Towards Generating High-Quality Knowledge Graphs by Leveraging Large Language Models

Morteza Kamaladdini Ezzabady[1], Frederic Ieng[2], Hanieh Khorashadizadeh[3], Farah Benamara[1,4], Sven Groppe[3], and Soror Sahri[2]

[1] IRIT, University of Toulouse, France
[2] Université Paris Cité, France
[3] University of Lübeck, Germany
[4] IPAL, CNRS-NUS-A*STAR, Singapore
{morteza.ezzabady,farah.benamara}@irit.fr
{frederic.ieng, soror.sahri}@u-paris.fr
{hanieh.khorashadizadeh,sven.groppe}@uni-luebeck.de

Abstract. Knowledge graph creation requires relation extraction (RE) tools often trained on annotated data either manually or by distant supervision. Recent approaches operate at the model level to handle new domains with unseen relations, relying on transfer learning or generative approaches in few/zero-shot learning scenarios. In this paper, we adopt a different strategy by operating instead at the level of dataset creation. We, for the first time to the best of our knowledge, investigate the ability of prompt-based models to build high-quality RE datasets relying on GPT4 to extract triples from sentences. Our approach is further enhanced by linking our knowledge graph to Wikidata, a step that enriches our dataset and ensures its interoperability. This strategy has been successfully employed in two use cases: COVID and health relation extraction.

Keywords: Knowledge Graph· Relation Extraction · Data Quality.

1 Introduction

Knowledge Graphs (KG) encode domain knowledge using a graph-based abstraction of knowledge where nodes are entities that can be real-world objects or abstract concepts and edges the relationships that represent the relation between these entities [15,13]. KGs have become a valuable resource for many downstream applications, such as reasoning and decision-making [21].

KG construction from texts heavily depends on the performances of entity recognition, entity linking, and relation extraction (RE) tools, the latter task being our focus here. To this end, RE datasets are crucial to train and evaluate RE classification models. Existing resources are either built manually following a predefined set of generic or domain-specific relations of interest [27] or by distant supervision leveraging knowledge bases such as Wikipedia pages or Wikidata [19]. While this last approach is an elegant solution to overcome human efforts

due to manual annotations, significant errors in automatic annotation may occur, leading to noisy training data, which may hurt models' precision [35].

To handle new domains with unseen relations, recent approaches operate *at the model level* relying on few/zero-shot learning strategies [10]. Recently, prompt-based models have been proposed as an alternative for entity-relation triples generation when evaluated on benchmarks RE datasets [5,6,42,32,16]. However, manual annotation is still needed to ensure wide-coverage data quality [3]. This is particularly salient when building domain-specific KGs in domains that lack annotated data and where the set of relations to be extracted has to be designed by experts [1].

To reduce the amount of supervision, we adopt a different strategy that addresses these shortcomings at the beginning of the process, i.e., *at the level of dataset creation*. We explore for the first time, as far as we know, the potential of recent advances in large language models (LLM) exemplified by GPT-4 [23] to answer the following questions: *Does a prompt-based approach work well for resource creation in domains with limited resources? Does this approach ensure the correctness and completeness of the generated KG? And more importantly, can this approach be transferred to other domains?* To this end, we design an iterative approach that first defines an initial domain-specific taxonomy of relations of interest and then generates triples from reliable sources while ensuring KG refinement. The core of our research lies in introducing a novel methodology that leverages LLMs to facilitate the entire KG creation process, from defining ontologies to annotating triples. It is important to note that our primary objective is not to build a comprehensive RE datasets but to show that (1) a generative approach for triple extraction is feasible, (2) a coherent set of triples can subsequently aid the RE task. Our contributions are as follows:

1. Investigate the ability of prompt-based models in building high-quality KG. We showcase the proposed approach on the specific use case of the COVID-19 domain in which a massive amount of unstructured (and potentially unreliable) data has been produced. However, most available KGs are (bio) medical-based with less effort on other aspects of the disease [17]. In addition, most current COVID-19-KGs rely on open RE without predefined relation taxonomy, showing many errors in RE prediction [14]. Therefore, RE COVID-19 datasets become primordial to improve the quality of KG and overcome the lack of evaluation standards [22]. Our approach results to **the first dataset annotated for COVID-19 relations**.[5]
2. Perform a qualitative intrinsic evaluation of our dataset relying on standard KG quality dimensions [12] that have never been used to evaluate a RE dataset as far as we know.
3. Demonstrate the utility of this novel resource for the task of relation classification. Our results show that prompt-based dataset creation achieves an accuracy of 87.43% compared to gold labels.
4. Show that the proposed approach is portable to a new domain while drastically reducing human efforts due to manual annotations.

