frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 27 August 2024
DOI 10.3389/frai.2024.1460065

OPEN ACCESS

EDITED BY
Aleksandr Raikov,
National Supercomputer Center, China

REVIEWED BY
Stefano Silvestri,
National Research Council (CNR), Italy

*CORRESPONDENCE
Luca Mariotti
luca.mariotti@unimore.it

RECEIVED 05 July 2024
ACCEPTED 09 August 2024
PUBLISHED 27 August 2024

CITATION
Mariotti L, Guidetti V, Mandreoli F, Belli A and Lombardi P (2024) Combining large language models with enterprise knowledge graphs: a perspective on enhanced natural language understanding. *Front. Artif. Intell.* 7:1460065. doi: 10.3389/frai.2024.1460065

Combining large language models with enterprise knowledge graphs: a perspective on enhanced natural language understanding

Luca Mariotti[1]*, Veronica Guidetti[1], Federica Mandreoli[1], Andrea Belli[2] and Paolo Lombardi[2]

[1]Department of Physics, Informatics and Mathematics, Università di Modena e Reggio Emilia, Modena, Italy, [2]Expert.ai, Modena, Italy

Knowledge Graphs (KGs) have revolutionized knowledge representation, enabling a graph-structured framework where entities and their interrelations are systematically organized. Since their inception, KGs have significantly enhanced various knowledge-aware applications, including recommendation systems and question-answering systems. Sensigrafo, an enterprise KG developed by Expert.AI, exemplifies this advancement by focusing on Natural Language Understanding through a machine-oriented lexicon representation. Despite the progress, maintaining and enriching KGs remains a challenge, often requiring manual efforts. Recent developments in Large Language Models (LLMs) offer promising solutions for KG enrichment (KGE) by leveraging their ability to understand natural language. In this article, we discuss the state-of-the-art LLM-based techniques for KGE and show the challenges associated with automating and deploying these processes in an industrial setup. We then propose our perspective on overcoming problems associated with data quality and scarcity, economic viability, privacy issues, language evolution, and the need to automate the KGE process while maintaining high accuracy.

KEYWORDS

LLMS, knowledge graph, relation extraction, knowledge graph enrichment, AI, enterprise AI, carbon footprint, human in the loop

1 Introduction

A Knowledge Graph (KG) represents real-world knowledge using a graph structure, where nodes denote entities and edges represent relationships between them (Hogan et al., 2021). Since Google introduced the Knowledge Graph in 2012, KGs have become essential in knowledge representation, enhancing various tasks Companies use them to improve product performance, boosting data representation and transparency in recommendation systems, efficiency in question-answering systems, and accuracy in information retrieval systems (Peng et al., 2023).

This work presents the perspective of Expert.AI, a leading enterprise in Natural Language Understanding (NLU) solutions, centered on meticulously created and curated KGs by expert linguists. While manual curation ensures high precision and data quality, it demands significant human effort, and the rapid evolution of real-world knowledge requires frequent updates to KGs.

Recent advancements in Large Language Models (LLMs) suggest potential for partial automation of this process. LLMs, deep learning architectures designed for natural language processing, have demonstrated impressive results in NLU tasks. Their advanced capabilities represent a promising avenue for automating and enhancing Knowledge Graph Enrichment (KGE), refining, and adding new entities and relationships in KGs. By leveraging the implicit knowledge embedded within pre-trained LLMs (PLMs), companies can streamline the identification of new entities and relationships in external corpora, enriching their KGs with minimal manual intervention (Valmeekam et al., 2024).

However, automating KGE from external text in an industrial context is far from straightforward. It is crucial to choose an appropriate methodological framework: various PLM-based KGE techniques require model finetuning, while others rely on prompting. We will discuss the advantages and disadvantages of both approaches. For instance, while finetuning is generally costly and requires large amounts of annotated data, prompting is more cost-effective but poses privacy-related risks.

We will also examine the primary challenges of implementing corporate KGE solutions, categorizing them into four areas: (i) the quality and quantity of public or automatically annotated data, (ii) developing sustainable solutions in terms of computational resources and longevity, (iii) adaptability of PLM-based KGE systems to evolving language and knowledge, and (iv) creating models capable of efficiently learning the KG structure.

We review existing solutions for each issue and identify promising options for automating KGE in industrial settings using PLMs while maintaining high quality. We recommend a hybrid approach that combines PLMs, KG structure understanding, and domain expertise, ensuring privacy compliance. To adapt to evolving LLMs, we suggest treating PLMs as plug-and-play components for versatility and longevity.

This paper is structured as follows: Section 2 presents Expert.AI and its research investment objectives. Section 3 discusses the state-of-the-art in PLM-based KGE. Section 4 provides our perspective on the challenges of deploying these methods in an enterprise environment. Finally, conclusions are drawn in Section 5.

