New Concept Placement in Ontologies

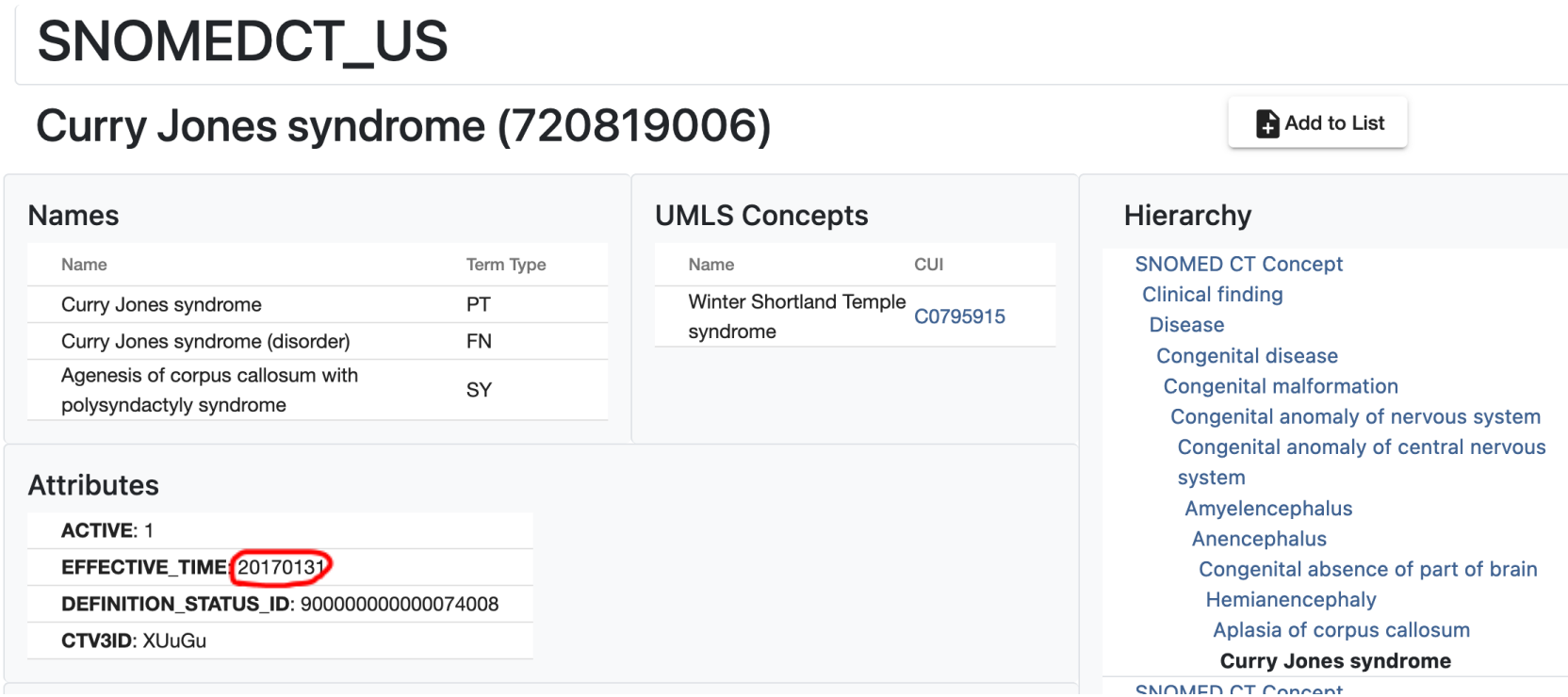
Hang Dong

Lecturer

University of Exeter

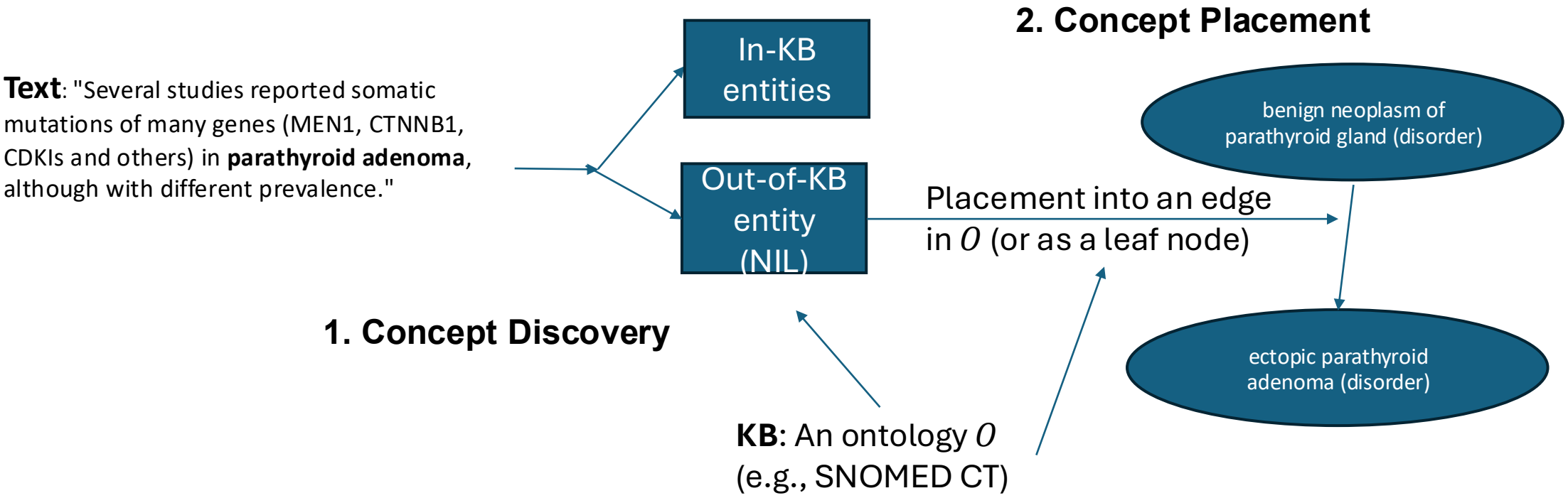
Emerging concepts to enrich Knowledge Bases

- Knowledge Bases (KB) are inherently incomplete.
- New entities emerge over time.
- **We focus on identifying new entities from texts in corpora.**
  - e.g., from scientific publications, and clinical notes.



Concept Discovery and Placement

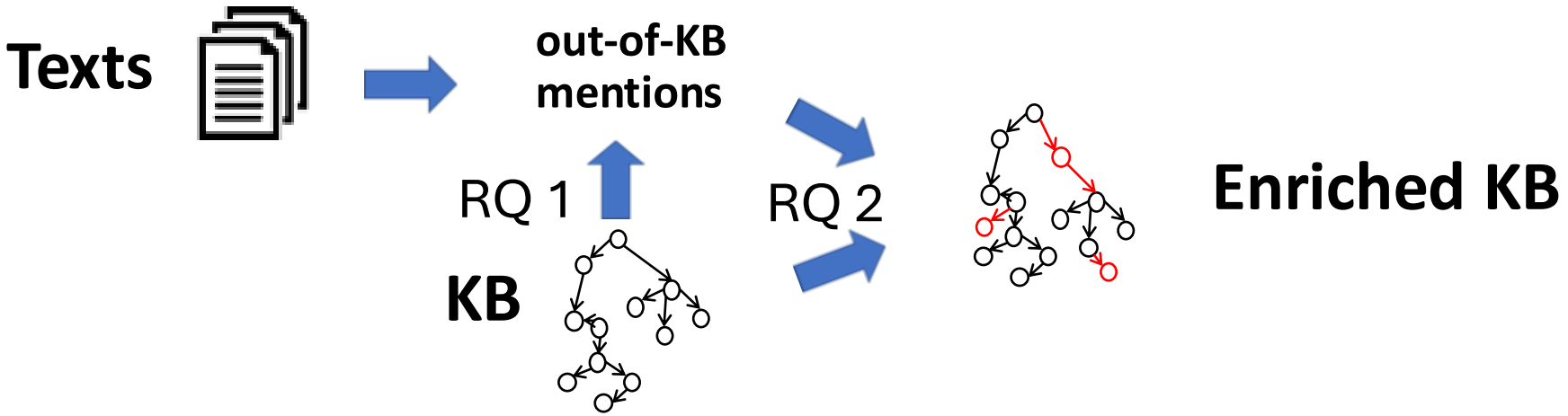
- Concept Discovery and Placement: inputs contextual in-KB and out-of-KB mentions in a corpus and a KB and outputs an enriched KB where each out-of-KB mention is inserted into a *directed edge*, i.e., < parent, child >, of the KB.
- Edge can be <parent, NULL> if it is a leaf node; parent can be a *complex concept*.



Research Questions (RQs)

1. How to identify out-of-KB mention (or NIL mention), i.e., uncaptured by a Knowledge Base, from texts?
   Dong et al., CIKM, 2023a

2. How to insert out-of-KB mentions as entities into a Knowledge Graph?
   Dong et al., CIKM, 2023b, ESWC, 2024



H. Dong, J. Chen, Y. He, Y. Liu, and I. Horrocks, 'Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking', in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, Birmingham, United Kingdom, Oct. 2023, pp. 452–462. doi: 10.1145/3583780.3615036.

Entity Linking for KBs

- Linking a mention (*m*) to a concept (*c*) in a KB (e.g., in UMLS)
  - "The patient has a small bowel obstruction." -> small bowel obstruction (C0235329)
  - "… male with ESRD on HD …" -> Kidney Failure, Chronic (C0022661), Hemodialysis (C0019004)

- Focus on KB, not just on documents
  - Extracting **new concepts** and merge them into an existing KB (Rao, McNamee, Dredze, 2013).
  - NIL entity problem, or *absence* problem.

Rao D, McNamee P, Dredze M. Entity Linking: Finding Extracted Entities in a Knowledge Base. In: Poibeau T, Saggion H, Piskorski J, Yangarber R, editors. Multi-source, Multilingual Information Extraction and Summarization. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 93–115. (Theory and Applications of Natural Language Processing).