[5] The annotated dataset will be available for research purposes upon request.

This paper is structured as follows. We first present related work on LLM-based approaches for resources creation with a focus on RE as well as knowledge graph validation. Section 4 presents our methodology for generating our knowledge graph. Section 5 presents the intrinsic evaluation while Section 6 our experiments on automatic relation classification. We discuss our results focusing in particular on qualitative error analysis together with a study of the portability of our approach to an unseen domain. We conclude this paper highlighting our main findings as well as limitations and directions for future work.

2 Related Work

2.1 LLMs for Resource Creation

The realm of synthetic data generation has witnessed transformative changes with the integration of LLMs, especially in contexts where authentic data is limited. Recently, LLMs have been used for generating training data in various NLP tasks, including database [25,2] and reasoning [26]. Some approaches augment existing datasets with automatically generated examples like ZeroShotDataAug [29] and AugGPT [8]. These approaches however require an already-existing labeled dataset. For example, when applied on the medical domain, AugGPT might yield inaccurate augmentations because most general-purpose LLMs lack specialized domain knowledge. The reader can refer to [18] for comprehensive overview of the use of LLMs as labeler or data generator.

2.2 LLMs for Relation Extraction

Relation extraction (which entails discerning semantic interrelations between textual entities) has been significantly augmented by the capabilities of LLMs. For example, [36] introduced two strategies to utilize LLMs for relation extraction: 1) in-context learning relying on prompts designed to consider task definition, relation labels, and entity types, and 2) data generation guided by instance descriptions along with some example instances. [31] evaluated the ability of large language models to perform few-shot relation extraction via in-context learning. They found out LLMs can achieve performance equivalent to SOTA methods by providing some demonstration examples. They augmented target RE labels with Chain of Thought (CoT) style explanations elicited from GPT-3 and used this to fine-tune Flan-T5. In [7], which is close to our work, the aim is to train a model for RE in the zero-shot setting. They proposed the RelationPrompt paradigm and showed that language models can effectively generate synthetic training data through relation label prompts to output triples. This is one difference with our work, because we extract triples from the text. Finally, [40] proposed LLMaAA as a new alternative to low-quality issues of the generated data but still requires hundreds of annotated examples to achieve good performances.

Compared to these works, and due to the lack of annotated resources (i.e., triples) in the COVID domain, our approach uses zero-shot learning, therefore

creating the KG from scratch, starting from ontology definition to triple annotation. The first advantage of our method is reducing the need for extensive pre-existing data, making it particularly beneficial in low resources domains. Furthermore, by automating the KG creation process, our approach minimizes the manual effort traditionally required in these tasks.

2.3 Data quality criteria for KG validation

Assessing KG is a crucial task to assure that the KG is exploitable, there are several quality metrics that can help assess KG quality [39]. The metrics can be grouped in different dimensions: accuracy, consistency, completeness, timeliness, trustworthiness, and availability [33]. Most metrics can be manually checked. However, it became tedious as the KG grew larger. Some work uses the sampling method instead of assessing the whole KG to make the assessment more realistically feasible [9]. In this paper, we manually assess every triplet to guarantee the highest quality assessment. We mainly focus on the most important issues: inaccurate triplet (accuracy) and incomplete triplet (completeness).

3 Methodology

Our methodology is automatic, in the sense that most parts of the process are driven by LLMs and guided by carefully crafted prompts. Manual inspection is needed but only at the validation step, which is essential to evaluate the effectiveness of our approach. We provide below a summary of the main steps, highlighting ones that rely on manual annotations:

1. *Ontology Definition*: This step generally requires a domain expert to design the main concepts and relations linking them. Although this approach will reduce the risk of mistakes, it is costly and experts are not always available. We therefore choose to rely on LLMs to build our initial ontology. This ontology is not evaluated per se, but the purpose of it was to guide the LLM to understand the domain we were focusing on.
2. *Ontology Expansion*: Utilizes the LLM alongside the preliminary ontology to extract triples from the data source.
3. *Ontology Pruning*: This phase emphasizes the refinement of relations, i.e., grouping relations that share similar characteristics or meanings under the same top-level relation. We streamline the ontology by merging semantically akin relations and eliminating those that are infrequent or nonsensical. Human intervention is mainly required in this stage. While some manual effort is essential, especially in the early stages of ontology building, it is notably less labor-intensive than starting data annotation from scratch. As we explain in depth in Section 4.3, because of $|relations| < |triples|$, we believe human efforts would be less than constructing a KG all manually.
4. *Intrinsic evaluation*: This step is not part of the process in itself and is only needed to report the quality of the generated triples.