2 Sensigrafo: an enterprise KG and its characteristics

Expert.AI, formerly known as Expert System, is a leading AI enterprise specializing in solving complex language challenges. With over 300 natural language solutions, Expert.AI has transformed language-intensive processes across various sectors. Central to Expert.AI's NLU solutions is a collection of large KGs called *Sensigrafo*s, meticulously built by linguists and domain experts and carefully modified to gain performance in downstream NLU tasks. [1] Each Sensigrafo includes attributes like grammatical role, semantic relation, definition/meaning, domain, and frequency that establish the characteristics of words and concepts (Buzzega et al., 2023). Terms with the same meaning are grouped into syncons, interconnected by millions of logical and linguistic links, organized hierarchically. For example, the English Sensigrafo contains about 440,000 syncons, grouping more than 580,000 words, and 80+ relation types, yielding around 7.1 million links.

In contrast, most collaborative open-source KGs are generated automatically, resulting in numerous triples. DBpedia, for instance, contains about 900 million triples. The number of entity classes varies across KGs, with Wikidata having over 110 million items and 500 million facts, and YAGO encompassing knowledge of more than 67 million entities and 343 million facts (Suchanek et al., 2023). The number of relation types also varies, with Freebase having 1,345 and YAGO holding only 140 (Suchanek et al., 2023). These KGs span diverse domains, primarily derived from text corpora like Wikipedia, aiming to cover extensive real-world knowledge. Conversely, each Sensigrafo is carefully constructed using only information sources from its intended domain, making the information extraction operation much more reliable and accurate.

However, the accuracy of Sensigrafo's information comes at a high maintenance cost. Adding new syncons and relations requires full human supervision, aided by Expert.AI's semantic engine, Cogito. Cogito uses a Sensigrafo to resolve ambiguities related to word meanings and can expand its knowledge through expert input or analyzing tagged content using ML algorithms.

As real-world information grows and the cost of upgrading Sensigrafo increases, Expert.AI plans to integrate symbolic and statistical technologies, combining expert-validated rules with AI methods to automate Sensigrafo updates. This hybrid approach is expected to reduce the costs of developing and maintaining symbolic AI solutions. Nevertheless, any AI solution should be accompanied by a high degree of explainability, robustness, and precision to make enrichment systems transparent and reliable.

To identify the crucial aspects in developing such a solution, we will present state-of-the-art KGE techniques based on PLMs.

3 Pretrained LLM for KG management and enrichment

Relation extraction (RE) and named entity recognition (NER) are key challenges in automatic KGE. RE identifies and categorizes relationships between entities in unstructured text, expanding the KG's structure. NER focuses on recognizing, classifying, and linking entities in the text to the knowledge base. These processes are crucial for accurately identifying entities and their interconnections, enhancing KGs. Recent literature highlights two approaches to NER and RE: creating large training sets with hand-curated or extensive automatic annotations to fine-tune LLMs, or using precise natural language instructions, replacing domain knowledge with prompt engineering efforts (Levy et al., 2017; Li et al., 2019; Soares et al., 2019; Peng et al., 2020; Agrawal et al., 2022; Wang et al., 2023).

Supervised methods for NER and RE usually include a pretraining stage followed by zero-shot learning (Wang et al., 2022) or the use of specialized architectures and training setups (Yu et al., 2020; Li et al., 2022b). Due to the lack of large annotated corpora, many approaches for RE and NER rely on distant supervision (DS), an automated data labeling technique that aligns knowledge bases with raw corpora to produce annotated data.

1 https://www.expert.ai/products/expert-ai-platform/knowledge-graph/

Early DS approaches to RE use supervised methods to align positive and negative pair relations for pre-training language models, followed by few-shot learning to extract relations (Soares et al., 2019; Peng et al., 2020). DS methods for NER involve tagging text corpora with external knowledge sources like dictionaries, knowledge bases, or KGs. A common DS method for NER is the teacher-student framework. For example, Liang et al. (2020) proposed a two-stage method: fine-tuning a LLM on DS labels, followed by teacher-student system self-training with pseudo soft labels to improve performance.

While DS is useful when labeled data is scarce or expensive, it can introduce incomplete and inaccurate labels. To address this, recent works have focused on mitigating DS label noise and improving results (Wan et al., 2022). A common method to address DS noise in RE is multi-instance learning (MIL) (Zeng et al., 2015), which groups sentences into bags labeled as positive or negative with respect to a relation, shifting the RE task from single sentences to bags. However, MIL is not data-efficient, leading to recent extensions into contrastive learning setups. These efforts aim to cluster sentences with the same relational triples and separate those with different triples in the semantic embedding space (Chen et al., 2021; Li et al., 2022a).