Problem Formulation

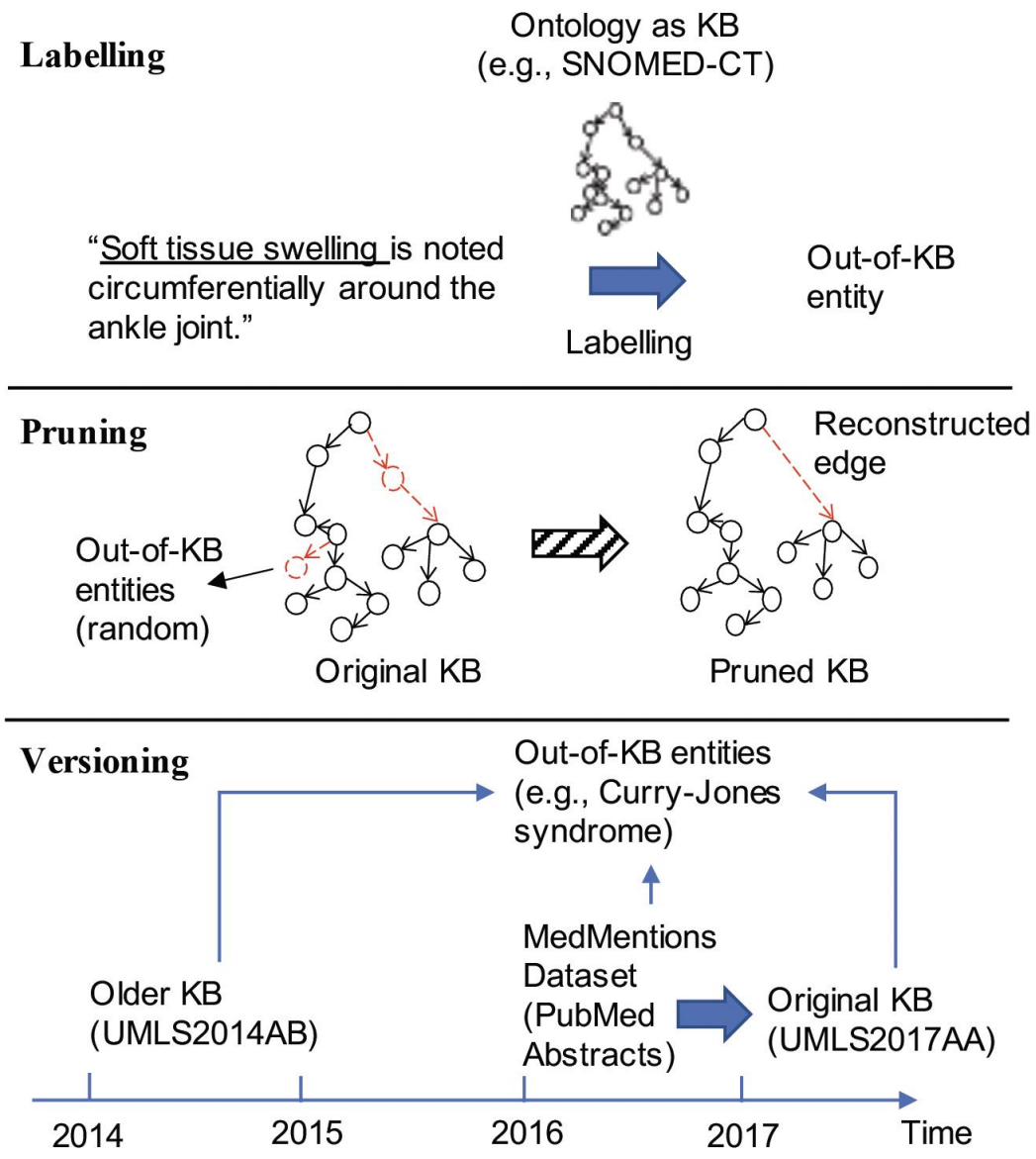
- **Out-of-KB mention discovery:** an extended entity linking task

Developing a function f such that each contextual mention (i.e., each in a context window in a document) d_m in D_M is mapped to an entity,

$$f(d_m) \in E \cup \{\text{NIL}\},$$

where NIL is a special entity denoting that there are no entities in E that can be matched with d_m.

Data Creation Strategies



Strategies for NIL-enhanced EL dataset construction: Manual Labelling, Ontology Pruning, and Ontology Versioning. (The specified UMLS ontologies are the subset linked to SNOMED CT under the semantic type of T047, "Disease or Syndrome".)

ShARe/CLEF 2013

Linked to

$\text{UMLS}_{disease\ group} \cap$ SNOMED-CT

MedMention

Linked to

$\text{UMLS 2017AA}_{T047} \cap$ SNOMED-CT

(T047: Disease or Syndrome)

NILK

Linked to

WikiData

Out-of-KB Labelled Datasets

	ShARe/CLEF 2013 train / test	MM-pruned-0.1 train / valid / test	MM-pruned-0.2 train / valid / test	MM-2014AB train / valid / test	NILK-sample train / valid / test
Out-of-KB mention creation	Manual Labelling	KB Pruning	KB Pruning	KB Versioning	KB Versioning
Domain or document type	Clinical notes	PubMed abstracts	PubMed abstracts	PubMed abstracts	Wikipedia articles
# of docs	199 / 99	2,635 / 878 / 879	2,635 / 878 / 879	2,635 / 878 / 879	81,334 / 10,587 / 10,528
# of mentions	5,816 / 5,351	6,201 / 2,121 / 2,000	6,201 / 2,121 / 2,000	6,181 / 2,112 / 1,988	86,379 / 10,688 / 10,613
# of out-of-KB mentions	1,639 / 1,723	456 / 201 / 155	1,031 / 405 / 415	307 / 82 / 103	1,312 / 154 / 158
% of out-of-KB mentions	28.2% / 32.3%	7.3% / 9.5% / 7.8%	16.6% / 19.1% / 20.8%	5.0% / 3.9% / 5.2%	1.5% / 1.4% / 1.5%
# of in-KB mentions	4,109 / 3,604	5,745 / 1,920 / 1,845	5,170 / 1,716 / 1,585	5,874 / 2,030 / 1,885	85,067 / 10,534 / 10,455
% mentions w/ zero-shot entities	- / 11.6%	- / 24.8% / 25.3%	- / 23.2% / 22.0%	- / 27.8% / 28.0%	- / 98.6% / 98.5%
# of entities in KB	88,150	35,392	31,460	35,398	79,412
# of entities & synonyms in KB	288,490	126,188	112,097	124,132	121,191
% of entities having synonyms	99.0%	99.4%	99.4%	99.3%	25.5%
Ave # syns per entity (having syns)	2.3 (2.3)	2.6 (2.6)	2.6 (2.6)	2.5 (2.5)	0.5 (2.1)

Table 1: Statistics of out-of-KB linking datasets. Slashes separate the statistics of training, validation, and testing sets. "MM" denotes MedMentions; pruned-0.1 and pruned-0.2 denote the percentage of pruned concepts in the ontology; "2014AB" refers to the older version of the UMLS applied; "NILK-sample" denotes the random sample of the NILK dataset that enriches WikiData.

Existing methods for out-of-KB mention discovery

- Threshold-based (Ji et al., 2020; Chen et al., 2021):
  - $S(d_m, E) \Longrightarrow m \leftarrow \mathrm{NIL}$, S is entity linking system; d_m, mention m in context; E, all in-KB entities; NIL, out-of-KB entity.
  - $\forall_{e \in E} s(d_m, e) < Th \Longrightarrow m \leftarrow \mathrm{NIL}$. Th is a threshold.

- Classification:
  - With features or embeddings related to out-of-KB entities

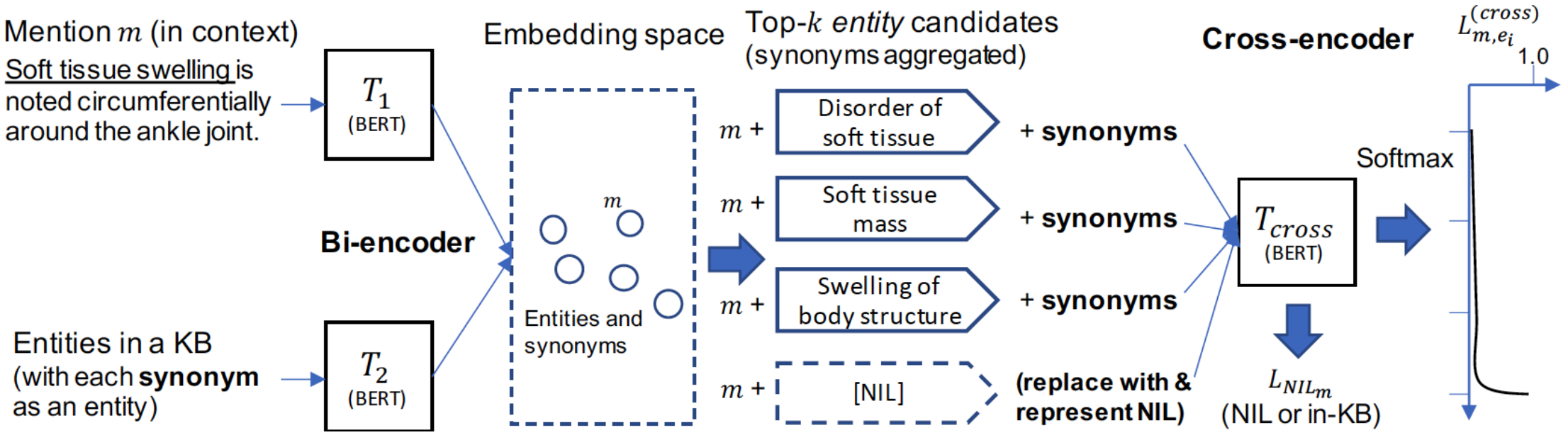
  (i) Adding *NIL* as a candidate with in-KB entities (Rao et al., 2013)

  (ii) Binary (or three-class) classification: Mention as in-KB, out-of-KB, (or not relevant) (Gu et al., 2021)

NIL entity representation

- How to represent an out-of-KB entity (to embeddings)?
  - Threshold-based – no need to represent it (Ji et al., 2020; Chen et al., 2021).
  - Classification-based – mostly do not represent NILs, but classify into NILs based on features or embeddings, using in-KB entities (Rao et al., 2013)
  - Using extra texts/corpora that contain the out-of-KB entity (Hoffart et al., 2014; Ran et al., 2022)
  - Clustering-based: need to detect out-of-KB mentions first,
    - Form clusters of mentions into entity (Kassner et al., 2022; Agarwal et al., 2021) and get embeddings of the cluster (Kassner et al., 2022).
    - Usually based on the above methods, threshold or classification.
- Fewer studies have explored NIL entity representation with neural LMs (Kassner et al., 2022; Agarwal et al., 2021).

