by entity linkers including SapBERT [17], which fine-tunes BERT on UMLS synonymy pairs to produce co-reference-aware representations. In place of running a trained linker at scale, we construct a lightweight in-house concept index from the UMLS 2025AA release: SapBERT [17] encodes each canonical surface string into a dense vector, and FAISS [18] provides efficient nearest-neighbour retrieval over the resulting index. The retrieved per-mention CUI assignments feed both a SAME_CONCEPT graph edge and a binary pair-level feature targeting the ALTERNATIVE_NAME relation. This choice reflects an empirical observation about the corpus rather than a general preference, and is one of several design decisions traced explicitly to the data analysis in Section 3.3.

Positioning. Against this background, the present work sits at the intersection of three lines: typed-marker pair classification, document-level relational GNNs, and biomedical knowledge-base grounding. Our specific contribution is to treat the structural properties of the BioNNE-R corpus — the dominance of nested entity configurations in particular — as the design driver, and to show empirically that explicitly typed graph edges over those configurations deliver consistent macro-F1 gains over a strong fully fine-tuned marker baseline across all three language tracks of the shared task.

3. Task and Data

3.1. Task Definition

BioNNE-R is a directed biomedical RE task over 8 entity types and 14 relation types (plus NO_RELATION). Crucially, entity spans may be nested: one span may fully contain another (HEAD_CONTAINS_TAIL / TAIL_CONTAINS_HEAD), partially overlap (PARTIAL_OVERLAP), or share exact boundaries (SAME_SPAN). Only 83 of the theoretical $8 \times 8 \times 14 = 896$ (head_type, tail_type, relation) triples are legal under the task schema; we derive this admissibility set directly from the official NEREL-BIO annotation configuration file.[2] Evaluation uses macro-F1 (equal weight per relation class) and micro-F1 over all 14 types, excluding NO_RELATION.

3.2. Dataset Statistics

We conducted a detailed analysis of the BioNNE-R corpus along five dimensions: split-level volume, document length, entity-type distribution, nesting structure, and relation-type distribution. Table 1 reports the headline figures and Figure 1 summarizes the distributional properties most relevant to model design.

Cross-lingual imbalance. The corpus exhibits a pronounced cross-lingual imbalance, summarized in Table 1: Russian training data contains roughly $13\times$ more documents and almost an order of magnitude

[2] https://github.com/nerel-ds/NEREL-BIO/blob/a7d0d16043850166a4e378190e53968fc4a2ab72/nerel-bio-v1.0/annotation.conf#L4

more annotated entities and relations than its English counterpart. The English training set, in particular, comprises only 55 documents — a regime in which naive full fine-tuning of a large encoder risks severe overfitting, and where every entity mention carries disproportionate supervisory weight. Documents are short paragraph-style passages: median 252 words for English train and 200 words for Russian train, with maxima of 433 and 365 words respectively. This compact length makes within-document modeling tractable inside a standard 384-token encoder window without long-document chunking.

Nesting is the dominant structural pattern. A defining property of BioNNE-R is the prevalence of nested entity mentions: 62.3% of Russian training entities and 63.2% of English training entities are nested inside or overlapping with another mention. The English dev set reaches 67.9%, confirming that nesting is not a corner case but the dominant structural pattern. The effect is even more pronounced at the relation level: between **87% and 90%** of all gold relations in both languages have at least one argument that participates in a nesting or overlap configuration with some entity in the document (87.5% English train, 90.5% English dev, 88.3% Russian dev). We distinguish this endpoint-level reading from the stricter reading in which the two arguments overlap *each other* pairwise, which holds for roughly half of all relations (49.7% English train, 51.5% Russian dev). The typed CONTAINS / INSIDE_OF / PARTIAL_OVERLAP edges connect exactly these pairwise-overlapping mentions, so the ∼50% figure most directly justifies the structural prior, while the 87–90% figure additionally captures relations whose endpoints are reachable from a nested neighbour through graph propagation. Under either reading, the majority of relations live inside the document's nested entity structure — the empirical foundation for the graph-based approach in Section 5. The containment structure is further concentrated in a small number of type-pair combinations, with DISO-in-DISO, ANATOMY-in-ANATOMY, and ANATOMY-in-DISO together accounting for roughly 45% of all nested mentions in Russian training data.

Long-tailed type and relation distributions. Both entity types and relation types are highly skewed (Figure 1). DISO and ANATOMY together account for roughly 50% of all entity mentions across both languages, while DEVICE and INJURY_POISONING each cover less than 2% of training entities. The relation distribution is similarly imbalanced: SUBCLASS_OF alone accounts for 23–32% of relations across splits, while APPLIED_TO, TREATED_USING, USED_IN, and ALTERNATIVE_NAME each represent fewer than 3% of training relations. The relation distribution is also not identical across languages: HAS_CAUSE is the second-most-frequent relation in English but only the third in Russian, where ASSOCIATED_WITH outranks it. Such language-specific differences argue against using a single shared classifier head and motivate per-language model configurations.

3.3. Corpus Concerns and Design Implications

The corpus analysis above directly shaped the architectural decisions described in Sections 4–5. Two clusters of concerns are particularly load-bearing. *Structural* properties of the corpus argue for an explicit graph-based representation: with 87–90% of gold relations holding between nested or overlapping entity pairs, any system that treats pairs as flat tuples in a token window concedes the central signal of the task, motivating the document-level graph model (Section 5) with typed CONTAINS / INSIDE_OF / PARTIAL_OVERLAP edges. Because three type-pair combinations account for nearly half of all nested mentions, basis-decomposed relational layers ($B = 4$ over 16 edge types) provide ample capacity without overfitting on rare combinations. The compact document length (medians under 260 words, maximum 433) further ensures that the entire candidate graph fits inside a standard 384-token encoder window for the typed-marker baseline and a 256-token contextual span window for cached graph embeddings, removing any need for long-document chunking.

