

Automatic Medical Knowledge Graph Construction

Joel Pardo–Ferrera

Ontology Engineering Group (OEG), Universidad Politécnica de Madrid (UPM), Madrid, Spain

Abstract

Knowledge Graphs are crucial for structuring and integrating large amounts of data, improving decision-making and data interoperability, especially in the healthcare domain. This PhD thesis aims to implement a unified end-to-end framework for building a cross-lingual KG for English and Spanish in the healthcare sector using NLP techniques. Addressing the reliance on traditional methods in KG construction and the limited non-English language resources, this work seeks to refine the information extraction process within unstructured medical texts and facilitate the (re)use of existent ontologies (schema to represent the real-world).

Keywords

Knowledge Graph Construction, Large Language Models, Information Extraction, Electronic Health Records

1. Introduction and Motivation

The need to effectively structure information has become increasingly important with the surge of data we experience today [1]. The idea of Knowledge Graphs (KGs) is that they organize real-world information into triples, consisting of at least two concepts connected by a semantic relation, encapsulated in nodes and edges, respectively [2, 3]. KGs form the core of many commonly used applications today, including recommendation systems [4], search engines [5], and question-answering systems [6].

As an example of a KG in the medical domain, a triplet might represent a patient (concept) diagnosed with (relation) a specific disease (concept) or a medication (concept) prescribed to (relation) a patient (concept). This structure allows for a detailed and interconnected representation of medical knowledge. However, a distinction arises between concepts and instances. A concept is a type or category, while an instance is a specific example of that concept. Given the last example, taking Joel (instance) as a patient (concept) diagnosed with (relation) diabetes (instance), a specific disease (concept).

Knowledge Engineering emerged as the discipline dedicated to designing, developing, and maintaining systems capable of effectively representing knowledge [7]. In a more updated definition, this concept encompasses the set of activities aimed at capturing, conceptualizing, and formalizing knowledge for use in information systems [8]. These systems often use ontologies,


Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

✉ joel.pardof@upm.es (J. Pardo–Ferrera)

🆔 0000-0001-8064-0128 (J. Pardo–Ferrera)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

a structure that represents the real world [9, 10] visualized with nodes representing concepts and edges denoting relationships.

The construction of a KG is key within Knowledge Engineering, identifying its nodes and edges in different dimensions; task-specific, domain-specific, or open-domain manner [11]. This process can be made by experts of the domain or automatically or semi-automatically from unstructured data such as texts. The automatic process usually relies on Natural Language Processing (NLP) tasks, specifically, Information Extraction (IE).

Recently, deep learning methods in NLP have enhanced the automation of IE tasks, including named entity recognition (NER) and Relation Extraction (RE). NER is the task of identifying and classifying named entities in text into predefined categories [12], while RE involves identifying and classifying relationships between entities within a text [13]. These advancements enable the automatic construction of KGs by automating various tasks needed for their construction [11]. Therefore, we can infer that Automatic KG construction is a crucial aspect of Knowledge Engineering [14].

Given the necessity of having accurate and structured medical data, which is critical for improving patient care, clinical decision-making, and medical research, the Automatic KG Construction process takes importance. This process involves integrating complex, interrelated data from sources such as Electronic Health Records (EHRs), Electronical Medical Records (EMRs), clinical trials, and medical literature. Also, well-established medical terminologies like Medical Subject Headings (MeSH) [15] and Unified Medical Language System (UMLS) [16] can be integrated into KGs to enhance their utility. An example is SNOMED-CT [17], which is a KG itself and provides a standardized vocabulary for medical terms, ensuring consistent terminology across the KG.

The challenges in KG construction within the current frameworks come from the dependence on traditional machine learning methods for IE [18, 19]. These frameworks often involve multiple tasks in a pipeline, leading to the propagation of errors at each step [20]. RE is typically based on classification tasks with predefined relation types [21]. Moreover, most of the approaches reviewed lack medical resources in non-English languages, such as specialized terminologies and annotated corpora, limiting the interoperability of KG [1].

The design of ontologies for domain-specific KGs is challenging, primarily based on existing ones [22]. The evaluation of a KG involves several dimensions, such as accuracy, consistency, timeliness, and trustworthiness, although there is no clear benchmark [23].

This PhD thesis focuses on the creation of cross-lingual KGs in the Medical Domain, using an end-to-end unified framework integrating NLP techniques. The aim is to automatically identify entities and relationships within the text and align them with predefined ontologies, thus ensuring the accuracy of domain-specific KG.