Methods: NIL Representation & Classification, with Synonym Enhancement

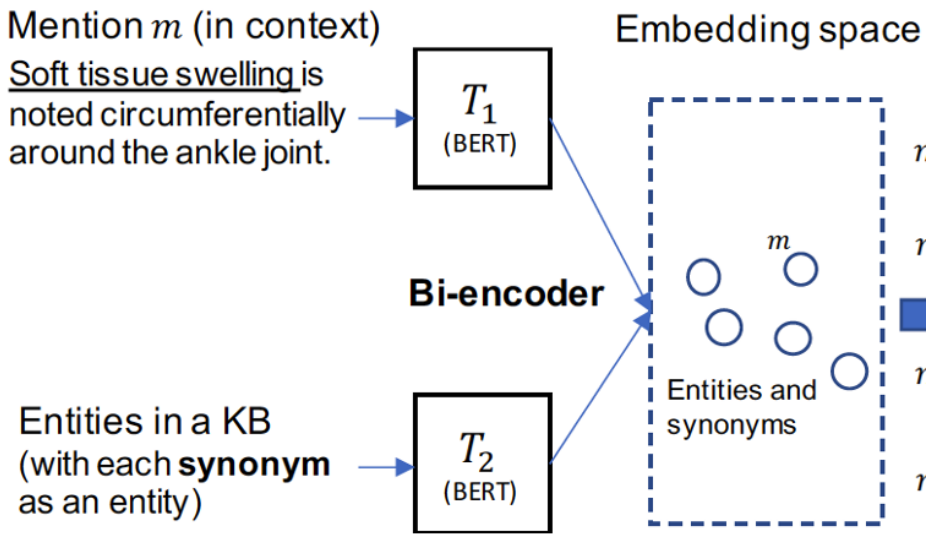


Bi-encoder (candidate generation)

Input mention d_m: **[CLS] ctxt_left [M_s] mention [M_e] ctxt_right [SEP]**
Input entity OR synonym d_e: [CLS] title [ENT] description [SEP]

- Train Bi-encoder

$$v_m = \text{red}(BERT_m(d_m))$$
$$v_e = \text{red}(BERT_e(d_e))$$
$$s(m, e) = v_m \cdot v_e$$
$$L_{m_i,e_i} = \sum_{j \neq i} [\alpha - s(m_i, e_i) + s(m_i, e_j)]_+$$



- Generate top-*k* entities for each mention
  - Aggregating over the synonyms into top-*k* entities (using the score of the most similar synonym of each entity)

Cross-encoder (candidate ranking)

```
Input mention (m):
        [CLS] ctxt_left [M_s] mention [M_e] ctxt_right

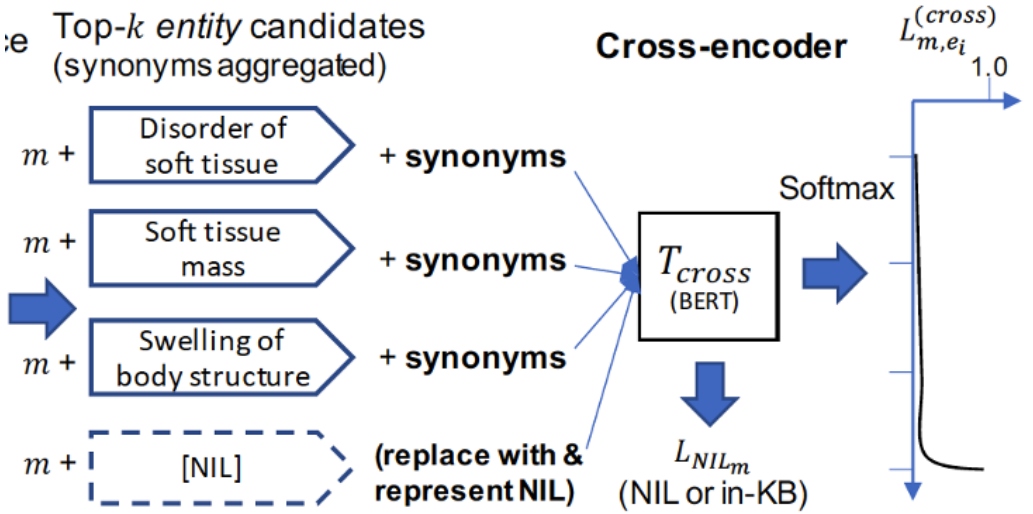
Input entity+synonym (es):
        title [ENT] syn_1 [SYN] syn_2 ... syn_n [SYN] description [SEP]

Input d_{m,es} <- Concat(Input mention, Input entity+synonym)
```

- Train Cross-encoder

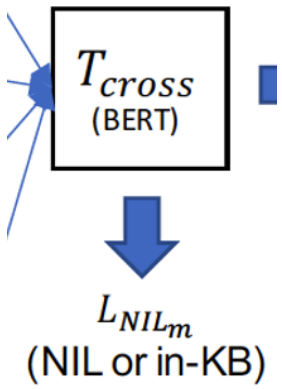
$$s_{m,e}^{(cross)} = v_{m,e}^{(cross)} w; \quad v_{m,e}^{(cross)} = \mathrm{red}(BERT_{cross}(d_{m,e}))$$

$$L_{m,e_i}^{(cross)} = -\log\left(\frac{\exp(s_{m,e_i}^{(cross)})}{\sum_{j \in E_{cand}} \exp(s_{m,e_j}^{(cross)})}\right)$$



Joint training for NIL verification

A binary classification to verify NIL, conditioned on mention representation



$$s_{isNIL_m} = \sigma(v_m^{(cross)} w_{NIL});\ v_m^{(cross)} = \mathrm{red}(BERT_{cross}(d_m))$$

$$L_{NIL_m} = -y_i \log(s_{isNIL_m}) + (1 - y_i)\log(1 - s_{isNIL_m})$$

$$L_m = L_{m,e_i}^{(cross)} + \lambda_{NIL} * L_{NIL_m}$$

Inspired by Gu, Y., Qu, X., Wang, Z., Huai, B., Yuan, N. J., & Gui, X. (2021, May). Read, retrospect, select: An MRC framework to short text entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 14, pp. 12920-12928).

Overall and out-of-KB results

Main contributors of BLINKout's performance:
(i) Fine-tuned NIL entity representation.
(ii) Synonym enhancement.
(iii) Domain-specific BERT, e.g., SapBERT.

	ShARe/CLEF 2013				MM-pruned-0.1				MM-pruned-0.2				MM-2014AB				NILK-sample			
	A	P_o	R_o	$F1_o$	A	P_o	R_o	$F1_o$	A	P_o	R_o	$F1_o$	A	P_o	R_o	$F1_o$	A	P_o	R_o	$F1_o$
Rule-based (Sieve-based)	90.0	84.7	92.6	88.5	**83.8**	36.3	**94.8**	52.5	**85.3**	63.4	**97.1**	76.7	**82.9**	26.6	**98.1**	41.9	72.1	6.6	**86.7**	12.2
Ft-based Classifier	36.0	49.1	71.6	58.2	19.1	11.2	69.7	19.2	25.8	27.5	63.9	38.4	20.1	7.2	37.9	12.1	24.2	1.6	70.3	3.0
Ft-based BLINKout $L_{NIL}^{(Ft)}$	60.9	61.4	**95.0**	74.6	42.3	14.0	92.3	24.2	48.0	39.3	95.9	55.7	43.7	8.9	85.4	16.1	45.2	1.8	61.4	3.5
Th-based BM25+BERT+syn	69.5	64.9	55.8	60.0	72.3	22.8	64.5	33.7	70.2	46.2	60.7	52.5	70.6	9.7	26.2	14.1	77.4	7.0	39.2	11.9
Th-based BM25+SapBERT+syn	76.2	71.3	64.7	67.8	75.6	28.6	84.7	42.2	74.7	51.3	72.5	60.1	71.4	16.9	81.6	28.0	75.8	7.6	70.3	13.8
BLINK (BERT)	53.0	-	-	-	67.0	-	-	-	57.2	-	-	-	62.4	-	-	-	88.0	-	-	-
+ syn	58.1	-	-	-	70.4	-	-	-	60.2	-	-	-	68.8	-	-	-	87.8	-	-	-
Th-based BLINK	72.4	65.4	67.7	66.5	68.1	21.9	42.6	28.9	71.1	51.6	77.6	62.0	65.7	21.7	74.8	33.6	88.1	17.8	49.4	26.2
+ syn	80.1	74.4	76.2	75.3	72.7	25.5	54.8	34.8	69.5	47.7	60.5	53.4	71.1	22.2	48.5	30.5	87.8	16.3	44.3	23.8
NIL-rep-based BLINK	74.8	67.6	89.2	76.9	69.7	24.8	47.1	32.5	71.9	56.9	74.9	64.7	65.7	19.1	80.6	30.9	87.9	16.3	24.7	19.6
+ syn	80.3	79.8	73.5	76.5	72.6	27.8	40.7	33.0	73.8	54.4	77.1	61.7	71.4	18.6	74.8	29.8	87.9	16.2	55.7	25.0
BLINKout (- syn, BERT)	77.7	74.1	91.2	81.8	62.3	18.4	62.6	28.5	68.8	51.1	80.7	62.6	68.2	25.2	78.6	38.1	88.1	24.3	49.4	32.6
BLINKout (BERT)	86.0	85.3	87.9	86.6	70.5	33.1	58.7	42.3	75.2	55.3	77.8	64.7	69.9	33.7	84.5	48.2	87.9	21.4	72.2	33.0
BLINKout (SapBERT)	89.9	90.0	89.8	89.9	81.8	53.5	54.8	54.1	81.4	**72.6**	81.0	76.5	81.5	57.3	79.6	66.7	86.6	22.3	69.0	33.7
+ k and NIL rep tuned*	90.8	**91.4**	90.3	**90.9**	83.3	**58.4**	65.2	**61.6**	84.8	**72.6**	88.7	**79.8**	**82.9**	**86.2**	72.8	**79.0**	**90.7**	33.0	70.3	44.9
+ L_{NIL}	**91.2**	88.5	92.5	90.5	83.1	47.0	75.5	57.9	83.0	65.2	91.6	76.2	82.5	59.3	83.5	69.4	90.4	**37.1**	67.1	**47.8**

Table 2: Comparison results for Entity Linking with *out-of-KB* mentions (Precision, Recall, F_1 scores w.r.t. out-of-KB mentions and *overall* Accuracy w.r.t. both in- and out-of-KB mentions). *The parameters k and NIL rep are tuned using BLINKout (SapBERT) for ShARe/CLEF 2013 and MM datasets and BLINKout (BERT) for NILK-sample.