Imbalance and rarity concerns drive the training-time and inference-time strategies. The order-of-magnitude gap between Russian and English training data motivates a split response: full fine-tuning works on the 717-document Russian set but risks overfitting on the 55 English documents, so the graph model operates over frozen PubMedBERT embeddings and uses per-language schedules (epochs, dropout, batch composition). The $30\times$ gap between the most and least frequent relation classes is addressed

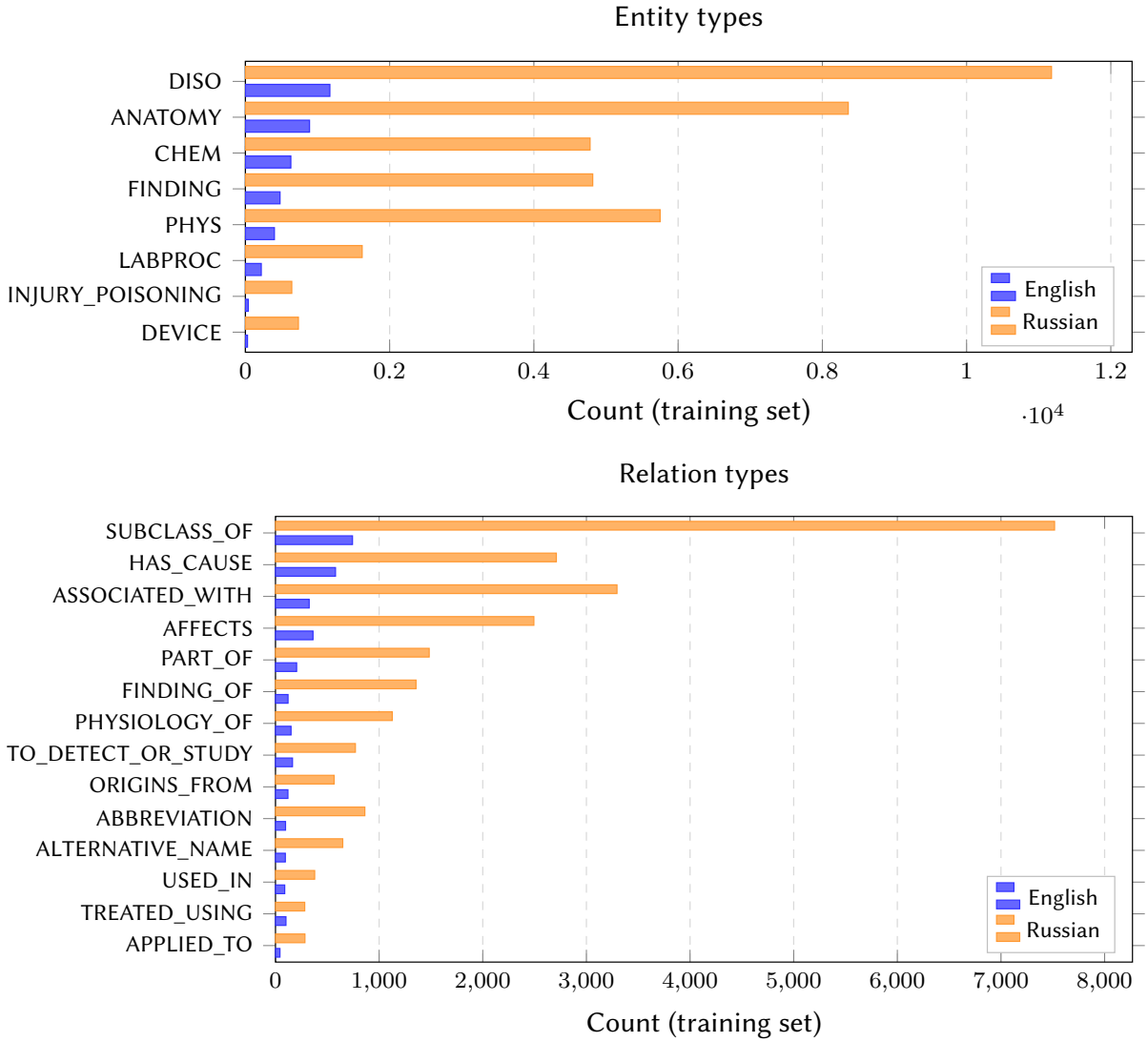

Figure 1: Entity-type (top) and relation-type (bottom) distributions over the English and Russian training sets. Both distributions are heavily long-tailed, with the top two categories accounting for roughly half of all instances. The ordering of mid-tail relations differs between languages (e.g., HAS_CAUSE outranks ASSOCIATED_WITH in English but not in Russian), motivating per-language classifier configurations.

through class-frequency–weighted cross-entropy, label smoothing, and stratified negative sampling that explicitly draws 30% of negatives from nested configurations to prevent trivial NO_RELATION predictions on structurally salient pairs. ALTERNATIVE_NAME, which is simultaneously rare (2.7–3.9%) and a co-reference signal unobservable from local context, additionally receives a dedicated UMLS-grounded pair feature whose construction is described in Section 5 and whose empirical effect is evaluated in Section 7. Finally, because only 83 of the $8 \times 8 \times 14 = 896$ type-pair–relation triples are schema-legal, we apply inference-time schema masking — a zero-cost intervention that eliminates an entire class of errors the classifier would otherwise have to learn to avoid.

3.4. Candidate Generation

All entity pairs within 80 tokens are enumerated as relation candidates. Nested pairs (containment, overlap) are retained regardless of token distance, reflecting the structural priority established above. Schema-illegal (head_type, tail_type, relation) triples — those that do not appear in the NEREL-BIO annotation configuration referenced in Section 3.1 — are filtered out, reducing candidate negatives by ∼25% without loss of recall on positive pairs. During training and development, gold positive pairs are

Table 2
Per-language configuration for `bionne_re`.