The structure of the paper is as follows: Section 2 explores the state-of-the-art in automatic and semi-automatic construction of KGs. Section 3 presents the Open Research Problem, along with the Hypothesis and Research Questions that form the foundation of this study. Section 4 clarifies the main objective and the related sub-objectives. The Research Methodology is detailed in Section 5. The paper concludes with Section 6, which presents the Conclusions and outlines Future Work.

2. State of the Art

KG construction is known by various terms, including KG construction, creation, acquisition, and building. Additionally, it is sometimes referred to as KG foundation and establishment. This diverse terminology reflects the breadth of approaches and perspectives in the field. Now and so on, this tasks will be referred as KG construction.

Rotmensch et al. [18] propose a framework to construct a KG that uses a string-matching tool to search for concepts defined by Google Health Knowledge Graph (GHKG) and UMLS. Then, it relates symptoms to diseases from EMRs with statistical models such as Naive Bayes, logistic regression and Noisy OR. Another end-to-end framework is Health Knowledge Graph Builder (HKGB), proposed by Zhang et al. [19] to construct disease-specific KGs by utilizing machine learning (ML) algorithms combined with clinicians' prior knowledge. The process involves two KGs: Concept KG and Instance KG, leveraging both structured and unstructured data through two different pathways. Structured data is transformed into Resource Description Framework (RDF) (a standard model for data interchange on the web, using a triple structure) using mapping languages, while unstructured data employs Long Short-Term Memory with Conditional Random Fields (LSTM-CRF) for NER and a pattern-based algorithm for RE. This works demonstrate a reliance on traditional machine learning methods for IE in KG construction.

Rossanez et al. [20] present KGen, a shorthand for Knowledge Graph Generation. KGen is a KG construction framework that employs several independent NLP tasks in a pipeline format. It processes sentence inputs to extract entities and relations, and links the extracted information to an existing ontology in the biomedical domain. A modular component carries out each NLP task. The use of pipelines can result in error propagation at each step.

Finally, Murali et al. [21] presents a survey in which critical aspects in KG construction, such as representation, extraction, or completion, are analyzed. The survey emphasizes ontology-based representations to organize and integrate knowledge from various sources, adhering to EHR standards. For extraction, it highlights entity and relation extraction techniques, noting that the most commonly used models are Bi-directional LSTM-CRF for NER, and BERT-based models, along with some reliance on GPT-2 and GPT-3, for RE. Here, RE is often approached as a classification task, where predefined relation types are used to determine the connections between entities. As a result, the dependency on predefined types can hinder the ability to capture nuanced or context-specific relationships, leading to potential inaccuracies in the constructed KGs.

Within the medical domain, Wu et al. [24] and Abu-Salih et al. [25] agree that the construction of medical KGs commonly involves extracting entities and relationships from various medical resources. Using Large Language Models (LLMs) for KG construction can significantly enhance this process by automating tasks such as NER and RE, thereby improving accuracy and efficiency [26, 27]. This approach could serve to streamline and optimize automatic KG construction in the medical field.

Three distinct approaches to perform this IE process for KG construction are identified:

1. **NER and RE as independent components**, requiring separate training for each. In this approach, RE functions as a Relation Classification (RC) task, determining the existence or the non-existence of any semantic relation between the extracted entities. It allows

for the fine-tuning and optimization of entity recognition and relation classification so that they interact effectively. However, this method demands substantial amounts of annotated data for optimal performance. Alimova and Tutubalina [28] use BERT-based models like BioBERT and Clinical BERT for binary relation extraction. They combine supervised methods with several features, including the distance between entities, word embeddings, sentence embeddings, entity co-occurrence, and semantic types from MeSH, among others.

2. **NER and RE configured as components within pipeline in a multitasking learning framework**, allowing for simultaneous training while maintaining them as distinct components. Park et al. [29] proposed two separate modules to perform either NER and either relation extraction. The NER module uses a pre-trained BERT model on scientific texts to predict the types of entities within spans. The RE module then concatenates four vectors: each span-based entity vector, a max-pooled vector from the embedded tokens located between two entities, and the attention score vector. This approach allows the model to learn contextual features of the input sentences collaboratively, improving its ability to predict the relationship between the entities identified by the NER module.
3. **End-to-end system** where the model is trained as a unique component. This model internally distinguishes components responsible for NER and RE tasks. Some approaches in this category aim to address the challenge through an autoregressive method with a seq2seq model to output each triplet present in the input text [30].

