Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

CUMULATIVE DISSERTATION

# Biomedical Knowledge Graph Question Answering

Xi Yan

Semantic Systems
Department of Informatics
Faculty of Mathematics, Informatics and Natural Sciences

Universität Hamburg
Hamburg, Germany

A thesis submitted for the degree of

*Doctor rerum naturalium (Dr. rer. nat.)*

Year of Disputation 2026

Biomedical Knowledge Graph Question Answering

Dissertation submitted by: Xi Yan

Supervisor(s):
Ricardo Usbeck, Leuphana University Lüneburg

Date of Disputation 04/03/2026

Universität Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Semantic Systems

# Affidavit

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

I hereby declare under oath that I have written this dissertation myself and have not used any sources or resources other than those indicated. If generative artificial intelligence (GenAI)-based electronic resources were used in the preparation of this dissertation, I confirm that my own contribution was paramount and that all resources used are fully documented in accordance with good scientific practice. I bear responsibility for any incorrect or distorted content, incorrect references, violations of data protection and copyright law, or plagiarism generated by the GenAI.

18-09-2025

Date

Signature

(Xi Yan)

# Acknowledgements


## Personal

I would like to thank my mentor, Ricardo, for his ongoing support, trust, and invaluable feedback throughout the course of my doctoral studies.

My sincere thanks go to my colleagues, especially those who began around the same time and experienced this journey alongside me — Angelie, Cedric, Junbo, Longquan, Debayan, Tilahun, and many others — for the thoughtful conversations and the productive, shared years.

I would like to express my heartfelt gratitude to the two cats in my life, Minus and Xiaopao, who reminded me that I am capable of taking responsibility and stronger than I ever expected. Without them, this thesis would not have been possible.

I am also deeply grateful to my friends, without whom this journey would have been far more difficult. Their unwavering support, understanding, and companionship during challenging times have meant a great deal to me. Johnny, Robert, Felix, Shen and many others — thank you for always being there to listen. I am also thankful for the climbing friends I've met along the way; the time spent climbing brought me balance, clarity, and strength when I needed it most.

To my little brother and my mom — I share this achievement with you. Everything from the past has shaped who I am today, we can do anything.

Ich versichere, dass dieses gebundene Exemplar der Dissertation und das in elektronischer Form eingereichte Dissertationsexemplar (über den Docata-Upload) und das bei der Fakultät (zuständiges Studienbüro bzw. Promotionsbüro Physik) zur Archivierung eingereichte gedruckte gebundene Exemplar der Dissertationsschrift identisch sind.

I, the undersigned, declare that this bound copy of the dissertation and the dissertation submitted in electronic form (via the Docata upload) and the printed bound copy of the dissertation submitted to the faculty (responsible Academic Office or the Doctoral Office Physics) for archiving are identical.

18-09-2025

**Date** **Signature**

# Use of Third-Party Software

For correcting grammar, and general writing improvement, I relied on Grammarly, an AI-based tool, as well as claude. I relied on the ShareLaTeX instance provided by the Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen for writing the thesis in LaTeX. The figures used in this thesis were created with Draw.io.

*Abeunt studia in mores*

— Publius Ovidius Naso, 1829

# Abstract

Automatic question answering in the biomedical domain is developing rapidly, driven by the growing need for quick access and up-to-date clinical evidence for the diagnosis and treatment of diseases. As healthcare professionals, researchers, and patients increasingly seek timely, precise, and explainable answers from vast structured data sources, traditional information retrieval (IR) methods—which rely on searching across large document collections—often fall short in terms of efficiency and accuracy. To address this, biomedical knowledge graphs (KGs) are emerging as a trustworthy and efficiently queryable structure for retrieving relevant information in the form of question answering (QA).