5. *Extrinsic evaluation* via RE Task: We employ our generated dataset to fine-tune a relation classification model. This step shows how our dataset can be used in a classification task.

Steps 2 and 3 are applied to all the input data as an iterative process, i.e., we split our dataset into batches then step 2 and 3 are looped until all the batches have been processed. This iterative approach ensures that each instance receives individual attention, allowing for more precise adjustments and refinements to the ontology. Overall, only step 3 requires manual validation as part of the process. Manual quality evaluation (i.e., step 4) allows for a better linguistic analysis of the generated triples. We illustrate the three key steps of our methodology on the COVID-19 domain (see Figure 1 and Table 1), but it is generic enough to be applied to other domains as well. We detail below each of these steps.

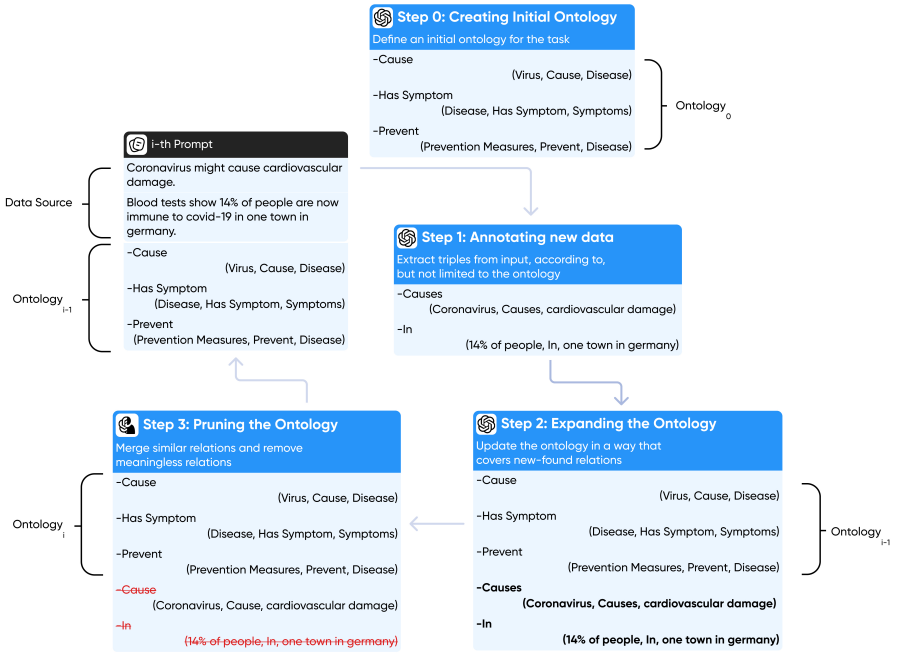

Fig. 1. KG Construction.

4 KG Construction

4.1 Ontology Definition

In traditional knowledge representation, an "ontology" typically refers to a structured framework that defines entities, concepts, and their inter-relationships

Table 1. Prompts and Responses.

Step	Prompt	Response
Initiate Ontology	I want to create a knowledge graph about COVID-19 facts and information. This knowledge graph contains triples (head entity, relation, tail entity). Build an ontology by defining possible relation and entity types.	... Entity types: - Virus - Disease - Symptoms ... Relation types: - Cause - Has_Symptom - Treat ...
Annotate Data	Now according to the ontology, for each sentence extract triples. [..., "Fujifilm tests favipiravir as covid-19 treatment", ...]	... (Fujifilm, Test, Favipiravir), (Favipiravir, Treat, Covid-19), ...
Expand Ontology	Update the ontology with the extracted relations.	... Relation types: - Cause - Has_Symptom - Treat - Test - In ...
Prune Ontology	Try to reduce the number of relations of ontology by grouping similar relations or changing them. Then update the ontology.	... Relation types: - Cause - Has_Symptom - Treat - Test ...
Augment Data	I'll give you a list of triples, generate a sentence for each representing that relation and containing both entities. [..., {'head': 'Moderna', 'tail': 'COVID-19 Vaccine', 'relation': 'develops'}, ...]	... Moderna, a biotechnology company, develops a COVID-19 Vaccine to help curb the spread of the virus. ...