Recent years have seen a significant increase in work on NER and RE involving prompt engineering. Prompting for NER includes using entity definitions, questions, sentences, and output examples to guide LLMs in understanding entity types and extracting answers (Ashok and Lipton, 2023; Kholodna et al., 2024). For RE, tasks are rephrased as question-answering (Levy et al., 2017), often injecting latent knowledge contained in relation labels into prompt construction (Chen et al., 2022) and iteratively fine-tuning prompts to enhance the model's ability to focus on semantic cues (Son et al., 2022). In general, zero-shot learning methods have been shown to perform better than supervised settings when the amount of training data is scarce.

Choosing between prompt engineering and fine-tuning is challenging. While prompting with large LLMs like GPTs is appealing for handling complex tasks with minimal data annotation, it can underperform in NER compared to fine-tuned smaller PLMs like BERT derivations, especially with more training data (Gutierrez et al., 2022; Keloth et al., 2024; Pecher et al., 2024; Törnberg, 2024). Large LLMs, such as GPT-3, struggle with specific information extraction tasks, including managing sentences without named entities or relations (Gutierrez et al., 2022). Prompting also faces hallucination issues, often overconfidently labeling negative inputs as entities or relations. Some approaches, such as Wang et al. (2023), address this by enriching prompts and reducing hallucinations via self-verification strategies. Other methods correct inaccurate NER and RE prompting results through active learning techniques (Wu et al., 2022) or by distilling large PLMs into smaller models for specific tasks (Agrawal et al., 2022).

4 Perspective

Summarizing the previous sections, the main challenges for enterprise LLM-based solutions for KGE include:

- *Computational resources and longevity:* creating tailored PLM-based KGE solutions can be costly and resource-intensive. There is a need for lightweight, sustainable, and durable training pipelines.
- *Data quality and benchmarking:* collaborative and Enterprise KGs have different structures, causing a mismatch between public benchmark datasets and enterprise use cases.
- *Evolving knowledge:* robust methods are needed to combine automated novelty detection (new links and nodes for the KG) with high-quality human-curated interventions.
- *Lack of adaptive hidden representations:* the learning paradigm should shift from classification to representation learning to accommodate novelty and efficiently encode KG features.

In the following sections, we will provide a comprehensive analysis of each of these challenges.

4.1 Computational resources and longevity of solutions

When developing enterprise-level NLU solutions, it's crucial to consider computational resources and carbon footprint due to the high environmental and economic costs of traditional model training (Patil and Gudivada, 2024). Fully fine-tuning PLMs, while effective for specific tasks, is often costly and inefficient, requiring substantial computational resources and time. These models are tailored for narrow applications, making updates challenging (Razuvayevskaya et al., 2023).

In contrast, in-context learning provides greater flexibility, facilitating adaptation to the rapidly evolving field of LLMs. However, prompt engineering is time-consuming and requires methods not universally applicable across models (Zhao et al., 2024). Balancing these factors is essential for creating sustainable and effective NLU solutions that meet the dynamic requirements of modern enterprises (Faiz et al., 2023).

Given the continuous release of new LLMs, we advocate for PLM-based KGE approaches that treat the LLM as a modular component, easily replaceable to integrate context-specific models trained on domain-specific knowledge, enhancing system relevance and accuracy.

The choice between fine-tuning and in-context solutions is closely tied to selecting an encoder or decoder architecture for NLU. The need for regularly updated tools favors encoder-based solutions. Generative models like ChatGPT, while user-friendly, can quickly become outdated or change unexpectedly, compromising the reproducibility and efficiency of prompting techniques (Törnberg, 2024). Additionally, the opacity of their training data makes these models less reliable in zero-shot scenarios. Ethical and legal considerations further limit the use of proprietary generative LLM APIs with private or confidential data, making them unsuitable for most enterprise environments (Törnberg, 2024). Prompt engineering for full KGE is also impractical due to the structural mismatch between natural language and KGs, complicating the creation of satisfactory, automated prompts for large KGs beyond simple proof-of-concept examples.

