A Resolution-Alignment-Completeness System for Diagnosis Code Imputation in Clinical Knowledge Graphs

Shervin Mehryar[1,*], Özge Erten[1], Tsvetan Asamov[2], Svetla Boytcheva[2] and Remzi Çelebi[1]

[1]Institute of Data Science, Maastricht University, Paul-Henri Spaaklaan 1, 6229 GT, Maastricht, Netherlands
[2]Ontotext, Sofia, Bulgaria

Abstract

The rapid growth of electronic health records (EHR) presents challenges in data integration and interoperability due to the incomplete nature of this information, limiting its effective utilization. While ontology-based data integration across diverse resources has been widely practiced, the process of codifying records remains error-prone and largely manual. Knowledge Graph Embeddings as an alternative solution can provide for efficient quality data representations. In this paper, we propose an embedding-based system that applies entity resolution and alignment across medical terminologies and ontologies for imputing codified data. Through experimentation we demonstrate the benefits of the proposed solution for semantic completion and consistency tasks in terms of NDCG@K and Sem@K.

Keywords

Entity Resolution, Entity Alignment, Knowledge Graph Completion, RAC, SPHN, SNOMED, EHR, ICD

1. Introduction

The amount of data stored as electronic health records (EHR) has grown significantly in recent years, now including an immense quantity of interactions, events and various medical information [1]. As such, data integration will play a transformative role in health information systems for the years to come, bridging the gap between research and applications, enabling integrative analyses, and improving clinical decision support systems. High-quality multi-sourced data rely heavily on codifying EHR records using standard medical ontologies (e.g. ICD, SNOMED CT). An important challenge to achieving ontology-based integration of patient data is semantic incompleteness whereby important encoded fields are missing and require records to be imputed.

Clinical knowledge graphs have been adopted to complement EHR modeling as a means to improve predictive performance and integrate expert knowledge with data-driven insights [2]. They enable effective learning of patterns from the clinical data through semantically rich relations between medical concepts (e.g. diagnoses, procedures, prescriptions, and lab measurements). To learn the graphical structure of data, embedding models have been successfully applied that can capture hidden hierarchies for downstream clinical tasks (e.g. comorbidity, readmission prediction) [3]. These models however suffer from the inherent incompleteness of knowledge graphs, rooted in missing values and inconsistent codification from legacy systems. Furthermore, such knowledge bases are predominantly developed relying on a single source ontology and lack the power of interoperability. Although alignment methods exist to consolidate complementary knowledge from disparate knowledge graphs [4], their application to EHR integration is not fully explored yet.

In this paper, we propose a robust system for consolidating tabular EHR into semantically consistent health knowledge graphs, using standard terminologies aligned with medical ontologies. Our framework (described in section 2) comprises multiple stages for extracting, linking, resolving, and

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
shervin.mehryar@maastrichtuniversity.nl (S. Mehryar)
0000-0002-9062-6925 (S. Mehryar)

completing the data. In particular, we focus on records from patients with Cardiovascular Diseases (CVD) with ICD[1] diagnosis codification. We graphically transform the source data according to SPHN RDF Schema[2] and apply state-of-the-art knowledge graph embedding methods to complete missing diagnosis codes, i.e. code imputation. In order to improve the semantic consistency and precision of predictions, the underlying patient subgraphs are aligned with SNOMED CT[3] medical ontology. Through experimentation (section 3), we demonstrate the contributions of RDF-schema based entity resolution and medical ontology-based entity alignment on the task of diagnosis code imputation using several KGE methods (section 4).

2. Methodology

In this section we provide a detailed description of the proposed resolution-alignment-completeness (**RAC**) system which follows a modular design as shown in Figure 1. Firstly, relevant patient entities are extracted from a relational data source and **resolved** via semantically equivalent identifiers based on a fixed RDF schema in order to generate a knowledge base. In the second phase, the result knowledge base is transformed and vectorized into translational embeddings, excluding class axiomatic and hierarchical information. In the third stage, semantics are extracted and **aligned** with a reference medical ontology in order to improve embedded representations as a KG **completion** task. Lastly, the imputed data are validated in terms of domain and range constraints and semantic consistency metrics.