Qualitative analysis

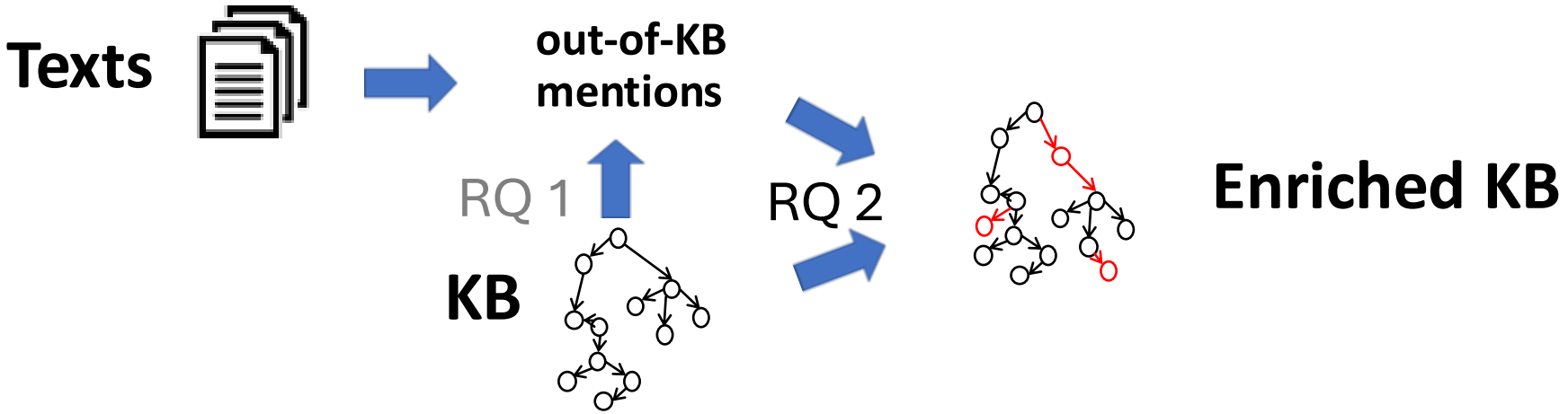
Data w/ row ID (starts from 0)	**Mention** in context	Gold	Th-based BLINK+syn	NIL-rep-based BLINK+syn	BLINKout (tuned)	BLINKout (tuned)+L_{NIL}
ShARe/ CLEF 2013 test-2353	Discharge Diagnosis: **PEA arrest** of unclear etiology	NIL	Electromechanical dissociation (C0340861) 1.000>Th_{cross} (0.95)	Electromechanical dissociation (C0340861) 1.000 (where NIL 0.000, 2nd)	Electromechanical dissociation (C0340861) 0.888 (where NIL 0.001, 2nd)	NIL 0.718
ShARe/ CLEF 2013 test-5083	Low dose beta blockade **decreased HR** to 95 in sinus and synthroid had been restarted.	Bradycardia (C0428977)	NIL (Cardiac murmur, intensity grade I/VI (C0232249) 0.191<Th_{cross} (0.95)) (Gold not in top-k)	NIL 0.973 (Gold not in top-k)	Bradycardia (C0428977) 0.989	NIL 0.947 (where Bradycardia (C0428977) 0.053, 2nd)
NILK-sample test-68	The ancient cities of Friesland are shown below: Dutch West Frisian Charter granted ... Workum Warkum 1399 **Bolsward** Boalsert 1455 Harlingen Harns 1234 Franeker Frjentsjer 1374	NIL	NIL (Ferdinand Bol (Q374039) 0.403<Th_{cross} (0.45))	Ferdinand Bol (Q374039) 0.480 (where NIL 0.276, 2nd)	NIL 0.377	NIL 0.191
NILK-sample test-2054	... **70720 Davidskillman** Named in honor of David R. Skillman (b. 1945) for his decades-long contributions to asteroid searching, stellar binary star systems and as lead systems engineer for the Hubble Space Telescope at Goddard Space Flight Center.	NIL	DAVID (Q5204342) 0.980>Th_{cross} (0.45)	DAVID (Q5204342) 0.987 (where NIL 0.019, 2nd)	NIL 0.798	NIL 0.965
NILK-sample test-755	TMS Entertainment (known for "Lupin III" and "Detective Conan") would animate the popular animated series "Tiny Toon Adventures" and "**Rainbow Brite**". ...	NIL	Rainbow Brite and the Star Stealer (Q2625526) 1.000>Th_{cross} (0.45)	Rainbow Brite and the Star Stealer (Q2625526) 1.000 (where NIL 0.000, 2nd)	Rainbow Brite and the Star Stealer (Q2625526) 1.000 (where NIL 0.000, 2nd)	Rainbow Brite and the Star Stealer (Q2625526) 0.999 (where NIL 0.001, 2nd)

Wrong predictions are marked with red. Normalised prediction score (after softmax) and the rank (if not 1st) are displayed after the predicted entity.

Research Questions

1. How to identify out-of-KB mention (or NIL mention), i.e., uncaptured by a Knowledge Base, from texts?

2. How to insert out-of-KB mentions as entities into a Knowledge Base?



H. Dong, J. Chen, Y. He, and I. Horrocks, 'Ontology Enrichment from Texts: A Biomedical Dataset for Concept Discovery and Placement', in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, Birmingham, United Kingdom, Oct. 2023, pp. 5316–5320. doi: 10.1145/3583780.3615126.

H. Dong, J. Chen, Y. He, Y. Gao, and I. Horrocks, 'A Language Model Based Framework for New Concept Placement in Ontologies', in *The Semantic Web*, ESWC 2024, Cham: Springer Nature Switzerland, 2024, pp. 79–99. doi: 10.1007/978-3-031-60626-7_5.

Datasets for ontology enrichment from texts

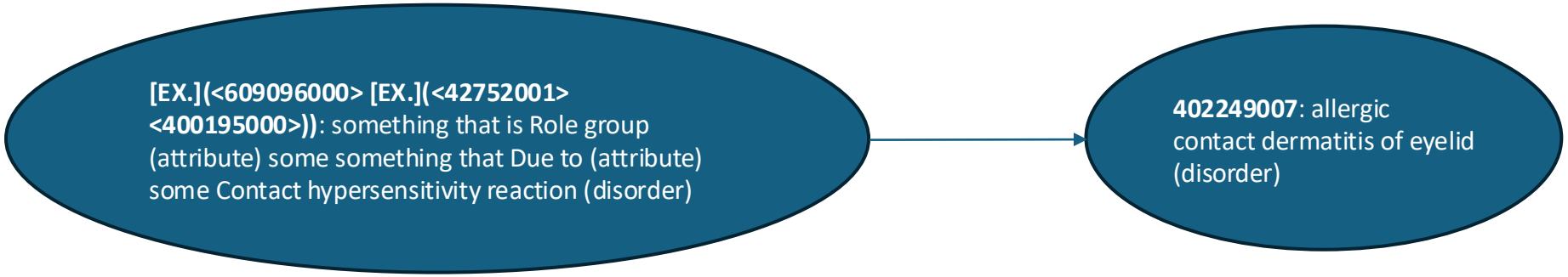
Datasets (with public access link in citations)	Task	NIL Discovery	Contextual Term	Concept Placement	Complex Concepts
MAG, WordNet [21, 30, 35]; OSConcepts, DroneTaxo, MeSH, SemEval [32, 34]	Taxonomy Completion	No	No	Yes	No
ChEBI_{500}, ChEBI^{+}_{500} [12]	Ontology Extension	No	No	Yes	No
SNOMED CT (English, manual, small-scale) [23]; SNOMED CT (Spanish, automated) [5]*	Concept Post-coordination	No	No	Yes	Yes
NILK [18]; ShARe/CLEF 2013 [31]; CLEF HIPE 2020 [11]; EDIN [22]; NEEL 2015-2016 [27, 29]	Out-of-KB Mention and Entity Discovery	Yes	Yes	No	No
MedMentions-SNOMED-CT-14 (-CPP, -Disease) [**this work**]	Concept Discovery and Placement	Yes	Yes	Yes	Yes

- **NIL Discovery**: discovering out-of-KB mentions (*cf.* in-KB mentions).
- **Contextual Term**: the input term has a context window in a text corpus.
- **Concept Placement**: the task finally places (or can be used to place) the term in the KB.
- **Complex Concepts**: the placement position in the KB includes *complex concepts*.

Complex Concept and Complex Edges

- Concepts that involve at least one logical operators, e.g., negation ($\neg$), conjunction ($\sqcap$), disjunction ($\sqcup$), existential restriction ($\exists r.C$), universal restriction ($\forall r.C$), etc.
- We use a rule-based verbaliser (in DeepOnto) to transform complex concepts into words.
- Complex *edges*: having one complex *concept* (usually as parent).