Biomedical knowledge graphs are heterogeneous networks that represent biomedical information by modeling biological entities as nodes and their relationships as edges. These graphs provide a structured, semantically rich, and computationally efficient framework for storing and querying large volumes of domain-specific knowledge, often accessed through query languages such as SPARQL. As a result, the task of biomedical question answering over KGs (KGQA) has received increasing attention. KGQA systems aim to enable users to pose complex natural language questions and retrieve meaningful answers by reasoning over the graph. Typically, a KGQA pipeline consists of several stages: identifying the entities and relations from the question, linking these entities and relations to corresponding nodes and edges in the KG, generating a query or traversing the graph to identify relevant paths, and finally extracting or synthesizing an answer based on the retrieved information.

One key challenge in KGQA is the lack of comprehensive and standardized datasets. The landscape is fragmented, with multiple knowledge graphs evolving independently across different versions and sources. As a result, we lack a unified system that provides an overview of existing datasets and KGQA systems, making it difficult to replicate or benchmark older systems, which often become irreproducible over time. This issue is particularly pronounced in the biomedical domain, where data annotation is expensive and time-consuming due to the need for domain expertise, making it hard to find available datasets. Furthermore, linking entities from natural language questions to the corresponding nodes in KGs remains a major challenge, largely due to the scarcity of high-quality training data for entity linking and disambiguation.

Another major challenge in linking and extracting knowledge from biomedical KGs in response to natural language questions lies in the lack of domain-specific semantic alignment. Biomedical language is highly specialized and differs significantly from general human language. However, most large language models (LLMs) we use to do such tasks are trained primarily on general-domain text, which limits their ability to accurately interpret biomedical terminology and relations. One promising research direction is the integration of symbolic knowledge from

KGs into LLMs. This can involve enriching LLM inputs with synonyms, definitions, or descriptions of biomedical entities and relations, thereby bridging the gap between natural language and structured domain knowledge.

Facing the above challenges, this dissertation aims to advance the development and scalability of KGQA systems in the biomedical domain by addressing key challenges related to dataset availability, system comparability, and data generation. More specifically, this dissertation contributes to the domain of biomedical knowledge graph question answering in the following key aspects, in terms of the resources, **Chapter 3**: This chapter developed a public leaderboard that compiles existing datasets and systems in KGQA, enabling standardized and transparent comparisons across benchmarks. **Chapter 4**: This work created a multilingual KGQA dataset containing questions of varying complexity, supporting research in multilingual reasoning and cross-lingual understanding. Built based on the above resources, we review and develop work on the KGQA systems: **Chapter 5**: This chapter proposed and evaluated a biomedical entity linking method that injects symbolic information from knowledge graphs to improve disambiguation and linking performance. **Chapter 6**: This work conducted a comprehensive review of existing relation extraction methods that combine both neural and symbolic approaches, providing insights into trends and gaps in the field. **Chapter 7** This chapter designed a generalizable framework for automatically generating biomedical KGQA datasets, and used this framework to build the first large-scale biomedical KGQA dataset.

# Zusammenfassung

Die automatische Beantwortung von Fragen im biomedizinischen Bereich entwickelt sich rasant, angetrieben durch den wachsenden Bedarf an selbstgesteuertem Zugang und aktuellen klinischen Erkenntnissen für die Diagnose und Behandlung von Krankheiten. Da medizinisches Fachpersonal, Forscher und Patienten zunehmend nach zeitnahen, präzisen und erklärbaren Antworten aus umfangreichen strukturierten Datenquellen suchen, sind herkömmliche Information-Retrieval-Methoden (IR), die sich auf die Suche in großen Dokumentensammlungen stützen, in Bezug auf Effizienz und Genauigkeit oft unzureichend. Um dieses Problem zu lösen, werden biomedizinische Wissensgraphen (KGs) als vertrauenswürdige und effizient abfragbare Struktur zum Abrufen relevanter Informationen in Form von Fragebeantwortung (QA) eingesetzt.