Setting	Encoder	Epochs	Dropout	neg ratio	LR (enc/head)
English	PubMedBERT	20	0.15	5	2e-5 / 1e-4
Russian	RuBERT	15	0.10	5	2e-5 / 1e-4
Bilingual	mBERT	20	0.15	5	2e-5 / 1e-4

force-added to guarantee a recall ceiling of 100% for positive instances.

4. System 1: Typed-Marker Pair Classifier

The `bionne_re` system casts relation extraction as pair classification over a fully fine-tuned encoder: it reads a candidate head–tail pair in shared context, forms a pair representation from typed span markers and a structural feature vector, and predicts one of the 14 relations or NO_RELATION under inference-time schema masking. The components are described below; per-language hyper parameters are summarized in Table 2.

Typed markers. Typed markers are inserted around the head and tail spans in the input. For each of the 8 entity types we add four span-delimiter tokens (`[H_TYPE]`, `[/H_TYPE]`, `[T_TYPE]`, `[/T_TYPE]`), for 32 tokens, plus three optional nesting-hint tokens (`[NESTED_HEAD_CONTAINS_TAIL]`, `[NESTED_TAIL_CONTAINS_HEAD]`, `[PARTIAL_OVERLAP]`) that may be prepended when the candidate pair is nested — 35 special tokens in total.

Encoders and context window. We use PubMedBERT [6] for English (`microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`), RuBERT [7] (`DeepPavlov/rubert-base-cased`) for Russian, and multilingual BERT (mBERT) [8] (`bert-base-multilingual-cased`) for the bilingual setting. All encoder layers are fully fine-tuned with a two-group learning-rate schedule (encoder LR $= 2 \times 10^{-5}$, classifier-head LR $= 1 \times 10^{-4}$). The context window is 384 tokens and encodes the head and tail spans in shared context.

Pair representation. The pair representation is built by endpoint pooling over each entity arm: for each arm we concatenate the opening-marker hidden state, the closing-marker hidden state, and a max-pool over the arm's token span (each of dimension H), yielding $3H$ per arm and a $6H$ pair vector, which is projected through a 2-layer MLP.

Structural features. A 24-dimensional numeric structural feature vector is concatenated at the classifier head, comprising:

- *10 continuous features*: character and token distances, sentence distance, number of entities between the arms, head and tail span lengths, length ratio, overlap length, token-overlap ratio, and normalized edit similarity;
- *14 binary features*: head-contains-tail, tail-contains-head, spans-overlap, partial-overlap, shared start boundary, shared end boundary, same-sentence, same surface form, textual containment in each direction, prefix match, suffix match, and an abbreviation-like flag for each arm.

Nesting status, distance bucket, and entity-type-pair identity are supplied separately as learned categorical embeddings rather than in this numeric vector.

Table 3
Best `bionne_re` (typed-marker) results on the challenge development set.

Setting	micro-F1	macro-F1	nested-F1	non-nested-F1
English	0.5592	0.5149	0.5364	0.4208
Russian	0.5760	0.5237	0.5570	0.4001
Bilingual	0.5432	0.5020	0.5411	0.3909

Training objective. The loss is class-weighted cross-entropy with `no_relation_weight=0.75` and label smoothing $= 0.1$. At inference, schema masking zeroes the logits of all schema-illegal (head_type, tail_type, relation) triples.

Table 3 reports the best single-model `bionne_re` results on the challenge development set. Full fine-tuning with the typed-marker approach provides a strong single-encoder baseline. Performance is strongest for Russian, where the larger training set (717 documents) enables effective encoder adaptation. Non-nested F1 is consistently lower than nested F1, reflecting the difficulty of distant-pair relations relative to the structurally constrained nesting cases.

5. System 2: Document-Level Graph RE

Typed markers encode structural properties (entity type, nesting) as token-level signals inside a BERT context window. However, the model must infer span containment from marker positions, which is indirect and competes with the task-specific representation learning. Our hypothesis is that making nesting and adjacency relations explicit as typed graph edges, and propagating context through R-GCN / R-GAT layers, should improve nested-F1 with a complementary inductive bias.

To avoid back-propagating through large BERT encoders during graph training, we pre-compute contextual entity embeddings once and cache them. The pooling mode is `contextual_endpoint_max` — concatenating the first token, the last token, and a max-pool over entity tokens within the full sentence context window (`max_length=256`), yielding a $3H$-dimensional vector (2304-dim for PubMedBERT). This richer pooling outperforms simple mean-pooling on nested entities because the first and last token positions capture span boundaries explicitly.

For each document, a heterogeneous entity graph is built. Nodes are unique entity mentions. Edges fall into the 16 typed categories listed in Table 4, covering identity self-loops, sequential adjacency within a four-position window, the three nesting relations (CONTAINS, INSIDE_OF, PARTIAL_OVERLAP), exact boundary matches (SAME_SPAN), short and long same-sentence co-occurrences, and an optional UMLS-grounded SAME_CONCEPT edge discussed below. The nesting labels of Section 3.1 map onto these directed edge names as HEAD_CONTAINS_TAIL $\rightarrow$ CONTAINS and TAIL_CONTAINS_HEAD $\rightarrow$ INSIDE_OF (the reciprocal edge is added in the opposite direction), while PARTIAL_OVERLAP and SAME_SPAN keep their names; the two namespaces are otherwise identical. Figure 2 illustrates this construction on a real development passage.