An example of the form $\exists r_1.(\exists r_2.A) \sqsupseteq B$

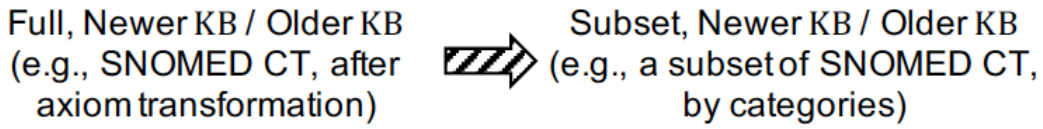


DeepOnto's Complex Concept Verbaliser: https://krr-oxford.github.io/DeepOnto/verbaliser/

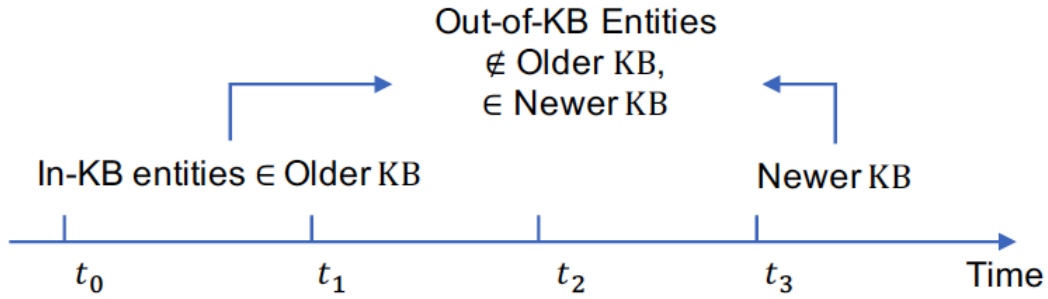
He Y, Chen J, Dong H, Horrocks I, Allocca C, Kim T, Sapkota B. DeepOnto: A Python package for ontology engineering with deep learning. Semantic Web Journal. 2023

Data Construction Pipeline

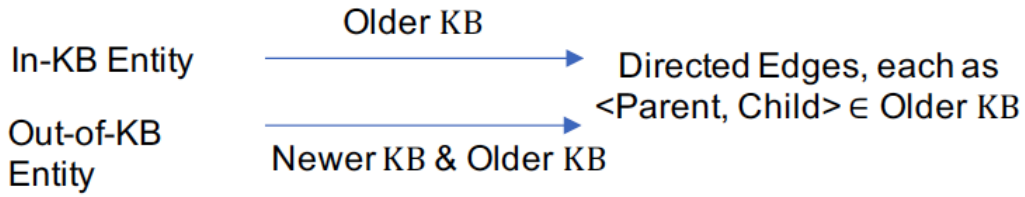
Step 0: KB and Subset Selection



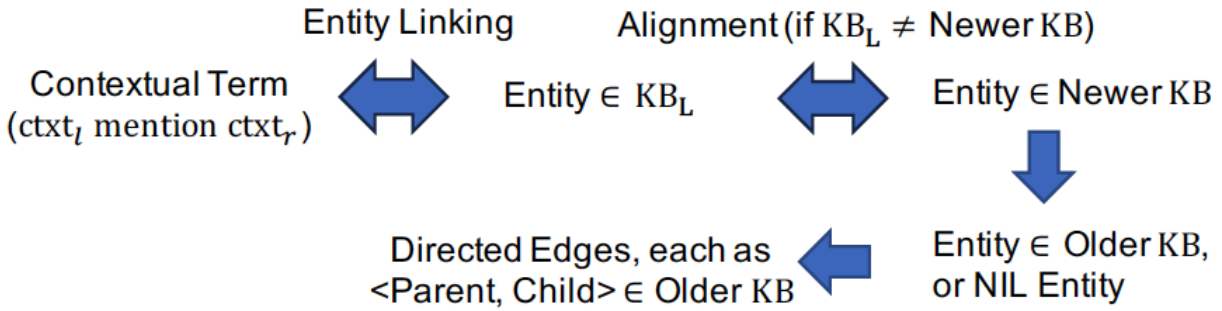
Step 1: KB Versioning to Synthesize Out-of-KB Entities



Step 2: Edge Extraction for In-KB and Out-of-KB Entities



Step 3: Mention-to-Edge Data Creation with Sources of Entity Linking and Alignment



SNOMED CT under different categories:

"**Disease**", and

"**CPP**", i.e., Clinical finding, Procedure, and Pharmaceutical / biologic product.

Older KB: ver 20140901, "S14" (included in UMLS 2014AB).

Newer KB: ver 20170301 (included in UMLS 2017AA) .

Entity Linking source: MedMentions, having around 5000 publication abstracts where mentions are linked to UMLS 2017AA.

Data Statistics

	Ontology: # all (# complex)		Corpus: # Mentions / # Mention-edge pairs / # Mention-edge pairs with *complex* edges			
	concepts	edges*	train, in-KB	valid, in-KB	test, in-KB	out-of-KB
MM-S14-Disease	64,900 (824)	237,826 (4,997)	11,812 / 887,840 / 917	4,248 / 383,457 / 203	3,970 / 316,319 / 393	605 / 1,637 / 13
MM-S14-CPP	175,895 (2,718)	625,994 (19,401)	34,704 / 1,398,111 / 9,475	11,707 / 548,295 / 4,305	11,564 / 478,424 / 4,129	1,000 / 2,131 / 22

A mention-edge pair denotes a mention (in a corpus) and one of its directed edges in the KB.

* The numbers of edges are those having one hop (including leaf nodes to NULL) and two hops from any paths in the ontology.

Data Format

- MM-S14-Disease
  - .DS_Store
  - mention-edge-pair-level
    - test-NIL-complex.jsonl
    - test-NIL.jsonl
    - test-in-KB-complex.jsonl
    - test-in-KB.jsonl
    - train-complex.jsonl
    - train.jsonl
    - valid-complex.jsonl
    - valid.jsonl
  - mention-level
    - .DS_Store
    - syn_attr
      - test.jsonl
      - train.jsonl
      - valid.jsonl
  - ontology

See more details at https://github.com/KRR-Oxford/OET/tree/main/dataset and https://zenodo.org/record/8043690

JSON keys for mention-level data:

```
"context_left"          # left context
"mention"               # mention
"context_right"         # right context
"label_concept_UMLS"    # label concept in UMLS
"label_concept"         # label concept in SNOMED CT older version
"label_concept_ori"     # label concept original, in SNOMED CT newer version if out of the older version
"label"                 # entity description
"parents_concept"       # parent concept SNOMED CT ID (or complex concept IDs)
"parents"               # parent concept title
"children_concept"      # child concept SNOMED CT ID (or complex concept IDs)
"children"              # child concept title
"label_id"              # row id in the entity catalogue ontology/xxx_syn_attr_hyp-all.jsonl file
"label_title"           # entity title
```

JSON keys for mention-edge-pair-level data:

```
"context_left"          # left context
"mention"               # mention
"context_right"         # right context
"label_concept_UMLS"    # label concept in UMLS
"label_concept"         # label concept in SNOMED CT older version
"label_concept_ori"     # label concept original, in SNOMED CT newer version if out of the older version
"entity_label_id"       # row id in the entity catalogue ontology/xxx_syn_attr_hyp-all.jsonl file
"entity_label"          # entity description
"entity_label_title"    # entity title
"parent_concept"        # parent concept SNOMED CT ID (or complex concept IDs)
"child_concept"         # child concept SNOMED CT ID (or complex concept IDs)
"parent"                # parent concept title
"child"                 # child concept title
"edge_label_id"         # row id in the edge catalogue ontology/xxx_edges.jsonl file
```

Concept placement

- Measured by micro-level (or mention-edge-pair level) precision@k , recall@k, and F1@k.
  - Other metrics: instance-level (or mention-level) metrics, e.g., percentage of mentions being corrected inserted in any of the edges (and in all the edges); ranking-based metrics, e.g., MRR.
- Methods:
  - Edge-Bi-encoder: for edge ranking
  - Edge-Bi-encoder + LLM (e.g., GPT-3.5-turbo):
    - (i) select top-k candidate edges (k as 50 in experiments) from Edge-Bi-encoder and
    - (ii) re-rank or select top-1 from top-k using LLM.
- GPT-3.5 may have limited knowledge in biomedical ontologies w.r.t. the prompt under zero-shot setting

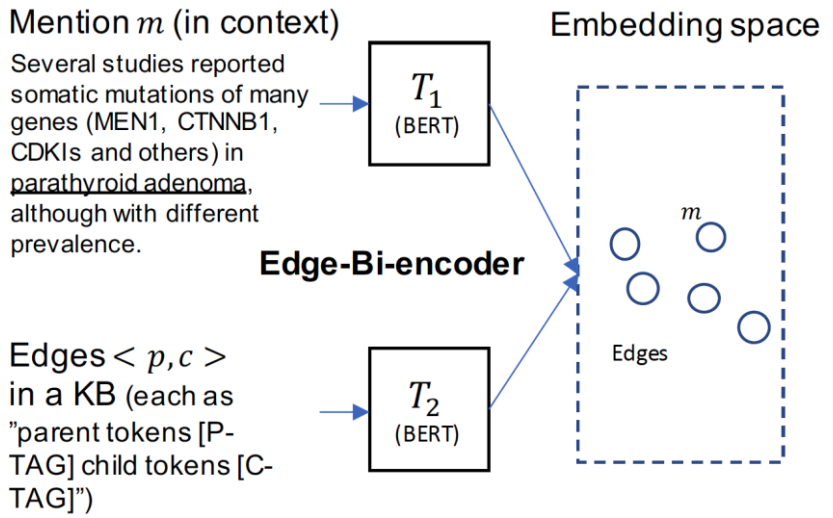
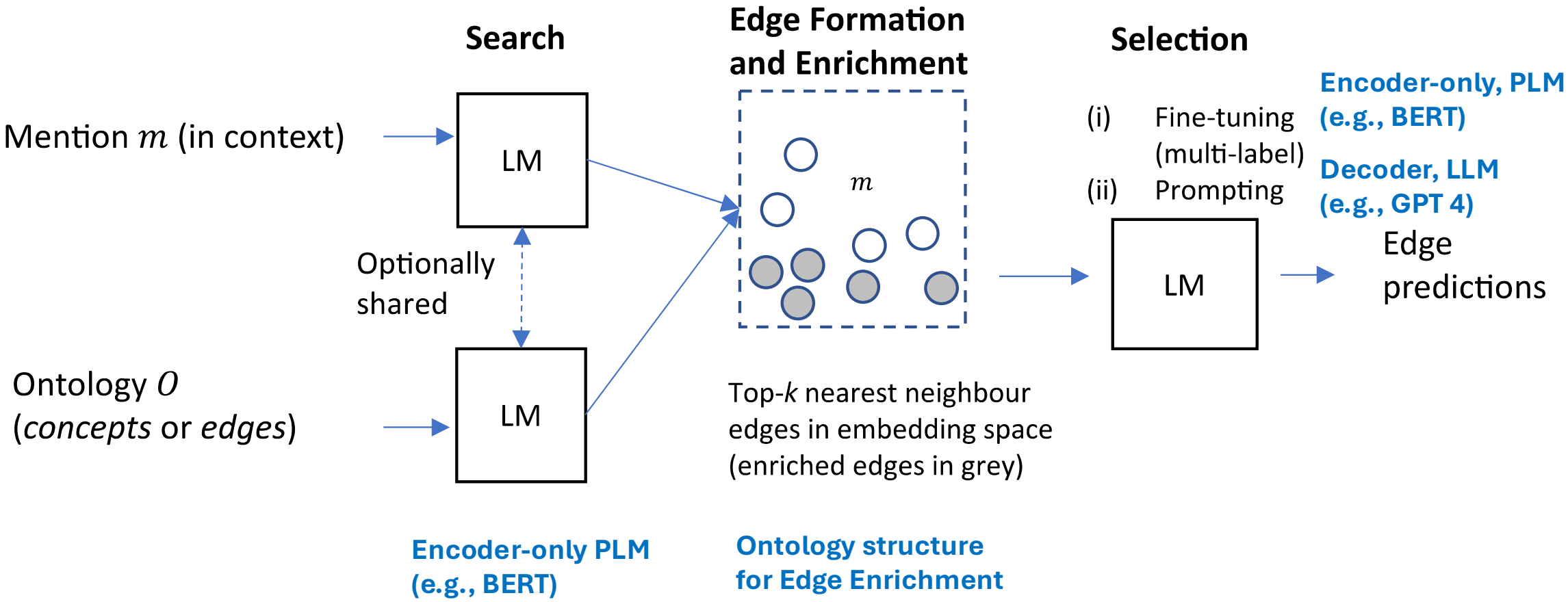