Cross-lingual KGs are KGs that incorporate data from multiple languages, allowing for the integration and retrieval of knowledge across different linguistic contexts [11]. Each of these three approaches needs to rely on domain-specific resources in languages other than English, such as specialized medical terminologies, annotated corpora, and language-specific tools [1], which are not always available.

Designing the schema for domain-specific KG construction is challenging [22]. The reuse of established ontologies provides a foundation that enhances consistency and interoperability across different datasets and applications [31]. However, this process requires careful adaptation and extension to meet the specific needs of the domain and application at hand while maintaining the integrity and accuracy of the integrated knowledge [32]. Due to this fact, reusing ontologies already created and depurated would serve to ensure comprehensive and accurate representation.

There is no clear evaluation benchmark for KG construction. The evaluation of a KG is divided into several dimensions, each representing a specific characteristic inherent in a KG [23]. The common dimensions identified are accuracy, consistency, timeliness and trustworthiness [23, 33]. Most of these surveys share common evaluation dimensions, but some introduce new ones depending on the granularity of the evaluation as in Wang et al. [34]. For example, the dimension accuracy can be divided into Syntactic, Semantic or Timeliness, focusing on the grammatical rules defined for the domain and/or data model, if data correctly represents real world facts, and how the knowledge graph is currently up-to-date with the real world state, respectively. A KG must be thoroughly evaluated in a comprehensive and domain-specific manner to ensure the KG effectively represents and integrates domain-specific knowledge.

3. Open Research Problem, Hypothesis and Research Questions

The **Open Research Problem (ORP)** in which the research objectives are based is “Developing an automated framework to generate domain-specific KGs for the electronic health sector, aimed at assisting medical professionals in the decision-making process”.

In this line, this research work raises the following **Research Hypothesis (RH)**: “The integration of existing ontologies with LLMs enables the automatic creation of domain-specific KGs from Electronic Health Records (EHRs). This approach enhances the organization of information in the electronic health sector by (re)using ontologies to structure data under real-world contexts and leveraging LLM-based tools to information extract tasks, thereby streamlining the process of converting unstructured text into structured, actionable knowledge.”

This research hypothesis intends to solve next **Research Questions (RQs)**:

- **RQ1:** How to automatically extract a KG from diverse unstructured text in different languages?
- **RQ2:** How to effectively (re)use existent ontologies from the Health Sector for KG construction?
- **RQ3:** How to evaluate the generated domain-specific KG in a real-world healthcare decision-making scenario?

4. Objective and Sub-Objectives

Taking the ORP as **Research Objective (RO)**, the **Research Sub-objectives (SO)** would be:

- **SO1. Text to Graph Construction:** To design and implement methodologies for extracting KGs from unstructured, free-format text within the Electronic Health domain in different languages, using advanced NLP techniques.
- **SO2. (Re)Usability of Ontologies:** To develop and implement strategies for utilizing and adapting existing ontologies in the health sector to enhance the construction and scalability of domain-specific KGs, ensuring that these ontologies are effectively integrated to support accurate and context-relevant decision-making processes.
- **SO3. KG Evaluation metrics:** To establish metrics and evaluation protocols for assessing the quality dimensions of generated KGs in the Health Sector, ensuring they accurately represent the underlying data form EHRs.
- **SO4. Framework for automatic KG creation:** To ease the usability and scalability of the proposed framework across different datasets, ensuring that it can be effectively applied to various types of medical texts and decision-making scenarios

5. Research Methodology

Let's explore the **Research Methodology (RM)** crafted to achieve the outlined objectives and sub-objectives while testing the Research Hypothesis as we can see in figure 1:

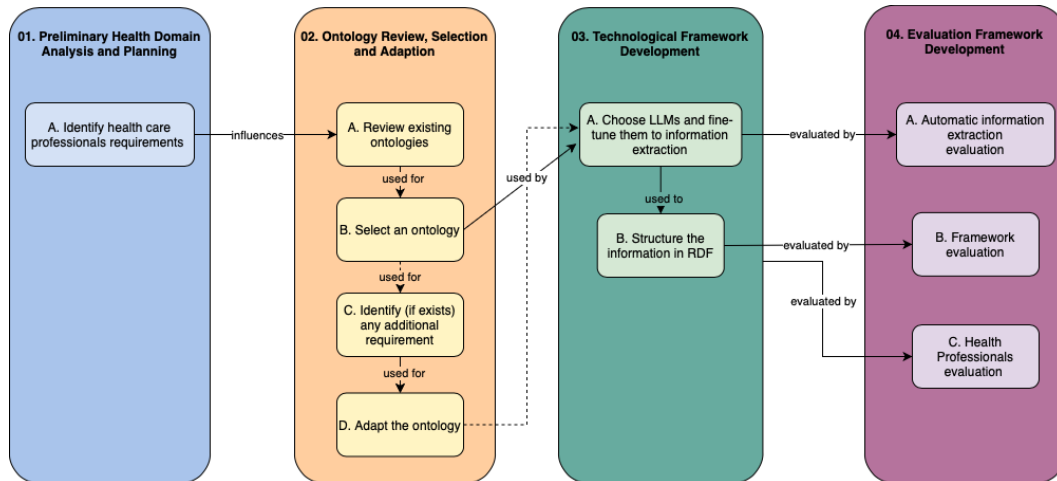


Figure 1: Iterative process of the work plan stages.

1. **Preliminary Health Domain Analysis and Planning:** In the initial stage, we have performed a preliminary analysis gathering requirements from the health professionals and analyzing EHRs authored by them. Some specific datasets explored yet include the E3C corpora [35], a multilingual dataset representing clinical histories in languages such as Spanish, Basque, English, Italian, and French. It is based on other corpora, including SPACCC, PubMed, and various other sources. It is related with SO1 and SO4.
2. **Ontology Review, Selection and Adaptation:** In this phase, we aim to (a) review existing ontologies that model the health domain to determine which one aligns with the previously established requirements, (b) select the most suitable ontology, and (c) identify (if exists) any additional specific requirements that are not yet addressed. Following this, we will (d) adapt the chosen ontology to comprehensively model the entire problem. It is related to SO2.
3. **Technological Framework Development:** In this step, with an ontology in place, we proceed to (a) choose LLMs and fine-tune them to incorporate the ontology knowledge. This helps in identifying text spans that correspond to ontology nodes and the relationships between these nodes. (b) After IE step, we structure this information into triplets using the RDF, thereby forming a formal and usable KG. It is related with SO1 and SO4.
4. **Evaluation Framework Development:** The final stage involves evaluating the entire workflow as well as assessing each individual IE task automatically. A test will be designed to be conducted where health professionals can input cases from their daily practice, and then apply the methodology to extract a KG specific to each case. It is related with SO1, SO3 and SO4.

6. Conclusions and Further Research

The challenges in KG construction derive from traditional methods for IE, reliance on pipelines that propagate errors at each step, the non-reuse of existing ontologies, and treating RE simply

as a classification task, which limits flexibility on KG semantics. Moreover, the limited interoperability of KGs due to the lack of non-English resources and the absence of a clear evaluation benchmark further complicate the KG construction process.

Then the goal of this work is to develop an end-to-end framework for constructing cross-lingual KGs in the medical domain, integrating advanced NLP techniques to automatically identify and align entities and relationships with predefined ontologies. This framework aims to further improve the KG construction process in healthcare data representation and ultimately improve patient care and clinical decision making in multiple languages.

As it is in an early stage, future works involve refining the NLP tasks with LLMs to handle more diverse data sources and complex medical terminologies in various languages and medical domains. Developing comprehensive evaluation benchmarks to assess the quality and performance of the constructed KGs. Additionally, collaboration with healthcare professionals will be essential to ensure that the framework meets the practical needs of health professionals and improves real-world medical outcomes.

Acknowledgments

This work has been funded by INESData (Infraestructura para la INvestigación de ESpacios de DATos distribuidos en UPM) project, from Ministerio para la Transformación Digital y de la Función Pública (PRTR, UNICO I+D CLOUD, EU NextGeneration).

I would like to thank my supervisors, Elena Montiel Ponsoda and Pablo Calleja Ibañez, and all the institutions that contribute to improve the daily life of people with health issues.

References

- [1] X. Zou, A survey on application of knowledge graph, in: *Journal of Physics: Conference Series*, volume 1487, IOP Publishing, 2020, p. 012016.
- [2] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, *Artificial Intelligence Review* (2023) 1–32.
- [3] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke, E. Rahm, Construction of knowledge graphs: State and challenges, *arXiv preprint arXiv:2302.11509* (2023).
- [4] X. Wang, X. He, Y. Cao, M. Liu, T.-S. Chua, Kgat: Knowledge graph attention network for recommendation, in: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 950–958.
- [5] N. Zhang, Q. Jia, S. Deng, X. Chen, H. Ye, H. Chen, H. Tou, G. Huang, Z. Wang, N. Hua, et al., Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba, in: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3895–3905.
- [6] X. Huang, J. Zhang, D. Li, P. Li, Knowledge graph embedding based question answering, in: *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 105–113.
- [7] R. Studer, V. R. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, *Data & knowledge engineering* 25 (1998) 161–197.