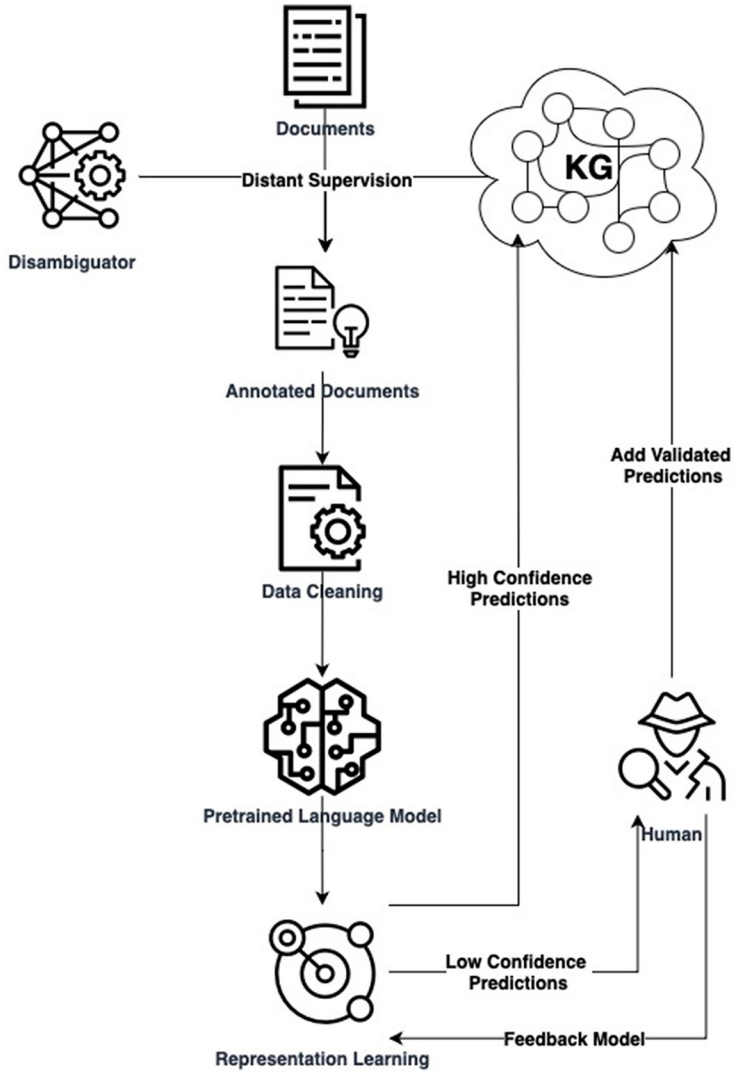


FIGURE 1
Flowchart illustrates the integration of human feedback in the Expert.AI process, from dataset preparation and disambiguation to knowledge graph querying, data processing, and representation learning.

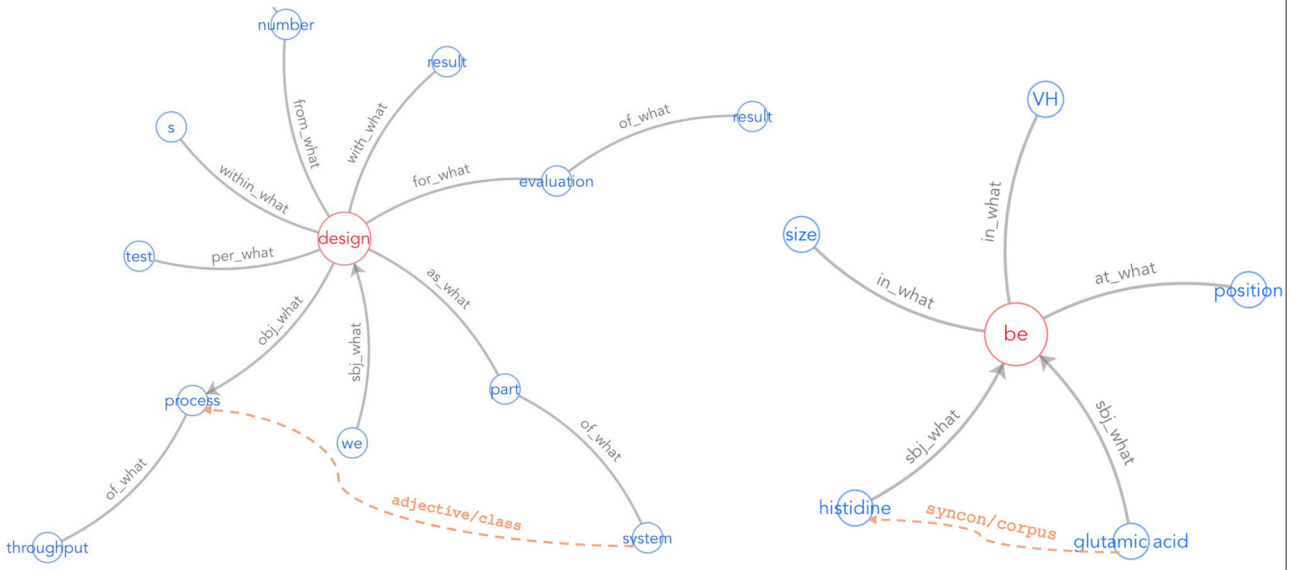
Thus, we advocate for adapter-based fine-tuning for efficient KGE solutions. Instead of modifying all LLM weights, this approach adds a small network to the existing encoder architecture and trains that module. Adapters are trained on task-specific data, while the original model's weights remain mostly unchanged, acting as prior knowledge. This method is more computationally efficient, allowing LLMs to be plug-and-play components in the data pipeline, making the system more flexible and easily updated. This approach would manage the carbon footprint of extensive computational processes and extend the lifespan of KGE solutions.