Table 4: Results on out-of-KB concept placement

MM-S14-Disease	P / R @1	P / R @5	P / R @10	P / R @50
Edge-Bi-encoder	**4.5 / 1.6**	6.0 / 11.0	5.4 / 19.9	2.1 / 38.4
+GPT-3.5	4.3 / 1.6	-	-	-
MM-S14-CPP	**P / R @1**	**P / R @5**	**P / R @10**	**P / R @50**
Edge-Bi-encoder	2.2 / 1.0	2.2 / 5.2	2.0 / 9.4	1.4 / 32.4
+GPT-3.5	**2.5 / 1.2**	-	-	-

The details of the prompts are available at https://github.com/KRR-Oxford/OET/tree/main/baseline-methods/concept-placement#prompting-gpt-35

A Language Model framework for new concept placement in ontologies



H. Dong, J. Chen, Y. He, Y. Gao, and I. Horrocks, 'A Language Model Based Framework for New Concept Placement in Ontologies',in *The Semantic Web*, A. Meroño Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, and P. Lisena, Eds, Cham: Springer Nature Switzerland, 2024, pp. 79–99. doi: 10.1007/978-3-031-60626-7_5.

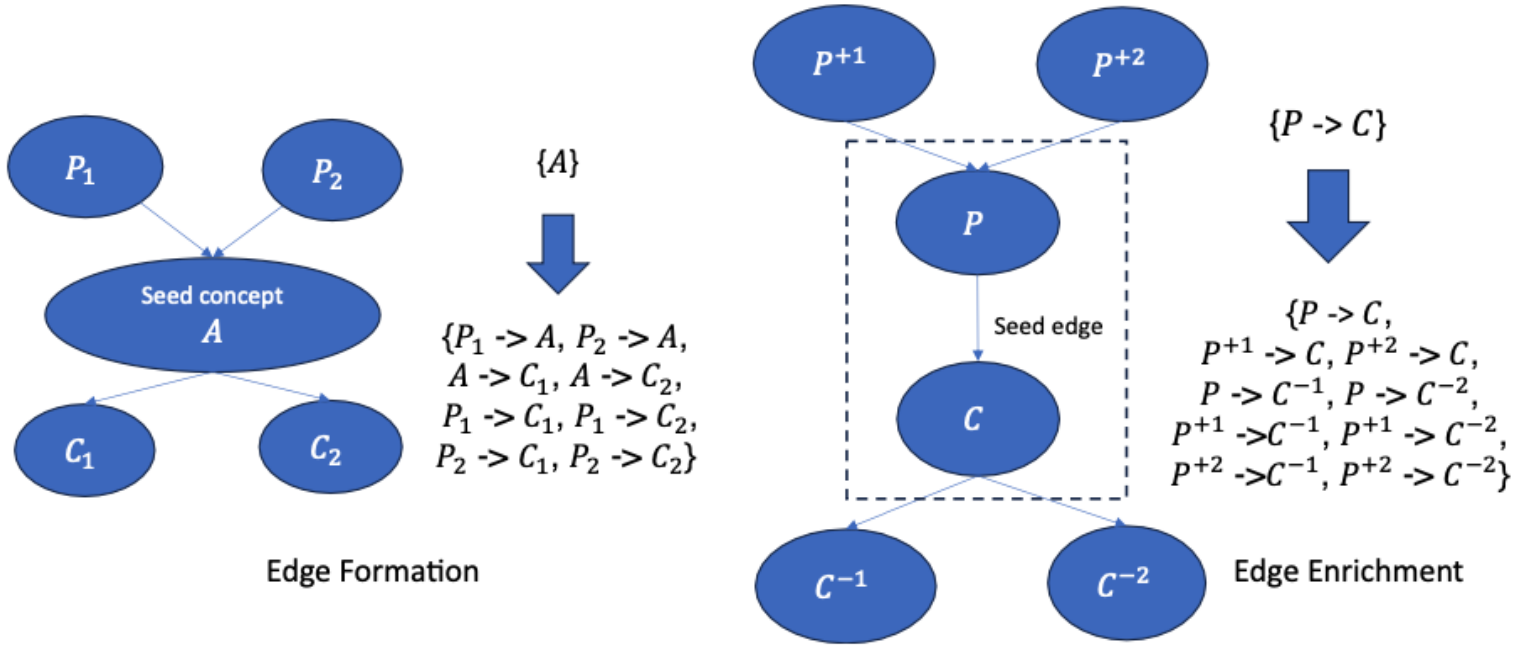
Edge Search

- Fixed BERT and its variations (e.g., SapBERT):
  - Concept search: Cosine similarity between embeddings of mention and concept label.

- Fine-tuning BERT: Edge Bi-encoder
  - Contextual Mention: [CLS] $ctxt_l$ [Ms] mention [Me] $ctxt_r$ [SEP]
  - Edge: [CLS] parent tokens [P-TAG] child tokens [C-TAG] [SEP]
    - Special token [NULL] to represent the child tokens of a leaf edge.

$$L_{m_i,e_i} = \sum_{j \neq i} [\alpha - s(m_i, e_i) + s(m_i, e_j)]_+; \quad s(m,e) = v_m \cdot v_e$$

Edge Formation and Enrichment



Edge Ranking after Edge Formation (with fixed BERT):

$$\text{Edge_score}_{fixed}(m, < P, C >) = \frac{\text{sim}(m, P) + \text{sim}(m, C)}{2}, \text{where } C \neq \texttt{NULL}$$

Edge Ranking after Edge Enrichment (with fine-tuned BERT):

$$\text{Edge_score}_{fine-tuned}(m, < P, C >) = s(m, < P, C >)$$

In Edge Formation: Prioritise leaf edge (when C is NULL) with the highest edge score if the most similar concept (A) of *m* is leaf.

Edge Selection

(i) Fine-tuning PLMs: Multi-label Edge-Cross-encoder (with PubMedBERT)

- Input as mention and edge concatenation:
  - [CLS] $ctxt_l$ [Ms] mention [Me] $ctxt_r$ [SEP] parent tokens [P-TAG] child tokens [C-TAG] [SEP]
  - Special token [NULL] to represent the child tokens of a leaf edge.

- Binary cross-entropy loss after a sigmoid activation

$$L_{m,e}^{cross} = -y_i \log(s_{m,e}^{(cross)}) + (1 - y_i)\log(1 - s_{m,e}^{(cross)})$$

$$s_{m,e}^{(cross)} = v_{cross} w$$

Edge Selection (ii) with LLMs: Prompt format and training instructions

Input:
Can you identify the correct ontological edges for the given mention (marked with *) based on the context? The ontological edge consists of a pair where the left concept represents the parent of the mention, and the right concept represents the child of the mention. If the mention is a leaf node, the right side of the edges will be NULL. If the context is not relevant to the options, make your decision solely based on the mention itself. There may be multiple correct options. Please answer briefly using option numbers, separated by commas. If none of the options is correct, please answer None.

mention in context:
There were no significant differences in age or sex distribution. Atrial fibrillation (*AF*) and previous unsuccessful ablation were more common in group I. A negative delta wave in lead II was the ECG finding with best sensitivity and specificity for the presence of a diverticulum. ...

options:
0.atrial arrhythmia -> atrial fibrillation and flutter
1.atrial arrhythmia -> lone atrial fibrillation
2.fibrillation -> lone atrial fibrillation
3.atrial arrhythmia -> NULL
4.cardiac arrhythmia -> NULL

Prompting: ### Input part + "### Response" -> ?
With explanations: ### Input part + "### Explanation" -> ?

Instruction tuning:
Input part + "### Response" -> 0,1,2
Input part + ### Explanation part + "###Response" -> 0,1,2 (with explanation)

("###" are Llama-2 model headers)

Explanation:
From the parents in the options above, including atrial arrhythmia (disorder), fibrillation (disorder), cardiac arrhythmia (disorder), the correct parents of the mention, AF, include atrial arrhythmia (disorder), fibrillation (disorder). Thus the options are narrowed down to 0, 1, 3, 2. From the children in the narrowed options, including atrial fibrillation and flutter (disorder), lone atrial fibrillation (disorder), NULL, the correct children of the mention, AF, include atrial fibrillation and flutter (disorder), lone atrial fibrillation (disorder). Thus, the final answers are 0, 1, 2.

Response:
0,1,2

Template for explanation generation

The highlighted parts are automatically generated given the candidate edges and the gold edges

$$\text{Expl_texts} = \text{Template}\,(\text{Edge}_{cand}, \text{Edge}_{gold})$$

Explanation by steps:
1. List all possible parents in the candidates
2. Find correct parents
3. Narrow the list of children
4. Find correct children
5. List the final answer

- From the parents in the options above, including xxx, xxx, xxx, the correct parents of the mention, xxx, include xxx, xxx, xxx. Thus the options are narrowed down to num1,num2,num3,...,numN. From the children in the narrowed options, including xxx, xxx, xxx, the correct children of the mention, xxx, include xxx, xxx, xxx. Thus, the final answers are num1,num2,num3,...,numN.