within a specific domain. However, in our approach, we deviate from this strict definition. We categorize our triples solely based on their relation token, completely ignoring the types of each instance's entities. This flexibility is a deliberate design choice to accommodate the dynamic nature of the data we handle. Given the vast and evolving nature of data, especially in domains like healthcare, it is essential to have a system that can accommodate new entities and relationships without requiring significant restructuring. Our framework can seamlessly integrate new information by not strictly defining entity types for a given relation. For example:

1. Relation: "Treat"
   - Traditional Ontology: Medicine → Disease (e.g., "Paracetamol" treats "Fever")
   - Our Framework: Entity → Entity (e.g., "Paracetamol" treats "Fever" or "Exercise" treats "Depression")
2. Relation: "Cause"
   - Traditional Ontology: Disease → Symptom (e.g., "COVID-19" causes "Fever")
   - Our Framework: Entity → Entity (e.g., "COVID-19" causes "Fever" or "Loud Noise" causes "Hearing Loss")

4.2 Ontology Expansion

The second step is to define a first core ontology. Existing Covid-KGs have been built for different purposes, either from scientific articles or biological databases covering various relations such as disease-symptom, drug-drug or drug-disease interactions [4]. In our case, we do not assume any prior set of relations as we aim to cover a wide range of relations beyond a medical use case, including regulations, policies, and everyday statistics about the pandemic. To this end, we used GPT-4, prompting it to define an initial core ontology covering domain entities (e.g.," virus"," country"," policy") and the potential relationships between them. This results in 12 initial relations (e.g.," cause"," have symptom"," affect").

Guided by this core ontology, we deploy the generative model to automate the extraction of triples (e_1, R, e_2) (i.e., entities and their relations) from external textual data while extending it to cover new relations through an iterative process. The data was chosen according to two criteria: (a) cover a comprehensive range of COVID-19-related information reflecting various aspects of the pandemic, and (b) contain verified information to guarantee the quality of the information encapsulated in the KG.

Among the huge amount of existing sources about Covid, we rely on CovidFact [24], a dataset of 4,086 sentences, among them 1,296 are scientific and simplified media claims, manually annotated as evidence for the claims, and 2,790 automatically generated refuted claims. For example, from the claim "*⌈Alcoholism treatment⌋*[1] *is potentially effective against ⌈covid-19⌋*[2]", the triple `(e1, Affect, e2)` was generated, $e1$ (resp. $e2$) being the entities denoted by the linguistic expressions "Alcoholism treatment" (resp. "covid-19").

Our study aims to show that triple generation, along with domain ontology building, is feasible only when a small proportion of input data is provided. Hence, in comparison to benchmarks RE datasets (e.g., 106,264 sentences in TACRED [41]), we only relied on the 1,296 verified evidence (i.e., each one composed of one sentence with an average of 12 tokens) from CovidFact and arrived at 1,556 generated triples corresponding to 170 relations, among which 1,435 appear more than twice for a total of 61 relations.

4.3 Ontology Pruning

As the iterative expansion progresses, the ontology captures a wide range of relations, leading to a significant increase, totaling 170 relations. While this comprehensiveness is advantageous, it can pose challenges for subsequent classification and knowledge representation. Therefore, pruning and structuring these relations into a coherent taxonomy become imperative [11]. To this end, we employed generation together with manual review to combine relations that are semantically similar (such as "cause" and "lead to") while eliminating others. The discarded relations were either uninformative about the pandemic, infrequent, or comprised of illogical or incomprehensible connections, such as `(e1,`

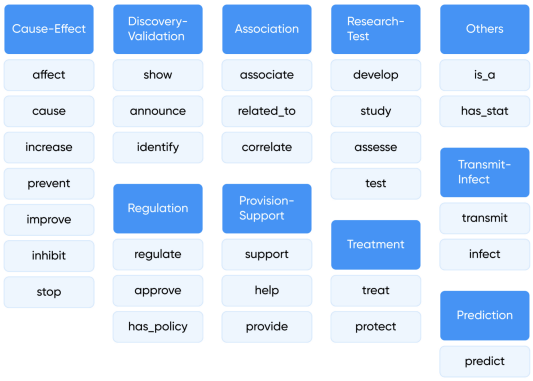

Fig. 2. CovidFact relations hierarchical structure.

`In, e2)`, generated from the statement "*Blood tests show [14% of people]*[1] *are now immune to covid-19 in [one town in Germany]*[2]".