4.2 Data quality and solution benchmarking

As mentioned, all supervised methods for KGE require creating large, annotated datasets. While leveraging benchmark datasets from literature would be ideal, most of these datasets are built from collaborative KGs. This can pose challenges and performance depletion when trying to export solutions to sparser KGs for NLUs. Additionally, the quality and properties of annotated corpora are significant concerns as manual annotation, the most reliable source, is scarcely available.

According to Bassignana and Plank (2022), cross-dataset and cross-domain setups for RE are particularly deficient. To combat data sparsity, several semi-automatically labeled datasets have been constructed, but they have issues, such as missing relation labels (NA). For example, the NYT10d dataset has 53% incorrect labels, while NYT10m and Wiki20m have 96 and 60% of triples labeled as "NA" (Gao et al., 2021; Lin et al., 2022). Datasets defined as manually annotated often only include human annotations in the test set.

Also popular NER datasets show limitations, such as the limited number of entity classes. For instance, the CoNLL 2003 dataset contains only four entity types, ACE 2005 has



As part of our diagnostic system, we designed a custom Google Glass RDT reader application for capturing RDT images, a Web portal for generating custom QR code-based identifiers and viewing test results, and the server processes for rapid and high-throughput digital evaluation of test results from any number of devices, with results ready within 8 s per test (see Figures 3 and 4).
TRUE RELATION: adjective/class
Relation ranking:
76.67% adjective/class
5.4% syncon/corpus
5.36% verb/object.
At position 56 in VH, glutamic acid (E) and histidine (H) are very similar in size, but the E56H substitution may cause side-chain changes of D368 in Env and form more stable hydrogen bonds with Env for more stable binding.
TRUE RELATION: syncon/corpus
Relation ranking:
35.98% syncon/corpus
24.96% supernomen/subnomen
12.81% noun + of/noun
9.67% adjective/class.

FIGURE 2
Sample of predicted links in a text. Results of Expert.ai Visualization Tool for RE, represented as solid lines. They show the relations expressing the semantic role of the terms (blue) that are connected to a verb (red). The LLM-based RE framework predicts the relations (ranked in confidence order) shown in the text boxes and the most probable relations are depicted in the corresponding Sensigrafo portions as dashed lines.

7, and Ontonotes v5 includes 18 entities (Zhang and Xiao, 2024). This scarcity challenges the extraction of diverse entity types for KGs. This can cause robustness problems leading to poor generalization in out-of-domain scenarios (Ma et al., 2023). Moreover, most NER datasets are not constructed from a KG, failing to capture complex KG structures and relationships, which affects the quality and completeness of extracted entities.

Given these challenges, generating KG-centered datasets via DS appears to be the safest choice for custom KGE solutions. However, DS can introduce errors due to its reliance on assumptions that are not always valid (Riedel et al., 2010), especially when KGs and the corpus do not align closely, leading to hallucinations. Furthermore, DS principles struggle to accommodate the evolving nature of knowledge in free texts, as text annotation is based on a static, pre-existing KG.

4.3 Ever evolving knowledge and LLMs

Maintaining and updating NLU solutions must account for the evolving nature of language and knowledge. KGE relying solely on DS may be inadequate, as weak annotations come from existing KG entities and relations, limiting the prediction of new types. Enterprises require precise solutions and cannot rely solely on self-/unsupervised tools, necessitating some level of human curation in KG updating methods.

This need can be addressed using the human-in-the-loop (HITL) paradigm, which integrates human expertise into the modeling process to manage ML models uncertainty (Wu et al., 2022). In NLU, HITL methods iteratively correct or predict text annotations. Typically, this involves starting with a small set of annotated data (human-curated or weakly labeled), selecting challenging samples for the model, having humans annotate these

samples, updating the model with the new annotations, and repeating the process.

HITL effectively handles scarce or sparse data for NER (Shen et al., 2017), can address mislabeling (Muthuraman et al., 2021) and enhance different stages of the ML pipeline, such as data processing, model training, and inference (Zhang et al., 2019; Klie et al., 2020; Wu et al., 2022). Moreover, this paradigm was already successfully employed to dynamically curate and expand databases based on subject matter expert feedback (Gentile et al., 2019). Although a comprehensive HITL method for KGE does not yet exist, Qian et al. (2020) provides a promising starting point. The authors focus on disambiguating entity names with various textual variations using non-annotated examples, DS to generate pseudo labels, and active learning to address DL models' data requirements. They rank predictions based on model confidence and involve users in labeling the top and bottom elements. This framework could be extended to simultaneously handle KG entities and relations, engaging with the full KG structure.