In order to generate a knowledge base form the input relational data, table entries are mapped to core concepts retrieved from the RDF schema with foreign keys establishing links between entities across tables (e.g. linking a diagnosis to a patient). Rows are identified with unique URIs as instances of the corresponding RDF class using the Admission numbers as the primary key, and column attributes (e.g. Diagnosis Code) are mapped to retrieved RDF predicates (e.g. hasCode). Biomedical codes (e.g. ICD, NDC, etc) are transformed into unique identifiers to resolve their semantically equivalent instances. After resolving and mapping all entities, the final knowledge graph is generated in N-Triple format with semantically consistent resources aligned with the entity's scope (e.g. lab test or prescription correctly associated with relevant admission).

The transformed RDF from above is embedded using multiple translation based algorithms extended to include subsumption and instance checking reasoning. In particular, we apply the algorithm in [5] which introduces a modified loss function to facilitate transitive and hierarchical knowledge. Formally, representations are learned for entities and relations in a knowledge graph $\mathcal{G}$ assuming the following vector algebra holds for a relation r between entities e_1 and e_2: $\vec{e_2} = \vec{e_1} + \vec{r}$. The loss for the single fact in this triple format is given by $d(e_1, r, e_2) = ||(\vec{e_1} + \vec{r}) - \vec{e_2}||_2$. The extension to K-hop reasoning, i.e. $(e_1, \{r_k\}_{k=1}^K, e_K) \in \mathcal{G}$, follows the path loss given by $d(e_1, \{r_k\}_{k=1}^K, e_K) = ||\vec{e_1} + \vec{r_1} + \cdots + \vec{r_K} - \vec{e_K}||_2$. Unseen (missing) values can be predicted (imputed) using this setup where the target node e_K is of the relevant type (e.g. diagnosis code).

In order to enrich the semantic soundness of predictions, the imputed types are subsequently aligned with a biomedical ontology, namely SNOMED CT (a comprehensive, multilingual healthcare terminology used to encode clinical information in a consistent and well-structured manner). The semantic richness of the latter enables capturing complex clinical relations - for instance "Pain in the left arm" with a type 'finding' predicated with attribute "finding site" and value "left arm", and thus facilitates more precise clinical data modeling through the extraction of structured information. Lastly, the semantic types for predictions are extracted and validated against the biomedical ontology (domain and range) constraints in order to add into the knowledge base.

[1] https://icd.who.int/
[2] https://www.biomedit.ch/rdf/sphn-schema/sphn
[3] https://www.snomed.org/

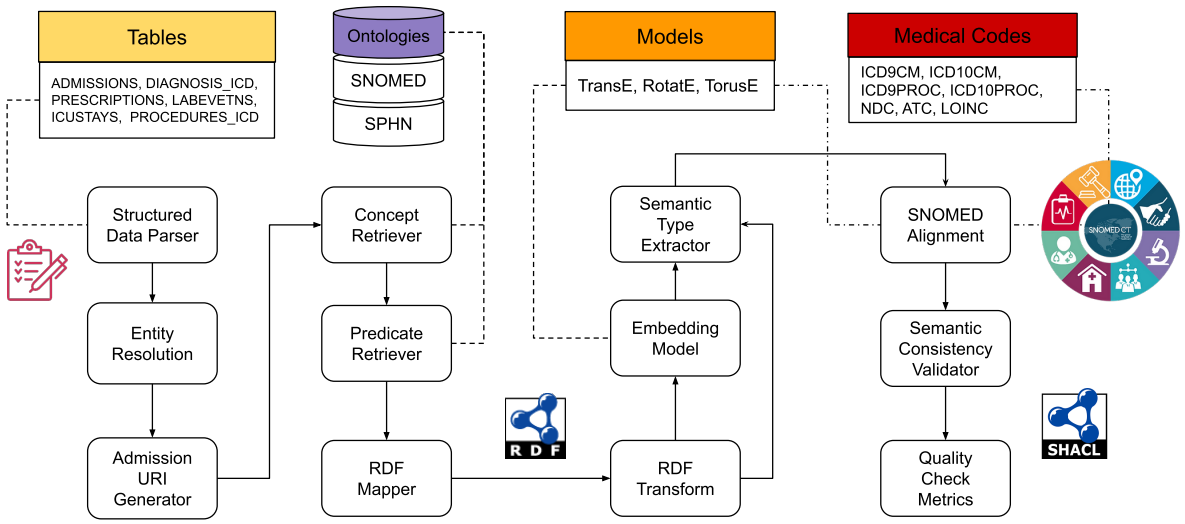

Figure 1: Resolution (R), Alignment (A), Completeness (C) system for imputing missing data in Clinical Knowledge Graphs.