Notably, the manual evaluation in this pruning step offers a more efficient approach than complete manual annotation. The focus shifts from extracting each triple to evaluating the relations, with $|relations| < |triples|$ streamlining the manual effort involved. The remaining relations have been hierarchically organized by grouping relations that share similar characteristics or meanings (e.g., "improve" and "increase") under the same top-level relation (e.g., "Cause-Effect"), as shown in Figure 2. Finally, since some relations are rare, we removed triples with their relations appearing in less than 25 instances to reduce noise due to extreme data imbalance. In the end, we arrived at a total of 1,164 instances corresponding to 10 top-level relations (see Table 2).

For a detailed account of the exact prompts used to interact with GPT-4 and the corresponding responses for different steps in our study, see Table 1.

5 KG Intrinsic Evaluation

Qualitative and rigorous evaluation is a significant challenge when constructing KGs [12]. Despite its importance, intrinsic evaluation has never been employed in NLP to evaluate a RE dataset. We rely on key determinant quality dimensions developed within the KG community [39], focusing, in particular, on those that can negatively impact the performances of RE models [34,28], namely:

(a) ***Semantic accuracy*** refers to the degree of correctness of facts in a KG [33]. For our cases, we have identified the following types of errors:

- *Incorrect entities* (Inc_E) are defined as errors in either the head or the tail of the triple. For example, in *"[stanford researchers]*[1] *test [3,200 people]*[2] *for covid-19 antibodies"* → `(e1, Tests, e2)`, the tail (e2) should be: "covid-19 antibodies" or "3,200 people for covid-19 antibodies".

- *Inaccurate relations* (INC_R) are defined as errors in the relation of the triple. For example, in *"[Dutch scientists]^1 find [new role for ACE2 receptors in Covid disease process]^2 "* → `(e1, Develop, e2)`, the relation "identify" should be more appropriate as "dutch scientists" did not develop anything but they identified "new role for ACE2 receptors in Covid disease process".
- *Incorrect triples* (INC_T) are defined as error in the whole triple, as in *"[Coronavirus vaccines]^1 leap through [safety trials]^2 "* → `(e1, Develop, e2)`, which suggests that coronavirus develops safety trials, which is incorrect.

(b) *Completeness* refers to the extent to which a KG covers the knowledge of interest [33]. For our case, we have identified the following issues:

- *Missing triples* (MIS_T): some information is not properly extracted; hence, some data is missing, as in *"[Antibody test]^1 for COVID-19 could help to [control virus spread]^2, says Singapore med-tech firm"*, where the triple `(e1, Helps, e2)` was not generated. This happens quite often for very long sentences.
- *Missing entities* (MIS_E): some information is only partially extracted and needs more information to be complete, like in *"[different mutations in SARS-CoV-2]^1 associate with severe and [mild outcome]^2"* where the underlined tokens are ignored for the triplet extraction. This often happens when multiple triples overlap on all but one element.

During the quality evaluation, we noticed that on the linguistic side, the most common errors concern entity extraction, especially when entities are acronyms or span over nonconsecutive tokens. Triples involving these entities (totaling 185) have been removed for the next classification experiments, leaving these complex cases for future work. We manually evaluated the remaining triples (970) in the next step according to these four quality dimensions. We observed that around 12.36% have quality issues split into correctness (6.54%) and completeness (5.82%): 3.12% are about INC_T, 1.25% INC_R while MIS_T concerns 5.1% of the triples vs. 0.52% for MIS_E. Overall, we observed that relation extraction in long sentences is more challenging. These observations align with those reported in generative KG construction [37].

It should also be noted that this step is not part of the process in itself and is only needed to report the quality of the generated triples.

6 KG Extrinsic Evaluation

We now show how the generated RE dataset can be used to train supervised RE classifiers. Following the general trend in the field, entities are given as input to the model. We rely on LUKE [38], a state-of-the-art RE model that we fine-tuned on our annotated dataset.

6.1 Relation Extraction on the Covid Domain

Evaluation Settings. Our dataset is highly imbalanced, with certain classes (or types of triples) being underrepresented (less than 20% of the triples). These minority classes posed a challenge to the model's learning process. To rectify this and ensure a balanced training set, we employed a data augmentation strategy. Specifically, we augmented all classes, expect the majority class and arrived at three training sets as follows:

- COV: Represents the original triples extracted using GPT-4. This setting serves as our baseline, capturing the raw knowledge extraction capability of GPT-4 without any augmentation or external data sources.
- GEN: Combines the original triples from COV with augmented data generated by GPT-4 (see the last row in Table 1).
- DB: Merges the original triples from COV with triples from an external data source. This setting provides a hybrid approach, leveraging both GPT-4's extraction capabilities and external, possibly domain-specific, knowledge. It uses a random set of 308 tuples from a publicly available structured database about the daily statistical updates and policy actions of different countries in response to the pandemic [20]. We then transform each tuple into coherent sentences, mimicking the data format found in the CovidFact dataset. For instance, from the tuple `(USA, 5000 new cases, 05-09-2023)`, the sentence "*On August 05, 2023, the USA reported 5000 new COVID-19 cases*" is generated.