Data Statistics

		MM-S14-Disease	MM-S14-CPP
Ontology: # all (# complex)	concepts	64,900 (824)	175,895 (2,718)
	edges	237,826 (4,997)	625,994 (19,401)
Corpus: # M / # L / # L_{comp}	train, in-KB	11,812 / 887,840 / 917	34,704 / 1,398,111 / 9,475
	valid, in-KB	4,248 / 383,457 / 203	11,707 / 548,295 / 4,305
	valid, out-of-KB	329 / 672 / 10	568 / 979 / 13
	test, out-of-KB	276 / 965 / 3	432 / 1,152 / 9

A mention-edge pair or link (in L) denotes a mention in M (from a corpus) and one of its directed edges in the KB.

L_{comp} denotes concept edges, i.e., the edges involving a complex concept.

* The numbers of edges are those having one hop (including leaf nodes to NULL) and two hops from any paths in the ontology.

Self-supervised learning process:

Using in-KB to fine-tune or instruction-tune LMs

Using out-of-KB for validation and testing

Results

- Most **decoder-only**, generative LLMs performed worse than **encoder-only** LMs (i.e., edge-bi-encoder + edge-cross-encoder).
- **GPT-4** performs best in some top-1 & top-5 settings, but <u>not</u> in top-10.

Table 3. Overall results after edge selection for MM-S14-Disease and MM-S14-CPP datasets. Each setting has validation and testing results, separated by a slash (/) sign.

MM-S14-Disease	k	InR_{any}@1	InR_{all}@1	InR_{any}@5	InR_{all}@5	InR_{any}@10	InR_{all}@10
Inverted Index	10	0.6 / 0.0	0.0 / 0.0	1.8 / 2.5	0.9 / 0.7	10.0 / 12.0	9.1 / 10.1
	50	0.6 / 0.0	0.0 / 0.0	0.9 / 1.8	0.0 / 0.0	3.3 / 4.0	0.9 / 1.5
Fixed embs	10	4.0 / 1.4	0.9 / 0.0	6.7 / 2.2	1.5 / 0.7	16.1 / 13.0	7.0 / **12.3**
	50	4.0 / 1.4	0.9 / 0.0	6.7 / 2.2	1.5 / 0.7	13.4 / 4.4	3.0 / 2.5
Edge-Bi-enc	10	4.0 / 11.6	0.3 / 0.0	9.7 / **17.4**	2.7 / 1.4	**31.9** / 25.7	**14.6** / 8.0
	50	4.0 / 11.6	0.3 / 0.0	9.7 / **17.4**	2.7 / 1.4	13.7 / 20.3	4.3 / 2.5
+ Edge-Cross-enc	10	0.6 / 2.2	0.0 / 0.4	12.2 / 14.1	1.5 / 3.6	**31.9** / 25.7	**14.6** / 8.0
	50	7.3 / 7.6	**1.8** / **1.5**	17.9 / 15.6	**7.3** / **4.7**	25.8 / **26.5**	10.6 / 8.7
+ GPT-3.5	10	4.0 / 4.0	0.0 / 0.0	5.5 / 4.3	2.4 / 1.4	5.5 / 4.3	2.4 / 1.4
	50	3.3 / 1.5	0.0 / 0.0	4.6 / 3.6	1.5 / 0.4	4.6 / 3.6	1.5 / 0.4
+ GPT-4	10	**7.6** / 2.5	0.0 / 0.0	**21.0** / 9.1	8.8 / 4.0	21.0 / 9.1	8.8 / 4.0
	50	5.8 / 1.8	0.0 / 0.0	20.4 / 9.4	8.5 / 2.9	21.0 / 11.2	9.1 / 4.0
+ FLAN-T5-XL	10	2.7 / 1.8	0.6 / 0.0	2.7 / 1.8	0.6 / 0.0	2.7 / 1.8	0.6 / 0.0
+ FLAN-T5-XXL	10	4.6 / 5.1	1.2 / 0.7	4.6 / 5.1	1.2 / 0.7	4.6 / 5.1	1.2 / 0.7
+ Llama-2-7B	10	2.7 / 4.3	0.3 / 0.0	5.8 / 6.2	2.1 / 0.0	8.8 / 7.2	3.3 / 1.1
	50	1.8 / 3.3	0.0 / 0.0	3.7 / 5.8	1.2 / 0.7	4.0 / 6.9	1.2 / 0.7
+ Llama-2-7B-tuned	10	5.2 / **13.8**	0.0 / 0.0	7.6 / 16.3	1.5 / 1.8	7.6 / 16.3	1.5 / 1.8
	50	6.1 / 13.0	0.0 / 0.0	8.5 / 15.2	1.5 / 1.1	8.5 / 15.6	1.5 / 1.5
+ Llama-2-13B	10	3.3 / 6.9	0.0 / 0.0	7.0 / 10.1	0.3 / 1.8	8.5 / 11.2	0.6 / 2.2
	50	2.4 / 5.8	0.0 / 0.0	5.5 / 9.4	1.8 / 0.4	7.0 / 9.8	2.4 / 0.4
+ Llama-2-13B-tuned	10	7.3 / 10.1	0.0 / 0.0	7.6 / 10.5	0.6 / 0.7	7.6 / 10.5	0.6 / 0.7
	50	4.6 / 7.2	0.0 / 0.0	6.1 / 8.0	1.2 / 1.4	6.1 / 8.3	1.2 / 1.4

See resource and time usage in Appendix 1:

- All results with an NVIDIA Quadro RTX 8000 GPU card (48GB GPU).
- Instruction-tuning inference slow, e.g., Llama-2-7B to generate explanation and answer taking around 1.5 minutes (78s) per mention.

Results (cont.)

Applicability:

- Potential to suggest edge ranking for terminologists.
- A balance to achieve: Larger k vs. more manual effort

MM-S14-CPP	k	InR_{any}@1	InR_{all}@1	InR_{any}@5	InR_{all}@5	InR_{any}@10	InR_{all}@10
Inverted Index	10	0.4 / 0.0	0.0 / 0.0	0.9 / 0.5	0.0 / 0.0	5.5 / 5.8	5.1 / 5.3
	50	0.4 / 0.0	0.0 / 0.0	0.4 / 0.0	0.0 / 0.0	0.9 / 1.4	0.0 / 0.5
Fixed embs	10	2.8 / 1.2	0.7 / 0.2	6.3 / 3.7	2.3 / 2.1	11.3 / 8.3	8.3 / 7.4
	50	2.8 / 1.2	0.7 / 0.2	6.3 / 3.7	2.3 / 2.1	7.9 / 6.0	3.9 / 4.6
Edge-Bi-enc	10	2.5 / 6.3	0.0 / 0.2	6.2 / 11.8	1.2 / 1.9	**32.0** / **27.8**	**19.7** / **14.4**
	50	2.5 / 6.3	0.0 / 0.2	6.2 / 11.8	1.2 / 1.9	8.6 / 14.4	3.0 / 3.5
+ Edge-Cross-enc	10	3.4 / **9.3**	0.2 / 0.0	7.8 / 13.7	2.1 / 2.3	**32.0** / **27.8**	**19.7** / **14.4**
	50	4.9 / 3.9	**2.1** / 0.2	15.3 / **17.6**	6.3 / **6.9**	24.8 / 26.6	13.2 / 14.4
+ GPT-3.5	10	5.1 / 3.9	0.0 / 0.0	7.9 / 6.0	3.3 / 3.5	7.9 / 6.0	3.3 / 3.4
	50	1.8 / 1.9	0.0 / 0.0	3.9 / 2.8	0.9 / 0.7	4.0 / 2.8	0.9 / 0.7
+ GPT-4	10	**6.2** / 2.6	0.0 / 0.0	**16.5** / 8.4	**9.3** / 5.8	16.7 / 8.4	9.7 / 5.8
	50	3.3 / 2.1	0.0 / 0.0	10.9 / 8.6	6.0 / 5.6	12.7 / 9.0	7.2 / 6.0
+ FLAN-T5-XL	10	2.6 / 1.9	0.5 / 0.7	2.6 / 1.9	0.5 / 0.7	2.6 / 1.9	0.5 / 0.7
+ FLAN-T5-XXL	10	3.5 / 4.2	1.2 / **0.9**	3.5 / 4.2	1.2 / 0.9	3.5 / 4.2	1.2 / 0.9
+ Llama-2-7B	10	1.8 / 4.6	0.0 / 0.2	4.8 / 7.2	0.7 / 1.9	8.8 / 10.4	3.9 / 3.9
	50	1.2 / 3.5	0.0 / 0.0	2.5 / 5.1	0.7 / 0.9	3.0 / 6.3	1.1 / 1.2
+ Llama-2-7B-tuned	10	2.6 / 7.2	0.0 / 0.0	6.5 / 10.6	1.9 / 1.2	7.6 / 12.7	2.5 / 3.2
	50	2.5 / 4.6	0.0 / 0.0	3.3 / 6.7	0.5 / 0.6	4.0 / 8.1	0.9 / 1.4
+ Llama-2-13B	10	1.8 / 3.9	0.0 / 0.0	5.5 / 8.3	2.6 / 1.4	7.2 / 9.7	3.9 / 2.8
	50	2.1 / 3.5	0.0 / 0.0	2.0 / 5.6	0.7 / 0.5	4.2 / 6.0	0.7 / 0.9
+ Llama-2-13B-tuned	10	4.2 / 2.5	0.0 / 0.0	5.3 / 3.5	2.3 / 1.2	5.5 / 3.7	2.5 / 1.2
	50	5.3 / 3.2	0.0 / 0.0	6.5 / 4.2	2.3 / 0.9	6.7 / 4.2	2.3 / 0.9

A case study

Mention in context

...Since no one had **CKD** in partial nephrectomized patients, we determined risk factors for CKD in radical nephrectomized patients...

http://snomed.info/id/709044004
Chronic kidney disease

Not available in SNOMED CT 2014.09
(although was modified from "Chronic renal impairment" ID 236425005)

Edge-bi-encoder candidates (top-10) – with correct ones in bold (incomplete edges)

options:
0.renal impairment → end stage renal disease
1.renal impairment → renal failure following molar and/or ectopic pregnancy
2.renal impairment → renal failure syndrome
3.renal impairment → chronic kidney disease due to hypertension
4.kidney disease → impaired renal function disorder
5.kidney disease → renal function impairment with growth failure
6.kidney disease → renal failure syndrome
7.renal impairment → chronic renal failure syndrome
8.disorder of the genitourinary system → impaired renal function disorder
9.renal impairment → hypertensive heart and renal disease with renal failure