3. Experimental Setup

We use the MIMIC III repository which contains data associated with 53,423 distinct hospital admissions for adult patients (aged 16 years or above) admitted to critical care units between 2001 and 2012 [6]. Specifically, we focus on a smaller subset related to Cardiovascular Disease (CVD) diagnoses with precodings in the range 410-430 (e.g. 428.22: Chronic systolic heart failure, 428.23: Acute on chronic systolic heart failure, 428.32: Chronic diastolic heart failure). These codes categorize various forms and severities realted to the dysfunction of the heart in ICD-9, and in ICD-10 are largely replaced by categories under I50 (Heart Failure). To generate multiple subsets, we identify and store admissions in increments of 100 (i.e. DS-100, DS-200, etc) for patients with one or more CVDs ICD codes associated with them. The scope is further narrowed down to associated medication prescribed, lab results and measurements taken, hospital procedures performed, diagnoses made, stored in separate tables from columns and relations selected by examination according to the target RDF schema as described in section 2. The final data covers 49,785 admission instances from 38,597 patients.

In order to map patient relational records to an RDF schema, we use the Swiss Personal Health Networkschema (SPHN) model. It provides an RDF schema to map core concepts and predicates for the purpose of seamless and extensible interoperability. In particular, we focus on 13 core concepts, namely, 'LabTestEvents', 'LabResult', 'Code', 'DrugPrescription', 'Drug', 'Substance', 'Diagnosis', 'BilledProcedure', 'AminstrativeCase', 'SubjectPseudoIdentifier', 'BirthDate', 'MedicalProcedure', 'BodySite', and 'AdministrativeGender'. As for predicates, we model a total of 9, namely, 'hasCode', 'hasLabTest', 'hasAdministrativeCase', 'hasSubjectPseudoIdentifier', 'hasDrug', 'hasActiveIngredient', 'hasAdministrationRoute', 'hasBirthdate', 'hasAdministrativeGender', to capture the relations between the entities. Additionally, we include 'is a' relation to indicate the type of entities, 'rdfs:subClassOf' to indicate subsumptions, and 'owl:sameAs' to indicate equivalent types. The transformations result in 67638, 129653, 197889, 260640, 333527 triples for each data subset DS-100 through DS-500, respectively.

The embeddings are generated with various models using the Pykeen[4] library. We use TransE, TorusE, and RotatE models with the RMSProp optimizer and hyperparameters set as `embeddig_dim=90`, `batch_size=128`, `learning_rate=0.0015`, and `num_epochs=100`. We use codes from our knowledge base which classify diagnoses, symptoms, and procedures to align with biomedical ontology identifiers using the ICD-9-CM to SNOMED CT Mapping Project (December 2022 version) [5]. We

[4] https://github.com/pykeen/
[5] https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html

encountered both one-to-one and one-to-many matches. For example, the ICD-9 code 427.69 maps to multiple disorders in SNOMED, including “Multifocal premature ventricular complexes" and “Supraventricular bigeminy". In our experiments, we considered these distinct medical conditions and retain all such mappings. The training and evaluations are done with a 8-to-2 split ratio for each dataset, and the performance is reported in terms of discounted cumulative gain NDCG@K for retrieval accuracy and Sem@K for semantic consistency [7].

Dataset	Model	Baseline		RC		RAC	
		NDCG@1	NDCG@10	NDCG@1	NDCG@10	NDCG@1	NDCG@10
DS-100	TransE	0.00	0.04	0.47	0.96	0.59	0.97
	RotatE	0.01	0.09	0.43	0.91	0.85	0.95
	TorusE	0.03	0.12	0.47	0.994	0.91	**1.0**
DS-200	TransE	0.02	0.11	0.28	0.78	0.20	0.70
	RotatE	0.02	0.23	0.36	0.83	0.67	0.95
	TorusE	0.08	0.27	0.44	0.94	0.71	**0.97**
DS-300	TransE	0.005	0.09	0.14	0.58	0.08	0.43
	RotatE	0.02	0.15	0.32	0.72	0.51	0.80
	TorusE	0.04	0.20	0.35	0.79	0.58	**0.91**
DS-400	TransE	0.004	0.09	0.11	0.57	0.04	0.40
	RotatE	0.04	0.18	0.37	0.77	0.55	0.89
	TorusE	0.07	0.31	0.44	0.88	0.65	**0.96**
DS-500	TransE	0.0	0.07	0.09	0.48	0.05	0.31
	RotatE	0.04	0.147	0.34	0.72	0.53	0.81
	TorusE	0.05	0.29	0.25	0.78	0.61	**0.93**

Table 1
Model performance on Diagnosis Code Imputation. TransE, RotatE, and TorusE are baseline. RC and RAC refer to variations with multi-hop resolutions and biomedical ontology alignment. Evaluated by NDCG@1 and NDCG@10 between 0 and 1. The best and second best scores are highlighted and underlined.