Each node is initialised by projecting $\mathrm{concat}(\mathrm{entity\_emb},\ \mathrm{type\_emb}[32],\ \mathrm{lang\_emb}[8])$ through a linear layer to $H = 256$. The graph body comes in two variants. The first (G1–G3) stacks three R-GCN layers with basis decomposition ($B = 4$ bases, $R = 16$ relation types) and residual connections; each layer applies

$$h_i' = \mathrm{GELU}\left(\mathrm{LayerNorm}\left(\sum_r W_r \cdot \sum_{j \in N_r(i)} \frac{h_j}{|N_r(i)|} + W_{\text{self}} \cdot h_i\right)\right) + h_i. \tag{1}$$

Layer normalization is applied to the aggregated messages (post-aggregation, post-norm), the GELU nonlinearity follows the normalization, and the input residual h_i is added afterwards. The second variant (G4) replaces the fixed in-degree normalization with learned per-edge attention: $\alpha_{ij} = \mathrm{softmax}_{\mathrm{per\_dst}}(\mathrm{LeakyReLU}(a_r \cdot [W_r h_j \| h_i]))$, adding only $R \times 2H$ attention parameters per layer

Table 4
Graph edge types used in `bionne_re_graph`. The 16 edge types comprise 1 self-loop, 8 sequential-adjacency edges (forward and backward at offsets 1–4), 2 containment edges (CONTAINS, INSIDE_OF), 1 partial-overlap edge, 1 same-span edge, 2 same-sentence edges (near/far), and 1 UMLS SAME_CONCEPT edge: $1 + 8 + 2 + 1 + 1 + 2 + 1 = 16$.

Edge Type	Count	Description
SELF_LOOP	= N nodes	Identity edge, always present
ADJACENT_FORWARD/BACKWARD_{1–4}	Window	Sequence adjacency within 4 positions
CONTAINS / INSIDE_OF	Nested pairs	Head contains tail / reciprocal
PARTIAL_OVERLAP	Overlap pairs	Partial span overlap
SAME_SPAN	Rare	Identical envelope spans
SAME_SENTENCE_NEAR ($\leq$80 tok)	Config	Co-sentence, short distance
SAME_SENTENCE_FAR ($\leq$160 tok)	Config	Co-sentence, long distance
SAME_CONCEPT (UMLS CUI match)	UMLS only	Co-referent entity mentions

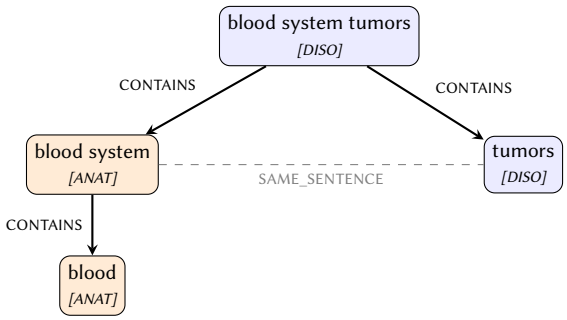

Figure 2: Worked example of the document graph for the development passage *"…pregnant women with blood system tumors"* (doc. 25823274). The nested mention BLOOD SYSTEM TUMORS contains both BLOOD SYSTEM and TUMORS, which are wired as directed CONTAINS edges (reciprocal INSIDE_OF edges omitted for clarity); co-sentence mentions are additionally joined by SAME_SENTENCE edges. For the candidate pair (BLOOD SYSTEM TUMORS, BLOOD SYSTEM) the two arguments are one CONTAINS edge apart, so a single R-GAT layer routes containment context between them and the pair head predicts AFFECTS (correct). A typed-marker model must instead infer the same containment implicitly from overlapping marker offsets in the token stream.

at $H = 256 - 7{,}680$ for the $R = 15$ edge set used by the submitted models, or $8{,}192$ when the UMLS SAME_CONCEPT edge ($R = 16$) is enabled. In both variants, the resulting head and tail node states (h_g, t_g) are combined with the same structural features used in System 1 and fed to the pair classification head. For each candidate pair, the pair MLP receives $\mathrm{concat}(h_g, t_g, |h_g - t_g|, h_g \odot t_g, \mathrm{type\_emb}_h, \mathrm{type\_emb}_t, \mathrm{type\_pair\_emb}, \mathrm{nesting\_emb}, \mathrm{bucket\_emb}, \mathrm{lang\_emb}, \mathrm{numeric\_MLP}(\text{24-dim features}), \mathrm{umls\_pair\_emb})$, projects through 2 hidden layers (width 512), and produces 15-class logits.

UMLS-grounded features. To inject canonical biomedical concept identity into the graph model, we construct an in-house lookup index from the UMLS 2025AA release. The index stores, for each Concept Unique Identifier (CUI), its canonical surface strings filtered to entries of at most 25 characters — a constraint that focuses the index on short, high-precision strings (acronyms, drug names, anatomy terms) while discarding long descriptive phrases whose embeddings are dominated by syntactic rather than semantic structure. Each retained string is encoded with SapBERT [17], a BERT variant fine-tuned on UMLS synonymy pairs that produces co-reference-aware biomedical entity vectors, and the resulting vectors are loaded into a FAISS [18] index supporting efficient inner-product nearest-neighbor search. At training and inference time, every entity mention in the BioNNE-R corpus is independently embedded with the same SapBERT encoder, queried against the FAISS index, and assigned the CUI of its top-1 neighbor. We treat this assignment as a deterministic mention-level feature rather than a probabilistic distribution over concepts.