Data in DB are regularly published by reputable health organizations and governmental bodies, which guarantees the reliability of the generated sentences. GEN resulted in 889 new sentences covering all the classes except "Cause-Effect", our majority class. DB, on the other hand, allows a significant augmentation of "Regulation" and "Others", which are among the less frequent (see Table 2).

Results and Error Analysis. Results are shown in Table 3 in terms of macro-averaged F1-score when tested on the COV test set, considering generated labels as gold. Best scores have been achieved by models trained on the augmented dataset, GEN being the most productive with 75.20. When compared to COV, the performances of all minority classes have been increased except for "Transmit-Infect" and "Prediction". This is more salient for "Treatment", "Regulation", and "Research-Test". DB achieved interesting results, particularly in classes like "Regulation" which has been considerably augmented. Although results on the class "Others" are less compared to COV, adding more instances from this class contributed to the other classes as well. These findings suggest incorporating reliable external structured data sources can be a reliable data augmentation strategy with less cost. A qualitative evaluation of the best classifier (i.e., GEN) shows that most errors occur between "Cause-Effect" and "Association", and "Research-Test" and "Discovery-Validation", suggesting potential linguistic similarities causing confusion.

Table 2. Original vs. augmented Train-Test split. We made the split randomly while keeping the same distribution of instances in each class.

	# Triples			Train			Test
	Cov	Gen	Db	Cov	Gen	Db	Cov
Cause-Effect	393	393	393	309	309	309	84
Disc.-Vali.	136	322	136	117	303	117	19
Association	109	287	109	78	256	78	31
Research-Test	103	258	103	87	242	87	16
Treatment	50	140	50	41	131	41	9
Others	43	111	200	35	103	153	8
Transmit-Infect	37	110	37	32	105	32	5
Provision-Supp.	37	84	37	26	73	26	11
Prediction	32	80	32	26	74	26	6
Regulation	30	74	181	25	69	152	5
Total	**970**	**1,859**	**1,278**	**776**	**1,665**	**1,021**	**194**

Table 3. Model performances on the COV test set. We repeated the process 5 times and these numbers are the average.

	Cov	Gen	Db
Cause-Effect	84.52	**87.65**	83.53
Discovery-Validation	75.00	80.95	**85.71**
Association	75.00	73.68	**75.41**
Research-Test	41.67	**71.43**	62.07
Treatment	31.58	**64.00**	35.29
Others	80.00	**87.50**	75.00
Transmit-Infect	61.54	55.56	**71.43**
Provision-Support	50.00	**66.67**	58.82
Prediction	**100.00**	90.91	92.31
Regulation	66.67	**90.91**	88.89
Macro F-score	66.60	**76.93**	72.85

To further evaluate our model, we manually labeled the whole 194 triples[6] from the Cov test set and then compared manual labels first with the ones generated by our approach, then with the ones predicted by Gen. We obtained an accuracy of 87.43% for GPT-4 vs. human and 81.15% for Gen vs. human.

In Table 4, we show examples from our dataset with labels that different settings predicted alongside the human annotation.

[6] Two of the authors of this paper annotated the data to have gold labels to evaluate performance of our approach to create a dataset. As we are not specialists we used common sense and if needed google search for this task.

Table 4. Examples from Test set with different model predictions. Blue and orange blocks indicate head and tail entities respectively.

	Human	GPT-4	LUKE (COV)	LUKE (GEN)
fda authorizes 15-minute coronavirus test.	Regulation	Regulation	Regulation	Regulation
high incidence of venous thromboembolic events in anticoagulated severe covid-19 patients.	Cause-Effect	Cause-Effect	Association	Cause-Effect
anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy.	Association	Cause-Effect	Cause-Effect	Association
polish doctor treats covid19 symptoms within 48h with amantadine.	Treatment	Treatment	Cause-Effect	Treatment
more than 80 clinical trials launch to test coronavirus treatments.	Research-Test	Research-Test	Research-Test	Research-Test

6.2 Portability to Unseen Domain

We now demonstrate the portability of our relation classifier to a new unseen domain, reducing therefore human efforts due to manual annotations. We rely on HealthFC [30], a curated collection of health claims designed explicitly for evidence-based medical fact-checking. Comprising 750 claims, this dataset offers a rich source of medical assertions that can be leveraged to test the robustness and adaptability of knowledge graph models in the medical domain. It is important to note that HealthFC was not part of the iterative phase and we just evaluated fine-tuned LUKE on our generated dataset (see previous section), on HealthFC.