4. Results and Discussions

The performance results of the proposed RAC system are shown in Table 1 in terms of five subsets (DS-100, DS-200, DS-300, DS-400, DS-500) that differ in the number of admission events as described above. The baseline models alone perform poorly by NDCG@1 and NDCG@10 metrics, due to the inability to capture the complex semantics of clinical knowledge graphs. Using TorusE for 100 patient admissions (total of 67638 triples), NDCG@1 and NDCG@10 are at their highest (0.91 and 1.0) with RAC due to extended interactions between nodes in alignment with structural information from biomedical node embeddings. Without alignment to SNOMED (TorusE plus RC), these metrics decrease by 0.45 and 0.006 points, respectively. Similar patterns can be observed using TransE and RotatE, although they generally underperform as compared to TorusE.

With increasing admission numbers to 400 and 500, TorusE and RotatE with RAC maintain high performance and experience a lower drop over TransE. It is worth noting since each admission event has more than one diagnosis code associated with it, that NDCG@1 is naturally lower (i.e. no single correct code). Overall, it can be seen that the addition of RAC contributes to improved performance and robustness for code imputation in clinical KGs. In terms of semantic consistency as shown in Table 2, the predictions by RAC models are accepted at a high rate as measured by Sem@K metrics, for K=1, 3, 5, and 10. This implies that in almost all cases the semantic types of output values after ranking, are consistent with the (domain and range) constraints of the entity completion task.

Dataset	Sem@1	Sem@3	Sem@5	Sem@10
DS-100	1	1	1	1
DS-200	0.98	1	1	1
DS-300	0.98	1	1	1
DS-400	1	1	1	1
DS-500	0.98	0.99	0.99	0.99

Table 2
RAC Semantic Consistency between 0 and 1 evaluated by Sem@K [7] on Diagnosis Code Imputation task.

5. Conclusions

Clinical knowledge graphs offer a valuable form of data representation through typified node connections, allowing efficient querying in EHR management systems. The presence of incomplete substructures however can severely degrade retrieval. In this work, we propose a RAC system that leverages knowledge graph embeddings to complete missing links. In particular, for imputing CVD diagnosis codes, our system improves retrieval performance by 0.31 to 0.88 points in NDCG@10 over moderately large graph sizes.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] I. de Zegher, K. Norak, D. Steiger, H. Mueller, D. Kalra, B. Scheenstra, I. Cina, S. Shulz, K. Uma, P. Kalendralis, E.-M. Lotmam, M. Dumontier, R. Celebi, Artificial intelligence based data curation: enabling a patient-centric european health data space, Frontiers in medicine 11 (2024).

[2] I. Y. Chen, M. Agrawal, S. Horng, D. Sontag, Robustly extracting medical knowledge from ehrs: a case study of learning a health knowledge graph, in: Pacific Symposium on Biocomputing 2020, World Scientific, 2019, pp. 19–30.

[3] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, A. Dai, Learning the graphical structure of electronic health records with graph convolutional transformer, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 606–613.

[4] H. Li, Z. Zhu, H. Zhu, B. Jin, Fusing attribute character embeddings with truncated negative sampling for entity alignment, Electronics 12 (2023) 1947.

[5] S. Mehryar, R. Celebi, Improving transitive embeddings in neural reasoning tasks via knowledge-based policy networks, in: CEUR Workshop Proceedings, volume 3337 of *CEUR Workshop Proceedings*, 2022, pp. 16–27. Joint 2nd Semantic Reasoning Evaluation Challenge and 3rd SeMantic Answer Type, Relation and Entity Prediction Tasks Challenge, SemREC-SMART 2022.

[6] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016) 160035. Publisher: Nature Publishing Group.

[7] N. Hubert, P. Monnin, A. Brun, D. Monticolo, Sem@ k: Is my knowledge graph embedding model semantic-aware?, Semantic Web (2023) 1–37.