4.4 Need for adaptive hidden representations

As previously described, KGE subtasks are often modeled as classification problems, which pose several issues.

Modeling NER or RE as classification outcomes forces the model to predict an entity or a relation, leaving little room for uncertainty. This is problematic especially given the risk of conceptual hallucinations in LLMs leading to false positives (Peng et al., 2022), an undesirable feature in non-transparent models that can compromise disambiguation tasks.

Furthermore, modeling KGE as a classification problem prevents the correct handling of KGs where multiple relations connect two entities. This affects both disambiguation, which must identify the correct triple in a sentence, and link prediction, which aims to detect the appropriate relation.

KGs are dynamic, frequently updating entities and relationships. While HITL can address this, systems must incorporate new classes or modify existing ones. Classification tasks constrain outcomes by a fixed structure, preventing real-time adaptation to evolving KGs and necessitating full retraining when new relations or entity classes are added, making the process inefficient.

We advocate for designing ML models for downstream tasks to consider KG structures. Instead of classification, representation learning methods should be used to minimize noise impact and manage uncertainty and evolving output structures. Methods like contrastive learning can mimic and learn the principles of symbolic KGs and disambiguation systems, leading to a consistent and dynamic deep-learning approach to KGE.

4.5 Outlining the process: a simplified pipeline for expanding knowledge graph relations

Our proposed KGE solution for enterprises involves creating customized datasets via DS, using lightweight supervised representation learning, and integrating human feedback for high-quality updates. Figure 1 illustrates this operational pipeline. Such a pipeline aligns with explainable AI in NLU, addressing computational efficiency, data quality, evolving knowledge, and adaptive representations simultaneously. To illustrate these steps, consider the task of enriching the life sciences-oriented Sensigrafo through RE. We use a collection of PubMed[2] documents. Entities in the text are marked leveraging Cogito, the Expert.ai's disambiguator, and a DS module grounded on Sensigrafo produces the possible relations in the annotated documents. We select a field-specific PLM, such as BioBERT (Lee et al., 2020). Annotated documents are transformed into contextualized embeddings using the PLM as a prior knowledge base. A small neural network added to the PLM performs adapter-based fine-tuning for RE using techniques like contrastive learning. After training, the model can recognize relations between marked entities in free text, with predictions ranked by model confidence. Figure 2 shows two RE use cases in text boxes. High-confidence predictions are injected into the KG, while low-confidence ones are reviewed by domain experts. Experts validate model results, insert new relations into the KG, and provide feedback by adding new data to the training set. They also assess data quality and identify potential disambiguation mistakes.

5 Conclusions

Integrating LLM solutions into enterprise environments reliant on KGs holds great potential for automated and data-driven maintenance and updates. Drawing on the experience of Expert.AI, a leader in NLU solutions, we identify critical issues in current approaches and outline future challenges. Key aspects to address include data quality, computational resources, the role of human expertise, and choosing the right technique to machine-learn the foundational principles of KG construction. Future efforts should aim to develop resilient frameworks that blend automated and human-involved processes, ensuring business applications of LLMs are effective, efficient, and sustainable.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LM: Investigation, Methodology, Writing – original draft, Writing – review & editing. VG: Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. FM: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. AB: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. PL: Conceptualization, Funding

2 https://pubmed.ncbi.nlm.nih.gov/

acquisition, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This paper was partially funded by Region Emilia-Romagna through the LR 14 year 2021 Project "IbridAI-Hybrid approaches to Natural Language Understanding".

Acknowledgments

We would like to extend our sincere gratitude to Expert.ai for their invaluable support and contributions to this research.

Conflict of interest

AB and PL were employed by Expert.ai.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. *arXiv [Preprint]*. arXiv:2205.12689. doi: 10.48550/arXiv.2205.12689

Ashok, D., and Lipton, Z. C. (2023). Promptner: prompting for named entity recognition. *arXiv [Preprint]*. arXiv:230515444. doi: 10.48550/arXiv.230515444

Bassignana, E., and Plank, B. (2022). "What do you mean by relation extraction? A survey on datasets and study on scientific relation classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, eds. S. Louvan, A. Madotto, and B. Madureira (Dublin: Association for Computational Linguistics), 67–83. doi: 10.18653/v1/2022.acl-srw.7

Buzzega, G., Guidetti, V., Mandreoli, F., Mariotti, L., Belli, A., Lombardi, P., et al. (2023). "Automated knowledge graph completion for natural language understanding: Known paths and future directions," in *CEUR Workshop Proceedings*, Vol. 3478 (CEUR-WS), 160–172.