The retrieved CUIs are used through two complementary mechanisms inside the graph model. First, a SAME_CONCEPT edge type (edge index 15) is added to the heterogeneous graph whenever two mentions in a document share a CUI; the relational layers then propagate co-reference into the corresponding node representations alongside the structural edges. Second, because graph propagation makes co-referent nodes similar but does not signal *why* the pair MLP is seeing similar inputs, we add a binary `pair_umls_same` feature (1 if the candidate head and tail share a CUI, 0 otherwise) as a dedicated `nn.Embedding(2, 8)` concatenated directly into the pair MLP input. This creates a direct gradient path tailored to the ALTERNATIVE_NAME relation, mirroring the role of `nesting_emb` for structural containment. Both mechanisms are active only when `umls.enabled=true` in the per-setting config, which we enable only for English and treat as an ablation (Section 7.2) rather than as part of the submitted systems, which use no UMLS features: SapBERT is trained on English UMLS strings, so Russian mentions cannot be linked reliably without an additional translation or cross-lingual linking step that we leave to future work. The empirical contribution of each UMLS mechanism is evaluated in Section 7.2.

6. Main Results

6.1. Experimental Setup

All experiments use seed 42. Graph training: AdamW LR $= 1 \times 10^{-3}$ (graph body and pair head only; the typed-marker encoder learning rates are given in Section 4), weight decay $= 0.01$, linear warmup (5%), gradient accumulation over 4 documents, `max_grad_norm`$= 1.0$, early stopping (patience$= 5$) on dev macro-F1. Negative sampling: stratified (`nested_fraction`$= 0.30$, `same_sentence`$= 0.30$, `type_pair`$= 0.20$, `nearby`$= 0.10$), resampled each epoch. Loss: weighted cross-entropy (`no_relation_weight`$= 0.75$, `label_smoothing`$= 0.1$). Schema masking at inference only.

Statistical reliability. All reported numbers are single-run results at seed 42; we did not perform multi-seed averaging. The small margins in particular — the English graph-vs.-marker test gap ($+0.006$ macro-F1, Table 5) and the SAME_CONCEPT edge effect (Section 7.2) — fall within plausible single-seed variation and should be read as indicative rather than as statistically established differences.

Compute resources. All models were trained on a single NVIDIA Quadro GV100 (32 GB) GPU. Because the contextual entity embeddings are pre-computed once and cached, each graph configuration trains in a few minutes and well within the 32 GB memory budget; the one-time embedding pre-computation over all documents is the dominant cost of the graph pipeline. The fully fine-tuned typed-marker models of System 1 are more expensive, training in well under an hour per language track on the same hardware.

6.2. Results

Table 5 reports development- and test-set results for our two submitted systems in each track, together with the official organizer baseline. The graph model outperforms the typed-marker baseline on every metric in every track, on both development and test, and both of our systems clear the organizer baseline by a wide margin (roughly $+0.20$ test macro-F1 on average). On the official test set the graph model is the top-ranked submission in the **Russian** and **bilingual** tracks and second in **English**, with the typed-marker baseline one rank below in each. The submitted graph variant per track is the development-best configuration, selected on dev macro-F1 before any test scoring: G4 (R-GAT) for English and bilingual and G2 (R-GCN) for Russian, all sharing the same edge set (self-loop, sequential adjacency ± 1–4, and the nesting/overlap edges CONTAINS, INSIDE_OF, PARTIAL_OVERLAP, SAME_SPAN), a negative-sampling ratio of 10, three graph layers, and basis decomposition ($B = 4$); same-sentence and UMLS edges are disabled in all submitted models.

Table 5
Development- and test-set results on BioNNE-R. For each track we report our two submitted systems (Typed-Marker; Graph, with the submitted variant in parentheses) and the official organizer baseline (an OpenNRE-based `bert-base-multilingual-cased` model with entity markers [1], which reports macro-F1 only). **Bold** marks the better of our two systems for each (track, metric); "Rank" is our official test placement.

		Development		Test		
System	Setting	macro-F1	micro-F1	macro-F1	micro-F1	Rank
Typed-Marker	English	0.5149	0.5592	0.4958	0.5775	3rd
Graph (G4)	English	**0.5321**	**0.5981**	**0.5016**	**0.6030**	2nd
Organizer baseline	English	0.3435	—	0.2825	—	—
Typed-Marker	Russian	0.5237	0.5760	0.5213	0.6237	2nd
Graph (G2)	Russian	**0.5558**	**0.6193**	**0.5283**	**0.6430**	1st
Organizer baseline	Russian	0.4844	—	0.3070	—	—
Typed-Marker	Bilingual	0.5020	0.5432	0.4853	0.5683	2nd
Graph (G4)	Bilingual	**0.5517**	**0.6161**	**0.5015**	**0.6229**	1st
Organizer baseline	Bilingual	0.4911	—	0.3027	—	—

Two patterns temper the headline numbers. First, the test-set margins between graph and typed-marker ($+0.006$, $+0.007$, and $+0.016$ macro-F1 on English, Russian, and bilingual) are smaller than the corresponding development gaps ($+0.017$, $+0.032$, $+0.050$), consistent with some overfitting to development-set idiosyncrasies on the small English split; the ordering between the two systems nonetheless holds across all tracks. Second, the typed-marker baseline attains the highest micro-recall in every track, whereas the graph model trades a little recall for markedly higher precision and a higher overall F1 — the test-time signature of the structural prior, most pronounced on bilingual, where graph micro-precision rises to 0.6207 from 0.4809. The design choices behind these numbers — which edges are active, whether UMLS features and schema masking are applied, and how negatives are sampled — are dissected one factor at a time in Section 7.

7. Ablation Studies

This section consolidates the four ablation experiments that justify the configurations reported in Table 5. Each subsection isolates a single factor while holding the rest of the pipeline fixed, so that the contribution of that factor to the headline results can be read directly.

7.1. Graph Edge Typology

We begin with the central design question for the graph model: which edge types carry the structural signal? Table 6 reports an English experiment ladder that builds the graph incrementally, starting from a pair-only baseline (G0, identity self-loops only) and adding one edge family at a time.