From the HealthFC dataset, we meticulously extracted 109 triples. The extraction process was guided by the criteria of compatibility of the claims with our model's training data, which means relations of these selected triples could be categorized within our model's relations. For the experiment, we employed our best-performing model which is fine-tuned on GEN, results are presented on Table 5. We obtain an average F-score of 65.05%, the best performing was the majority relation Cause-Effect. Except Association which achieves low score (28.57%– which can be explained by the very low number of instances in the test set), all the remaining relations achieved very good score ($> 80.00\%$).

Table 5. Model performances on the unseen HealthFC dataset.

	# **Triples**	**Macro F-score**
Cause-Effect	76	93.06
Discovery-Validation	7	80.00
Association	1	28.57
Research-Test	1	00.00
Treatment	6	80.00
Others	4	85.71
Transmit-Infect	0	-
Provision-Support	14	88.00
Prediction	0	-
Regulation	0	-
Average	109	65.05

Overall, we can conclude that our approach can be a good alternative to alleviate the need of human annotation in a zero-shot evaluation scenario.

7 Future Work

Future work will explore the use of open-source models, which could offer significant improvements. While we focused on key determinant quality dimensions developed within the KG community, we also identified issues with entity consistency. Emphasis will be placed on entity linking to unify related entities under common concepts, thereby enhancing the knowledge graph's coherence and effectiveness. Additionally, Entity Evaluation will be a critical area of focus, as entities extracted by LLMs can sometimes be of low quality and include pseudo-sentences. Refining entity extraction methods to better capture complete contexts, especially in complex sentences, will be another important direction for improvement.

8 Conclusion

In this paper, we proposed a generative approach for RE dataset creation leveraging recent generative models. Our approach has been illustrated on the COVID-19 domain and evaluated relying on two quality dimensions criteria. We also demonstrated that with a very small subset of existing unstructured data, we can build a new dataset that can be used to train a robust state-of-the-art RE classifier. While our method involves some manual efforts, it is judiciously optimized for maximum effectiveness, representing a significant advancement in knowledge graph creation in under-resourced domains. However, there are also drawbacks. Our reliance on a specific LLM means that our results are contingent on the capabilities and limitations of that model. Additionally, while our method reduces manual labor, it doesn't eliminate it entirely, as human validation is still crucial, particularly in the ontology pruning phase. We believe these minimal efforts are still needed when dealing with new domains which lack good quality linguistic resources.

Our approach relies on external source of data (CovidFact claims) that have been used as input to prompt the model. Therefore our RE dataset can be potentially biased towards those claims. However, the approach has been designed to be domain-agnostic and can be easily transferred to other use cases for which reliable external sources are available, such as scientific papers, news articles, or encyclopedic knowledge (e.g., Wikipedia). Its generalization to generic as well as other domain-specific relations are important directions for future work.

Acknowledgments

This work is jointly funded by the French National Agency QualityOnt ANR-21-CE23-0036-01, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 490998901.