Chen, T., Shi, H., Tang, S., Chen, Z., Wu, F., Zhuang, Y., et al. (2021). Cil: contrastive instance learning framework for distantly supervised relation extraction. *arXiv [Preprint]*. arXiv:2106.10855. doi: 10.4550/arXiv.2106.10855

Chen, X., Zhang, N., Xie, X., Deng, S., Yao, Y., Tan, C., et al. (2022). "Knowprompt: knowledge-aware prompt-tuning with synergistic optimization for relation extraction," in *Proceedings of the ACM Web Conference 2022* (New York, NY: ACM), 2778–2788. doi: 10.1145/3485447.3511998

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., et al. (2023). Llmcarbon: modeling the end-to-end carbon footprint of large language models. *arXiv [Preprint]*. arXiv:2309.14393. doi: 10.48550/arXiv.2309.14393

Gao, T., Han, X., Bai, Y., Qiu, K., Xie, Z., Lin, Y., et al. (2021). "Manual evaluation matters: Reviewing test protocols of distantly supervised relation extraction," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, eds. C. Zong, F. Xia, W. Li, and R. Navigli (Association for Computational Linguistics), 1306–1318. doi: 10.18653/v1/2021.findings-acl.112

Gentile, A. L., Gruhl, D., Ristoski, P., and Welch, S. (2019). "Explore and exploit. dictionary expansion with human-in-the-loop," in *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings 16* (Cham: Springer), 131–145. doi: 10.1007/978-3-030-21348-0_9

Gutierrez, B. J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., et al. (2022). Thinking about GPT-3 in-context learning for biomedical IE? Think again. *arXiv [Preprint]*. arXiv:2203.08410. doi: 10.48550/arXiv.2203.08410

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., et al. (2021). Knowledge graphs. *ACM Comput. Surv.* 54:71. doi: 10.1145/3447772

Keloth, V. K., Hu, Y., Xie, Q., Peng, X., Wang, Y., Zheng, A., et al. (2024). Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics* 40:btae163. doi: 10.1093/bioinformatics/btae163

Kholodna, N., Julka, S., Khodadadi, M., Gumus, M. N., and Granitzer, M. (2024). Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. *arXiv [Preprint]*. arXiv:2404.02261. doi: 10.48550/arXiv.2404.02261

Klie, J.-C., de Castilho, R. E., and Gurevych, I. (2020). "From zero to hero: human-in-the-loop entity linking in low resource domains," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds. D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, (Association for Computational Linguistics), 6982–6993. doi: 10.18653/v1/2020.acl-main.624

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240. doi: 10.1093/bioinformatics/btz682

Levy, O., Seo, M., Choi, E., and Zettlemoyer, L. (2017). Zero-shot relation extraction via reading comprehension. *arXiv [Preprint]*. arXiv:1706.04115. doi: 10.48550/arXiv.1706.04115

Li, D., Zhang, T., Hu, N., Wang, C., and He, X. (2022a). Hiclre: a hierarchical contrastive learning framework for distantly supervised relation extraction. *arXiv [Preprint]*. arXiv:2202.13352. doi: 10.48550/arXiv.2202.13352

Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., et al. (2022b). Unified named entity recognition as word-word relation classification. *Proc. AAAI Conf. Artif. Intell.* 36, 10965–10973. doi: 10.1609/aaai.v36i10.21344

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J., et al. (2019). A unified mrc framework for named entity recognition. *arXiv [Preprint]*. arXiv:1910.11476. doi: 10.48550/arXiv.1910.11476

Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., et al. (2020). "Bond: bert-assisted open-domain named entity recognition with distant supervision," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (New York, NY: ACM), 1054–1064. doi: 10.1145/3394486.3403149

Lin, Y., Xiao, H., Liu, J., Lin, Z., Lu, K., Wang, F., et al. (2022). Knowledge-enhanced relation extraction dataset. *arXiv [Preprint]*. arXiv:2210.11231. doi: 10.48550/arXiv.2210.11231

Ma, R., Wang, X., Zhou, X., Zhang, Q., and Huang, X. (2023). "Towards building more robust NER datasets: an empirical study on NER dataset bias from a dataset difficulty view," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 4616–4630. doi: 10.18653/v1/2023.emnlp-main.281

Muthuraman, K., Reiss, F., Xu, H., Cutler, B., and Eichenberger, Z. (2021). "Data cleaning tools for token classification tasks," in *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, eds. E. Dragut, Y. Li, L. Popa, and S. Vucetic (Association for Computational Linguistics), 59–61. doi: 10.18653/v1/2021.dash-1.10

Patil, R., and Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (LLMS). *Appl. Sci.* 14:2074. doi: 10.3390/app14052074

Pecher, B., Srba, I., and Bielikova, M. (2024). Fine-tuning, prompting, in-context learning and instruction-tuning: How many labelled samples do we need? *arXiv [Preprint]*. arXiv:2402.12819. doi: 10.48550/arXiv.2402.12819