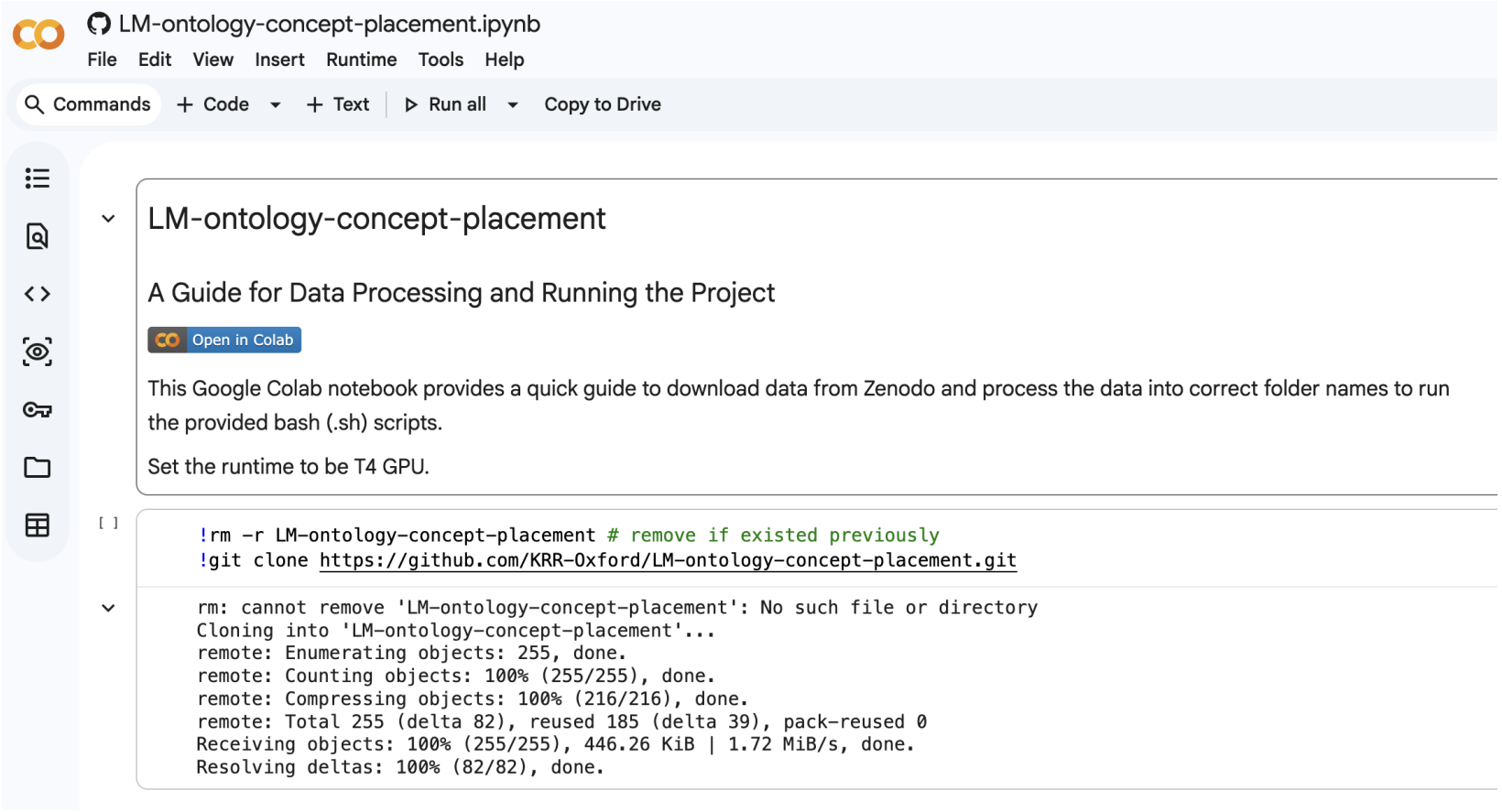
Explainable instruction-tuned Llama-2-7B

Explanation:
From the parents in the options above, including renal impairment , kidney disease , disorder of the genitourinary system , the correct parents of the mention, CKD, include **renal impairment** , kidney disease , disorder of the genitourinary system . Thus the options are narrowed down to 0, 1, 2, 3, 7, 9, 4, 5, 6, 8. From the children in the narrowed options, including end stage renal disease , renal failure following molar and/or ectopic pregnancy , renal failure syndrome , chronic kidney disease due to hypertension , chronic renal failure syndrome , impaired renal function disorder , renal function impairment with growth failure , the correct children of the mention, CKD, include **chronic kidney disease due to hypertension** , **chronic renal failure syndrome** , impaired renal function disorder , renal function impairment with growth failure . Thus, the final answers are **3, 7**, 5, 6, 8.

Response: **3,7**,5,6,8

OET data Jupyter Notebook

- https://github.com/KRR-Oxford/LM-ontology-concept-placement/blob/main/LM-ontology-concept-placement.ipynb -> “Open in Colab”



Taxonomy completion

J. Shi, H. Dong, J. Chen, Z. Wu, and I. Horrocks, 'Taxonomy Completion via Implicit Concept Insertion', in *Proceedings of the ACM Web Conference 2024*, in WWW '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 2159–2169. doi: 10.1145/3589334.3645584.



Figure 1: Implicit concept examples ("Clothing" and "Shoes")

Code and data available at https://github.com/jingcshi/ICON

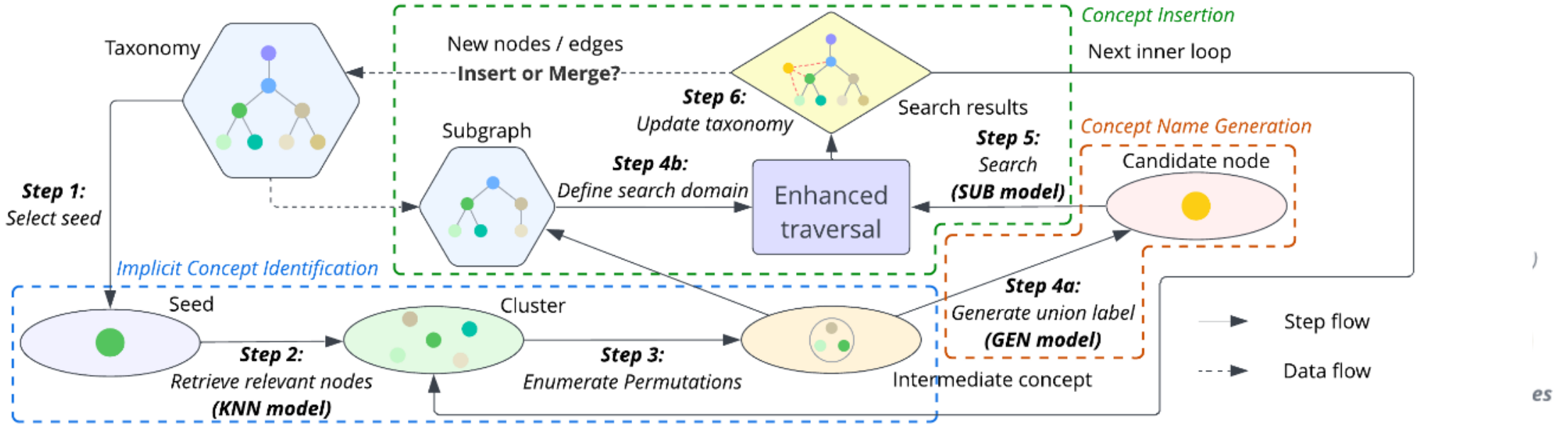


Figure 3: Workflow of ICON. In steps 1 - 2, it retrieves a cluster of nodes for a given seed. In steps 3 - 6, it traverses some subsets of the cluster, generates a labelled concept for each subset and inserts it into the taxonomy. After looping steps 3 - 6 over all the intended subsets, it goes back to step 1 and selects a new seed. Steps 1 - 2 form the outer loop, and steps 3 - 6 form the inner loop.

Open access links

RQ1 – identify out-of-KB mentions from texts

- BLINKout, GitHub: https://github.com/KRR-Oxford/BLINKout, Dataset: https://zenodo.org/record/8228371

RQ2 – insert out-of-KB mentions as entities into an ontology

- OET, GitHub: https://github.com/KRR-Oxford/OET, Dataset: https://zenodo.org/record/8043690
- LM-ontology-concept-placement, GitHub: https://github.com/KRR-Oxford/LM-ontology-concept-placement, Dataset: https://zenodo.org/records/10432003

Taxonomy completion, implicit - ICON, GitHub: https://github.com/jingcshi/ICON

This work was supported by EPSRC projects, including ConCur (EP/V050869/1), OASIS (EP/S032347/1), UK FIRES (EP/S019111/1); Samsung Research UK (SRUK); eBay Inc..

Conclusion

- We formulated the problem of new concept placement as two sub-questions (identification & insertion) and provided datasets with our solutions.

- Future studies:
  - **Multimodality**:
    - Can cross-modal information, e.g. images, genes, tables, help enrich new concepts into a Knowledge Graph?

  - **Mechanisms of explanations in LLM**:
    - Can KG evaluate & enhance LLMs' generated explanation?
    - Mechanistic explanation for ontology enrichment

  - **Human+AI collaboration**: working with terminologists and clinical coders (assigning concepts to clinical texts).

- Acknowledgement
  - Other key contributors:
    - Yuan He (Oxford, now in Amazon; DeepOnto, BERTMap, Bio-ML, HiT)
    - Mathias Jackermeier (Oxford; Box^{2}EL),
    - Jingchuan Shi (Oxford, now in eBay; ICON)
    - Yiping Song (Manchester; GenOM)
    - Ian Horrocks (Oxford)
  - Collaborators
    - SNOMED International, Sansumg Research UK, eBay
  - Main funders (projects):
    - EPSRC (OntoEm & ConCur), Samsung Research UK (Personal KG), eBay
- **Feel free to contact us**
  - hui.yang-2@manchester.ac.uk
  - jiaoyan.chen@manchester.ac.uk
  - h.dong2@exeter.ac.uk