Three observations follow from the ladder. *(i)* The headline $+0.134$ macro-F1 from G1 to G2 combines two interventions — adding nesting/overlap edges *and* raising the negative ratio from 3 to 10 — and the G2† row lets us separate them. Holding the negative ratio fixed at 3, adding nesting and overlap edges (G1 $\rightarrow$ G2†) raises macro-F1 by $+0.052$ ($0.374 \rightarrow 0.426$); raising the negative ratio from 3 to 10 on the same G2 edge set (G2† $\rightarrow$ G2) adds a further $+0.092$ ($0.426 \rightarrow 0.518$). The same-ratio comparison is thus *consistent with* explicit nesting and overlap edges — not sequential adjacency alone — carrying a substantial part of the structural signal for BioNNE-R, alongside a comparably large effect from denser negative sampling (Section 7.4). This is the empirical realization of the corpus observation in Section 3.3 that most relations live between or beside nested entity pairs. *(ii)* G3 (adding same-sentence edges) *hurts* English performance, likely because same-sentence-far edges introduce noise in the small 55-document

Table 6
Graph edge typology ablation on English. All runs use frozen PubMedBERT embeddings with `contextual_endpoint_max` pooling; neg ratio = negative sampling ratio. UMLS features are disabled here to isolate the effect of structural edges. Row G2† repeats the G2 edge set at the G1 negative ratio (3) so that the edge-typology effect can be separated from the negative-sampling effect.

ID	Body	Edges	Neg ratio	EN macro-F1	EN nested-F1
G0	pair_mlp	SELF_LOOP only	3	0.3832	0.4675
G1	R-GCN	G0 + adjacency	3	0.3740	0.4705
G2†	R-GCN	G1 + nesting + overlap	3	0.4257	0.5019
G2	R-GCN	G1 + nesting + overlap	10	0.5176	0.5513
G3	R-GCN	G2 + same-sentence	10	0.4863	0.5356
G4	R-GAT	G2 edges + learned attention	10	0.5321	0.5470
G4b	R-GAT	G3 edges + learned attention	10	0.5092	0.5554

English training set; the same edges are useful for Russian and bilingual settings (Table 5), where the larger training data absorbs the additional structural variance. *(iii)* G4 (R-GAT over G2 edges) recovers and surpasses G2 by $+0.015$ macro-F1, showing that learned attention can re-weight typed edges beneficially, but only when same-sentence noise is excluded — G4b (R-GAT over G3 edges) regresses by -0.023.

7.2. UMLS Integration

The UMLS components described in Section 5 target ALTERNATIVE_NAME, the rarest and structurally most difficult relation (F1 of 0.091 for the typed-marker baseline and 0.115 for graph G4 without UMLS on English). We evaluated two configurations on top of G4: the SAME_CONCEPT edge alone (graph propagation of co-reference) and the full configuration that additionally enables the `pair_umls_same` binary embedding in the pair MLP.

The SAME_CONCEPT edge alone produced macro-F1 changes within run-to-run variance on the development set, so the submitted graph models do not enable this edge by default. The pair-level embedding, by contrast, provides a direct and consistent gradient path for ALTERNATIVE_NAME and is the mechanism we recommend for co-reference-style relations whose signal is non-contextual. This asymmetry — propagation alone being insufficient, but a binary pair feature being effective — is a methodologically useful finding revisited in Section 8. The ablation is restricted to English because our SapBERT-based FAISS index is built over English UMLS strings; extending it to Russian requires either a multilingual biomedical entity encoder or a translation step into English, both of which we leave to future work.

7.3. Schema Masking

Schema masking zeroes the logits of all (head_type, tail_type, relation) triples that are inadmissible under the BioNNE-R schema (only 83 of 896 triples are legal). Disabling masking at inference reduces English macro-F1 by $\sim$0.008–0.012 across all systems, with the largest effect on the graph models where rare-class predictions benefit most from the removal of impossible alternatives. The intervention requires no change to training and incurs no computational cost at inference; we therefore apply it universally in Table 5.

7.4. Negative Sampling

Two negative-sampling factors materially affect graph-model performance: the negative-to-positive ratio and the stratification across hardness buckets. Increasing the ratio from 3 (G0/G1) to 10 (G2–G4) is critical: G2 trained at ratio 3 (row G2† in Table 6) achieves macro-F1 of 0.4257, compared with 0.5176 at ratio 10 — a $+0.092$ effect comparable in magnitude to the edge-typology effect isolated in Section 7.1.

Table 7
English per-relation F1, typed-marker vs. Graph G4.

Relation	Typed-Marker	Graph G4	Δ
PHYSIOLOGY_OF	0.647	0.753	+0.106
FINDING_OF	0.678	0.718	+0.040
ABBREVIATION	0.636	0.670	+0.034
SUBCLASS_OF	0.754	0.749	−0.005
PART_OF	0.598	0.566	−0.032
AFFECTS	0.597	0.613	+0.016
ASSOCIATED_WITH	0.508	0.587	+0.079
HAS_CAUSE	0.441	0.443	+0.002
TO_DETECT_OR_STUDY	0.431	0.463	+0.032
TREATED_USING	0.428	0.537	+0.109
APPLIED_TO	0.371	0.425	+0.054
USED_IN	0.465	0.341	−0.124
ORIGINS_FROM	0.566	0.469	−0.097
ALTERNATIVE_NAME	0.091	0.115	+0.024

Stratified sampling — in particular the 30% nested-negative fraction — prevents the model from trivially predicting NO_RELATION on nested pairs, which would otherwise be the loss-minimizing prediction given the rarity of positive relations among nested candidates. The resample-each-epoch strategy provides fresh negative variety without dataset bloat.