References

1. Abu-Salih, B.: Domain-specific knowledge graphs: A survey. Journal of Network and Computer Applications (2021)
2. Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., Kasneci, G.: Language Models are Realistic Tabular Data Generators (2023)
3. Cabot, P.L.H., Tedeschi, S., , Navigli, R.: RED^{FM}: a filtered and multilingual relation extraction dataset. In: ACL (2023)
4. Chen, C., Ebeid, I.A., Bu, Y., Ding, Y.: Coronavirus knowledge graph: A case study (2020)
5. Chen, X., Zhang, N., Xie, X., Deng, S., Yao, Y., Tan, C., Huang, F., Si, L., Chen, H.: Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In: Proceedings of the ACM Web Conference 2022 (2022)
6. Chia, Y.K., Bing, L., Poria, S., Si, L.: RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In: Findings of the Association for Computational Linguistics: ACL 2022 (2022)
7. Chia, Y.K., Bing, L., Poria, S., Si, L.: RelationPrompt: Leveraging Prompts to Generate Synthetic Data for Zero-Shot Relation Triplet Extraction. In: Findings of the Association for Computational Linguistics: ACL 2022 (2022)
8. Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., Li, X.: AugGPT: Leveraging ChatGPT for Text Data Augmentation (2023)
9. Gao, J., Li, X., Xu, Y.E., Sisman, B., Dong, X.L., Yang, J.: Efficient knowledge graph accuracy evaluation. arXiv preprint arXiv:1907.09657 (2019)
10. Gao, T., Han, X., Zhu, H., Liu, Z., Li, P., Sun, M., Zhou, J.: FewRel 2.0: Towards more challenging few-shot relation classification. In: EMNLP-IJCNLP (2019)
11. Han, X., Yu, P., Liu, Z., Sun, M., Li, P.: Hierarchical relation extraction with coarse-to-fine grained attention. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)
12. Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., Rahm, E.: Construction of knowledge graphs: State and challenges (2023)
13. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.d., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al.: Knowledge graphs. ACM Computing Surveys (CSUR) (2021)
14. Jaradeh, M., Singh, K., Stocker, M.e.a.: Information extraction pipelines for knowledge graphs. Knowledge Information Systems (2023)
15. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A survey on knowledge graphs: Representation, acquisition and applications (2020)
16. Jimenez Gutierrez, B., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., Su, Y.: Thinking about GPT-3 in-context learning for biomedical IE? think again. In: Findings of the Association for Computational Linguistics: EMNLP 2022 (2022)
17. Khorashadizadeh, H., Tiwari, S., Groppe, S.: A survey on covid-19 knowledge graphs and their data sources. In: Proceedings of the EAI International Conference on Intelligent Systems and Machine Learning (EAI ICISML 2022) (2022)
18. Lee, D.H., Pujara, J., Sewak, M., White, R., Jauhar, S.: Making large language models better data creators. In: EMNLP 2023. pp. 15349–15360 (2023)
19. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: A survey. Semantic Web (2018)
20. Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., Roser, M.: Coronavirus pandemic (covid-19). Our World in Data (2020)

21. Melnyk, I., Dognin, P., Das, P.: Knowledge graph generation from text. In: Findings of the Association for Computational Linguistics: EMNLP 2022 (2022)
22. Nguyen, H., Chen, H., Chen, J., Kargozari, K., Ding, J.: Construction and evaluation of a domain-specific knowledge graph for knowledge discovery. Information Discovery and Delivery (2023)
23. OpenAI: Gpt-4 technical report (2024)
24. Saakyan, A., Chakrabarty, T., Muresan, S.: COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (2021)
25. Schick, T., Schütze, H.: Generating Datasets with Pretrained Language Models. In: Proceedings of the EMNLP 2021 (2021)
26. Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., Chen, W.: Synthetic prompting: generating chain-of-thought demonstrations for large language models. ICML'23 (2023)
27. Stoica, G., Platanios, E.A., Póczos, B.: Re-TACRED: Addressing Shortcomings of the TACRED Dataset (2021)
28. Trajanoska, M., Stojanov, R., Trajanov, D.: Enhancing knowledge graph construction using large language models (2023)
29. Ubani, S., Polat, S.O., Nielsen, R.: ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT (2023)
30. Vladika, J., Schneider, P., Matthes, F.: HealthFC: A Dataset of Health Claims for Evidence-Based Medical Fact-Checking (2023)
31. Wadhwa, S., Amir, S., Wallace, B.: Revisiting Relation Extraction in the era of Large Language Models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2023)
32. Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., Kurohashi, S.: Gpt-re: In-context learning for relation extraction using large language models (2023)
33. Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., Wu, T., Chen, H.: Knowledge graph quality control: A survey. Fundamental Research (2021)
34. Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., Wu, T., Chen, H.: Knowledge graph quality control: A survey. Fundamental Research (2021)
35. Xie, C., Liang, J., Liu, J., Huang, C., Huang, W., Xiao, Y.: Revisiting the negative data of distantly supervised relation extraction. CoRR (2021)
36. Xu, X., Zhu, Y., Wang, X., Zhang, N.: How to Unleash the Power of Large Language Models for Few-shot Relation Extraction? (2023)
37. Xu, X., Zhu, Y., Wang, X., Zhang, N.: How to unleash the power of large language models for few-shot relation extraction? arXiv preprint arXiv:2305.01555 (2023)
38. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: Deep contextualized entity representations with entity-aware self-attention. In: EMNLP 2020 (2020)
39. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., Hitzler, P.: Quality assessment methodologies for linked open data. Submitted to Semantic Web Journal (2013)
40. Zhang, R., Li, Y., Ma, Y., Zhou, M., Zou, L.: LLMaAA: Making large language models as active annotators. In: EMNLP 2023. pp. 13088–13103 (2023)
41. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017) (2017)
42. Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., Zhang, N.: Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities (2023)