Peng, C., Xia, F., Naseriparsa, M., and Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. *arXiv [Preprint]*. arXiv:2303.13948. doi: 10.48550/arXiv.2303.13948

Peng, H., Gao, T., Han, X., Lin, Y., Li, P., Liu, Z., et al. (2020). Learning from context or names? an empirical study on neural relation extraction. *arXiv [Preprint]*. arXiv:2010.01923. doi: 10.48550/arXiv.2010.01923

Peng, H., Wang, X., Hu, S., Jin, H., Hou, L., Li, J., et al. (2022). Copen: Probing conceptual knowledge in pre-trained language models. *arXiv [Preprint]*. arXiv:2211.04079. doi: 10.48550/arXiv.2211.04079

Qian, K., Raman, P. C., Li, Y., and Popa, L. (2020). Partner: human-in-the-loop entity name understanding with deep learning. *Proc. AAAI Conf. Artif. Intell.* 34, 13634–13635. doi: 10.1609/aaai.v34i09.7104

Razuvayevskaya, O., Wu, B., Leite, J. A., Heppell, F., Srba, I., Scarton, C., et al. (2023). Comparison between parameter-efficient techniques and full fine-tuning: a case study on multilingual news article classification. *PLoS ONE* 19:e0301738. doi: 10.1371/journal.pone.0301738

Riedel, S., Yao, L., and McCallum, A. (2010). "Modeling relations and their mentions without labeled text," in *Proceedings of the 2010th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III, ECMLPKDD'10* (Berlin: Springer-Verlag), 148–163. doi: 10.1007/978-3-642-15939-8_10

Shen, Y., Yun, H., Lipton, Z. C., Kronrod, Y., and Anandkumar, A. (2017). Deep active learning for named entity recognition. *arXiv [Preprint]*. arXiv:1707.05928. doi: 10.48550/arXiv.1707.05928

Soares, L. B. FitzGerald, N., Ling, J., Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. *arXiv [Preprint]*. arXiv:1906.03158. doi: 10.48550/arXiv.1906.03158

Son, J., Kim, J., Lim, J., and Lim, H. (2022). Grasp: guiding model with relational semantics using prompt for dialogue relation extraction. *arXiv [Preprint]*. arXiv:2208.12494. doi: 10.48550/arXiv.2208.12494

Suchanek, F. M., Alam, M., Bonald, T., Paris, P.-H., and Soria, J. (2023). Yago 4.5: A large and clean knowledge base with a rich taxonomy. *arXiv [Preprint]*.

Törnberg, P. (2024). Best practices for text annotation with large language models. *arXiv [Preprint]*. arXiv:2402.05129. doi: 10.48550/arXiv.2402.05129

Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., and Kambhampati, S. (2024). "Planbench: an extensible benchmark for evaluating large language models on planning and reasoning about change," in *NIPS '23* (Red Hook, NY: Curran Associates Inc).

Wan, Z., Liu, Q., Mao, Z., Cheng, F., Kurohashi, S., Li, J., et al. (2022). Rescue implicit and long-tail cases: nearest neighbor relation extraction. *arXiv [Preprint]*. arXiv:2210.11800. doi: 10.48550/arXiv.2210.11800

Wang, C., Liu, X., Chen, Z., Hong, H., Tang, J., Song, D., et al. (2022). Deepstruct: pretraining of language models for structure prediction. *arXiv [Preprint]*. arXiv:2205.10475. doi: 10.48550/arXiv.2205.10475

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., et al. (2023). Gpt-ner: named entity recognition via large language models. *arXiv [Preprint]*. arXiv:2304.10428. doi: 10.48550/arXiv:2304.10428

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L., et al. (2022). A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* 135, 364–381. doi: 10.1016/j.future.2022.05.014

Yu, J., Bohnet, B., and Poesio, M. (2020). Named entity recognition as dependency parsing. *arXiv [Preprint]*. arXiv:2005.07150. doi: 10.48550/arXiv.2005.07150

Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, eds. L. Màrquez, C. Callison-Burch, and J. Su (Lisbon: Association for Computational Linguistics), 1753–1762. doi: 10.18653/v1/D15-1203

Zhang, S., He, L., Dragut, E., and Vucetic, S. (2019). "How to invest my time: lessons from human-in-the-loop entity extraction," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (New York, NY: ACM), 2305–2313. doi: 10.1145/3292500.3330773

Zhang, Y., and Xiao, G. (2024). Named entity recognition datasets: a classification framework. *Int. J. Comput. Intell. Syst.* 17:71. doi: 10.1007/s44196-024-00456-1

Zhao, H., Andriushchenko, M., Croce, F., and Flammarion, N. (2024). Is in-context learning sufficient for instruction following in LLMS? *arXiv [Preprint]*. arXiv:2405.19874. doi: 10.48550/arXiv.2405.19874