8. Discussion

This section integrates analysis and discussion. We first break down the development-set results along three axes — per-relation F1 (Section 8.1), nested vs. non-nested pairs (Section 8.2), and per-language behavior (Section 8.3) — to surface where the graph model helps and where it does not. We then synthesis these observations into a small number of cross-cutting findings (Section 8.5).

8.1. Per-Relation F1 (English)

Large gains appear for PHYSIOLOGY_OF (+0.106), TREATED_USING (+0.109), and ASSOCIATED_WITH (+0.079) — all relations that frequently occur between entities in adjacency or nesting configurations. The graph's explicit structural encoding is most beneficial here. USED_IN and ORIGINS_FROM regress, likely because these relations depend on broader discourse context that the typed-marker encoder captures through the full 384-token window but the frozen graph embeddings miss.

ALTERNATIVE_NAME remains challenging (F1 = 0.115 for G4 vs 0.091 for typed-marker). The direct UMLS pair embedding evaluated in Section 7.2 provides the most targeted intervention for this class, supplying a direct binary signal for co-reference at the pair MLP level.

8.2. Nested vs. Non-Nested Breakdown

Graph models consistently improve nested-F1 over the typed-marker baseline (EN: +0.015 for G2, bilingual: +0.035 for G4). This validates the core hypothesis: explicit nesting edges propagate containment context more effectively than implicit marker-position encoding. Non-nested F1 is slightly lower for graph models in English (−0.011 for G4 vs typed-marker), likely because the 55-doc English training set provides insufficient signal for the graph body to learn long-distance non-nested patterns from frozen embeddings.

Table 8
Representative English development predictions from the graph model (Graph G4). "nested" marks pairs joined by a CONTAINS/INSIDE_OF edge. Rows 1–2 are correct; rows 3–6 are the dominant error modes.

Head → Tail (types)	Gold	Predicted
blood system tumors → blood system (DISO→ANAT), nested	AFFECTS	AFFECTS
ocular tension → ocular (PHYS→ANAT), nested	PHYSIOLOGY_OF	PHYSIOLOGY_OF
inflammation → acute-phase protein levels (DISO→PHYS)	HAS_CAUSE	NO_RELATION
ultrasonography → ultrasound (LABPROC→LABPROC)	ALTERNATIVE_NAME	NO_RELATION
azathioprine → AZA (CHEM→CHEM)	ALTERNATIVE_NAME	ABBREVIATION
lower extremities → extremities (ANAT→ANAT), nested	PART_OF	SUBCLASS_OF

8.3. Language Analysis

Russian G2 achieves the highest macro-F1 (0.5558) despite fewer graph experiments, because the 717-document Russian training set provides abundant supervision for the R-GCN body even with frozen embeddings. The bilingual G4 model is competitive with Russian G2, suggesting that the graph architecture partially compensates for the multilingual encoder's lower per-language specialization. Russian ALTERNATIVE_NAME F1 is 0.40 for G2 vs 0.39 for typed-marker[3] — the larger training set enables better rare-class learning even without UMLS.

8.4. Qualitative Error Analysis

Table 8 lists representative English development predictions from the graph model that make the aggregate numbers concrete. Two patterns from the quantitative analysis reappear at the example level. First, the graph is reliably correct on *nested* pairs whose relation is licensed by the containment itself (rows 1–2): a disease affecting the anatomy named inside it, or a physiological property of a nested organ, are both one CONTAINS edge apart and predicted with high confidence. Second, the errors concentrate exactly where the corpus statistics predicted they would.

Long-distance causal relations are missed. Row 3 is the single largest error mode in the confusion matrix (HAS_CAUSE ↔ NO_RELATION): a cause and its effect that sit in different clauses share no structural edge, so the frozen graph embeddings — which see only a 256-token span — default to NO_RELATION. This is the example-level face of the non-nested-F1 gap in Section 8.2.

Co-reference remains hard. Rows 4–5 are ALTERNATIVE_NAME pairs, the rarest and lowest-F1 class. ULTRASONOGRAPHY / ULTRASOUND is a synonym pair with no shared substring and no structural edge, and is missed entirely; AZATHIOPRINE / AZA is instead labelled ABBREVIATION — a genuinely confusable alias-vs.-abbreviation distinction. Both are precisely the cases the UMLS `pair_umls_same` feature of Section 5 is designed to reach, and motivate its future extension to the full corpus.

Fine-grained type confusion on same-type nesting. Row 6 shows a subtler failure: the model correctly detects that the nested ANATOMY–ANATOMY pair is related (via the CONTAINS edge) but chooses SUBCLASS_OF over the gold PART_OF — the structural edge signals *that* a relation holds without fully disambiguating *which*, a limitation shared with the type-pair priors in both systems.

8.5. Key Findings

Three observations stand out from the results above. First, typed markers and graph propagation play complementary structural roles rather than competing for the same one. The typed-marker baseline encodes structural information implicitly through marker positions inside a 384-token window, which serves it well on long-distance non-nested relations where broader discourse context matters — the typed-marker remains the strongest model on English non-nested-F1. The graph model makes containment explicit through typed CONTAINS / INSIDE_OF / PARTIAL_OVERLAP edges, which is

[3] From the per-relation breakdown of the Russian development-set evaluation (Graph G2 vs. typed-marker); the full Russian per-relation table is omitted for space and available in the released code.

exactly the geometry that dominates the corpus, and accordingly outperforms the marker baseline on every macro-F1 column and on nested-F1 across all three tracks. Late-fusion ensembling of the two models is a natural next step that we leave to future work.