Biomedizinische Wissensgraphen (KGs) sind heterogene Netzwerke, die biomedizinische Informationen darstellen, indem sie biologische Einheiten als Knoten und ihre Beziehungen als Kanten modellieren. Diese Graphen bieten einen strukturierten, semantisch reichhaltigen und rechnerisch effizienten Rahmen für die Speicherung und Abfrage großer Mengen von domänenspezifischem Wissen, auf das häufig über Abfragesprachen wie SPARQL zugegriffen wird. Infolgedessen hat die Aufgabe der Beantwortung biomedizinischer Fragen über KGs (KGQA) zunehmende Aufmerksamkeit erhalten. KGQA-Systeme sollen es Benutzern ermöglichen, komplexe Fragen in natürlicher Sprache zu stellen und sinnvolle Antworten zu erhalten, indem sie über den Graphen schlussfolgern. Typischerweise besteht eine KGQA-Pipeline aus mehreren Stufen: Identifizierung der Entitäten und Relationen aus der Frage, Verknüpfung dieser Entitäten und Relationen mit den entsprechenden Knoten und Kanten im KG, Generierung einer Anfrage oder Durchlaufen des Graphen, um relevante Pfade zu identifizieren, und schließlich Extraktion oder Synthese einer Antwort auf der Grundlage der abgerufenen Informationen.

Eine zentrale Herausforderung bei der KGQA ist der Mangel an umfassenden und standardisierten Datensätzen. Die Landschaft ist zersplittert, mit mehreren Wissensgraphen (KGs), die sich unabhängig voneinander in verschiedenen Versionen und Quellen entwickeln. Infolgedessen fehlt ein einheitliches System, das einen Überblick über die vorhandenen Datensätze und KGQA-Systeme bietet, was die Replikation oder den Vergleich älterer Systeme erschwert, die mit der Zeit oft nicht mehr reproduzierbar sind. Dieses Problem ist im biomedizinischen Bereich besonders ausgeprägt, wo die Annotation von Daten teuer und zeitaufwändig ist, da Fachwissen erforderlich ist und es schwierig ist, verfügbare Datensätze zu finden. Darüber hinaus stellt die Verknüpfung von Entitäten aus natürlichsprachlichen Fragen mit den entsprechenden Knoten in KGs nach wie vor eine große Herausforderung dar, was vor allem auf den Mangel an hochwertigen Trainingsdaten für die Verknüpfung und Disambiguierung von Entitäten zurückzuführen ist.

Eine weitere große Herausforderung bei der Verknüpfung und Extraktion von Wissen aus biomedizinischen KGs als Antwort auf natürlichsprachliche Fragen liegt im Fehlen eines domänenspezifischen semantischen Abgleichs. Biomedizinische Sprache ist hochspezialisiert und unterscheidet sich erheblich von der allgemeinen menschlichen Sprache. Die meisten großen Sprachmodelle (LLMs), die wir für solche Aufgaben verwenden, sind jedoch in erster Linie auf Texte aus allgemeinen Bereichen trainiert, was ihre Fähigkeit zur genauen Interpretation biomedizinischer Terminologie und Beziehungen einschränkt. Eine vielversprechende Forschungsrichtung ist die Integration von symbolischem Wissen aus KGs in LLMs. Dies kann die Anreicherung von LLM-Eingaben mit Synonymen, Definitionen oder Beschreibungen biomedizinischer Entitäten und Relationen beinhalten, wodurch die Lücke zwischen unstrukturierter Sprache und strukturiertem Domänenwissen überbrückt wird.