- [8] B. P. Allen, F. Ilievski, Standardizing knowledge engineering practices with a reference architecture, arXiv preprint arXiv:2404.03624 (2024).
- [9] T. R. Gruber, A translation approach to portable ontology specifications, *Knowledge acquisition* 5 (1993) 199–220.
- [10] N. Guarino, Formal ontology, conceptual analysis and knowledge representation, *International journal of human-computer studies* 43 (1995) 625–640.
- [11] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, *ACM Computing Surveys* 56 (2023) 1–62.
- [12] B. Jehangir, S. Radhakrishnan, R. Agarwal, A survey on named entity recognition–datasets, tools, and methodologies, *Natural Language Processing Journal* (2023) 100017.
- [13] Z. Xiaoyan, D. Yang, Y. Min, W. Lingzhi, Z. Rui, C. Hong, L. Wai, S. Ying, X. Ruifeng, A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers, arXiv preprint arXiv:2306.02051 (2023).
- [14] I. Melnyk, P. Dognin, P. Das, Grapher: Multi-stage knowledge graph construction using pretrained language models, in: *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- [15] C. E. Lipscomb, Medical subject headings (mesh), *Bulletin of the Medical Library Association* 88 (2000) 265.
- [16] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, *Nucleic acids research* 32 (2004) D267–D270.
- [17] K. Donnelly, et al., Snomed-ct: The advanced terminology and coding system for ehealth, *Studies in health technology and informatics* 121 (2006) 279.
- [18] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, D. Sontag, Learning a health knowledge graph from electronic medical records, *Scientific reports* 7 (2017) 5994.
- [19] Y. Zhang, M. Sheng, R. Zhou, Y. Wang, G. Han, H. Zhang, C. Xing, J. Dong, Hkgb: an inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians’ expertise incorporated, *Information Processing & Management* 57 (2020) 102324.
- [20] A. Rossanez, J. C. Dos Reis, R. d. S. Torres, H. de Ribaupierre, Kgen: a knowledge graph generator from biomedical scientific literature, *BMC medical informatics and decision making* 20 (2020) 1–24.
- [21] L. Murali, G. Gopakumar, D. M. Viswanathan, P. Nedungadi, Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study, *Journal of biomedical informatics* (2023) 104403.
- [22] M. Kejriwal, Knowledge graphs: A practical review of the research landscape, *Information* 13 (2022) 161.
- [23] A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, *ACM Computing Surveys (Csur)* 54 (2021) 1–37.
- [24] X. Wu, J. Duan, Y. Pan, M. Li, Medical knowledge graph: Data sources, construction, reasoning, and applications, *Big Data Mining and Analytics* 6 (2023) 201–217.
- [25] B. Abu-Salih, M. Al-Qurishi, M. Alweshah, M. Al-Smadi, R. Alfayez, H. Saadeh, Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities, *Journal of Big Data* 10 (2023) 81.

- [26] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, arXiv preprint arXiv:2305.13168 (2023).
- [27] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [28] I. Alimova, E. Tutubalina, Multiple features for clinical relation extraction: a machine learning approach, *Journal of biomedical informatics* 103 (2020) 103382.
- [29] M. Park, C. U. Jeong, Y. S. Baik, D. G. Lee, J. U. Park, H. J. Koo, T. Y. Kim, Screener: Streamlined collaborative learning of ner and re model for discovering gene-disease relations, *Plos one* 18 (2023) e0294713.
- [30] P.-L. H. Cabot, R. Navigli, Rebel: Relation extraction by end-to-end language generation, in: *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2370–2381.
- [31] E. Simperl, Reusing ontologies on the semantic web: A feasibility study, *Data & Knowledge Engineering* 68 (2009) 905–925.
- [32] M. Fernández-López, M. Poveda-Villalón, M. C. Suárez-Figueroa, A. Gómez-Pérez, Why are ontologies not reused across the same domain?, *Journal of Web Semantics* 57 (2019) 100492.
- [33] B. Xue, L. Zou, Knowledge graph quality management: a comprehensive survey, *IEEE Transactions on Knowledge and Data Engineering* 35 (2022) 4969–4988.
- [34] X. Wang, L. Chen, T. Ban, M. Usman, Y. Guan, S. Liu, T. Wu, H. Chen, Knowledge graph quality control: A survey, *Fundamental Research* 1 (2021) 607–626.
- [35] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanolini, The e3c project: European clinical case corpus, *Language* 1 (2021) L3.