Second, the choice between R-GCN and R-GAT is setting-specific. R-GAT improves macro-F1 by $+0.015$ on English, where the per-document graph is small ($\sim$20–80 nodes) but richly structured, allowing learned attention to allocate capacity to the few high-signal nesting edges. On Russian, where the graph is denser and degree-based normalisation is already a good aggregation strategy, plain R-GCN edges out R-GAT by 0.002. We interpret R-GAT as most useful when edge density varies sharply across edge types within a document, and R-GCN as preferable when training data is abundant and structure is broadly homogeneous. Third, our UMLS integration confirms a distinction that is easy to overlook: SAME_CONCEPT graph edges propagate co-reference into the node representations but leave the pair MLP without a direct indicator of why two nodes have become similar, whereas a binary `pair_umls_same` embedding threaded into the pair head supplies exactly the gradient path that ALTERNATIVE_NAME requires. The pattern — propagation alone is insufficient when the downstream classifier needs a localized gradient — is likely to generalize to other rare classes whose defining signal is non-contextual.

9. Conclusion

We presented two progressively richer systems for multilingual nested biomedical relation extraction on the BioNNE-R shared task. Beginning with a fully fine-tuned typed-marker pair classifier, we developed a document-level relational graph model combining R-GCN, R-GAT, and UMLS entity-linking.

Our main contributions are: (1) explicit nesting edges in an entity graph lift nested-F1 by up to $+1.5$ points over the typed-marker baseline; (2) R-GAT attention over typed edges provides a further $+0.015$ macro-F1 gain on English; (3) a direct UMLS pair embedding gives a specific gradient path for the rare ALTERNATIVE_NAME relation; and (4) schema-constraint masking is a zero-cost inference improvement applicable to all systems. We also note a negative result worth recording: propagating UMLS-based co-reference through SAME_CONCEPT graph edges (i.e. combining mention aliases via the graph rather than at the pair head) did not yield improvements beyond run-to-run variance, indicating that for rare co-reference relations a direct pair-level signal is more effective than an indirect propagation route.

On the official BioNNE-R test set, the graph approach was our top-ranked submission, taking **first place in the Russian and bilingual tracks** (test macro-F1 of 0.5283 and 0.5015) and **second place on English** (0.5016); the typed-marker baseline was our second submission and placed one rank below the graph in every track (2nd in Russian and bilingual, 3rd in English). Development-set results corroborate this ordering, with the graph model reaching macro-F1 of 0.5321 (English), 0.5558 (Russian), and 0.5517 (bilingual), improving over the typed-marker baseline by $+0.017$, $+0.032$, and $+0.050$ respectively. Future directions include extending UMLS-linked annotations to Russian, end-to-end encoder fine-tuning through the graph, and a late-fusion ensemble of typed-marker and graph model logits. A further avenue is to replace our lightweight SapBERT + FAISS concept lookup with a graph-augmented biomedical entity encoder such as BERGAMOT [19], which jointly trains a text encoder and a graph neural network over the UMLS concept graph. Whereas our SAME_CONCEPT edge and `pair_umls_same` feature inject concept identity as a post-hoc structural signal on top of frozen embeddings, BERGAMOT bakes UMLS graph structure directly into the mention representations; evaluating it as a drop-in replacement for our cached entity embeddings — especially in the cross-lingual Russian setting, which our English-only SapBERT index does not currently reach — is a natural next step.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling check, paraphrasing, and minor translation; and Claude Code for code documentation and code review. After using these tools, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] I. Rozhkov, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-R Task on Nested Biomedical Relation Extraction at CLEF 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[3] S. Lim, J. Kang, BioREL: A large-scale biomedical relation extraction benchmark, Briefings in Bioinformatics (2020).

[4] H. Khachatrian, L. Nersisyan, K. Hambardzumyan, T. Galstyan, A. Hakobyan, A. Arakelyan, A. Rzhetsky, A. Galstyan, BioRelEx 1.0: Biological relation extraction benchmark, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, 2019, pp. 176–190.

[5] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, Proceedings of the BioNLP Workshop (2019) 58–65.

[6] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare 3 (2021) 1–23. doi:10.1145/3458754.

[7] Y. Kuratov, M. Arkhipov, Adaptation of deep bidirectional multilingual transformers for Russian language, arXiv preprint arXiv:1905.07213, 2019. Proceedings of the International Conference "Dialogue 2019".

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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 (NAACL-HLT), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[9] N. Loukachevitch, S. Manandhar, E. Baral, I. Rozhkov, P. Braslavski, V. Ivanov, T. Batura, E. Tutubalina, Nerel-bio: a dataset of biomedical abstracts annotated with nested named entities, Bioinformatics 39 (2023) btad161.

[10] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Association for Computational Linguistics, 2021, pp. 50–61. doi:10.18653/v1/2021.naacl-main.5.

[11] F. Christopoulou, M. Miwa, S. Ananiadou, Connecting the dots: Document-level neural relation extraction with edge-oriented graphs, 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), Association for Computational Linguistics, 2019, pp. 4925–4936. doi:10.18653/v1/D19-1498.

[12] D. Wang, W. Hu, E. Cao, W. Sun, Global-to-local neural networks for document-level relation extraction, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 3711–3721. doi:10.18653/v1/2020.emnlp-main.303.

[13] G. Nan, Z. Guo, I. Sekulić, W. Lu, Reasoning with latent structure refinement for document-level relation extraction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2020, pp. 1546–1557. doi:10.18653/v1/2020.acl-main.141.

[14] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: The Semantic Web - 15th International Conference (ESWC), Springer, 2018, pp. 593–607. doi:10.1007/978-3-319-93417-4_38.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903, 2018. International Conference on Learning Representations (ICLR).

[16] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.

[17] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, 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 (NAACL-HLT), Association for Computational Linguistics, 2021, pp. 4228–4238. doi:10.18653/v1/2021.naacl-main.334.

[18] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2021) 535–547. doi:10.1109/TBDATA.2019.2921572.

[19] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Biomedical entity representation with graph-augmented multi-objective transformer, Findings of the Association for Computational Linguistics: NAACL 2024, 2024.