Angesichts der oben genannten Herausforderungen zielt diese Dissertation darauf ab, die Entwicklung und Skalierbarkeit von KGQA-Systemen im biomedizinischen Bereich voranzutreiben, indem sie sich mit den wichtigsten Herausforderungen in Bezug auf die Verfügbarkeit von Datensätzen, die Vergleichbarkeit von Systemen und die Datengenerierung befasst. Insbesondere leistet diese Dissertation in den folgenden Schlüsselbereichen einen Beitrag zum Bereich der biomedizinischen Wissensgraphen-Fragenbeantwortung, was die Ressourcen betrifft: **Kapitel 3:** In diesem Kapitel wurde eine öffentliche Rangliste entwickelt, die bestehende Datensätze und Systeme in KGQA zusammenfasst und standardisierte und transparente Vergleiche zwischen Benchmarks ermöglicht. **Kapitel 4:** In dieser Arbeit wurde ein mehrsprachiger KGQA-Datensatz mit Fragen unterschiedlicher Komplexität erstellt, der die Forschung im Bereich des mehrsprachigen Schlussfolgerns und des sprachübergreifenden Verstehens unterstützt. Auf der Grundlage der oben genannten Ressourcen überprüfen und entwickeln wir Arbeiten zu KGQA-Systemen: **Kapitel 5:** In diesem Kapitel wurde eine Methode zur Verknüpfung biomedizinischer Entitäten vorgeschlagen und bewertet, bei der symbolische Informationen aus Wissensgraphen eingefügt werden, um die Disambiguierung und die Verknüpfungsleistung zu verbessern. **Kapitel 6:** In dieser Arbeit wurde eine umfassende Überprüfung bestehender Methoden zur Relationsextraktion durchgeführt, die sowohl neuronale als auch symbolische Ansätze kombinieren, und Einblicke in Trends und Lücken in diesem Bereich gegeben. **Kapitel 7** In diesem Kapitel wurde ein verallgemeinerbarer Rahmen für die automatische Generierung biomedizinischer KGQA-Datensätze entworfen und dieser Rahmen zur Erstellung des ersten groß angelegten biomedizinischen KGQA-Datensatzes verwendet.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **BART** | Bidirectional and Auto-Regressive Transformers |
| **bioKGQA** | Biomedical Knowledge Graph Question Answering |
| **bioNER** | Biomedical Named Entity Recognition |
| **CNN** | Convolutional Neural Network |
| **DL** | Deep Learning |
| **EL** | Entity Linking |
| **FFNN** | Feed-forward Neural Network |
| **IR** | Information Retrieval |
| **IE** | Information Extraction |
| **KG** | Knowledge Graph |
| **KGQA** | Knowledge Graph Question Answering |
| **LLM** | Large Language Model |
| **LSTM** | Long Short-Term Memory network |
| **MLP** | Multilayer Perceptron |
| **ML** | Machine Learning |
| **NER** | Named Entity Recognition |
| **NLP** | Natural Language Processing |
| **NN** | Neural Network |
| **PPO** | Proximal Policy Optimization |
| **QA** | Question Answering |
| **RDF** | Resource Description Framework |
| **RE** | Relation Extraction |
| **ReLU** | Rectified Linear Unit |
| **RHLF** | Reinforcement Learning from Human Feedback |
| **RNN** | Recurrent Neural Network |

# 1 Introduction

**This introduction chapter outlines the motivation, related work, key publications, and main contributions of this dissertation on biomedical knowledge graph question answering (KGQA). Motivated by the need for interpretable and efficient access to biomedical knowledge, this thesis reviews existing efforts in KGQA, including datasets and systems. This chapter further highlights key challenges, such as limited datasets and domain-specific language, and positions the contributions of this research within that context.**

## Contents



## 1.1 Motivation

Knowledge graphs are heterogeneous graphs that contain different types of vertices and edges. In a biomedical knowledge graph, core biological components such as genes, diseases, and drugs are modeled as entities (nodes), while the interactions or associations between them are modeled as relations (edges). This

graph-based structure enables the integration of heterogeneous biomedical data into a unified framework, where each node represents a unique concept and each edge encodes a specific type of biological or pharmacological interaction (Bonner et al., 2022), as presented in Figure 1.3. The resulting graph enables efficient querying and reasoning, supporting tasks like personalized diagnosis of disease or search for up-to-date domain-specific knowledge (Jin et al., 2022).

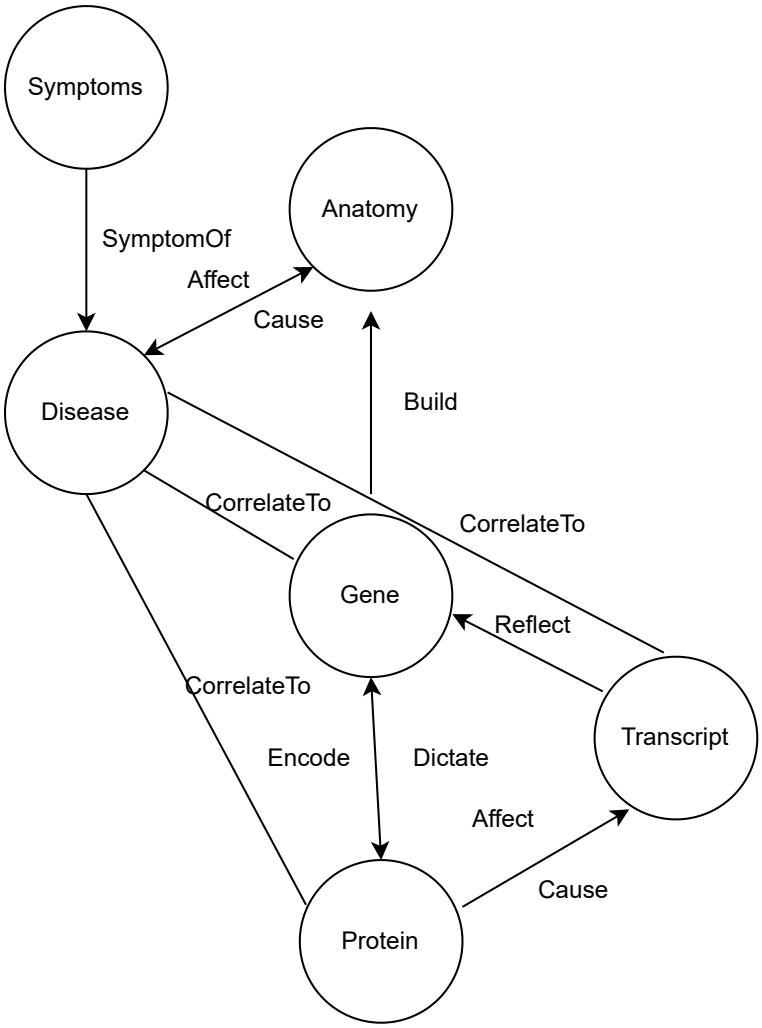



**Figure 1.1**: A knowledge Graph example, with instances borrowed from PrimeKG (Chandak et al., 2023a).

Knowledge Graph Question Answering (KGQA) is the task of automatically answering natural language questions by leveraging structured information stored in one or more knowledge graphs (KGs). It involves interpreting a user's question, mapping it to the corresponding entities and relations in the graph, and retrieving or reasoning over relevant triples to produce an accurate answer. This process often requires several sub-tasks. Recent KGQA systems typically follow two main paradigms:

1. **Information retrieval-style**, which link the entities to the graph and use neural models to retrieve and rank candidate answers from the graph; and

2. **Semantic parsing-style**, which translate the input question into a formal query language (such as SPARQL or a logical form) that can be executed directly over the KG to retrieve the answer.

Despite differences in paradigm, both KGQA approaches require recognizing and linking entities mentioned in the natural language question to their corresponding nodes in the graph, i.e., *Named Entity Recognition* (NER) and *Entity Linking* (EL). NER refers to the process of identifying entity spans in the input text. In biomedical KGs, entities represent identifiable concepts of interest, such as genes, proteins, or anatomical structures. After NER, the extracted span is passed to the EL step, which maps it to the corresponding node in the KG. Due to ambiguity and lexical variation, multiple candidate entities may be retrieved, requiring contextual disambiguation to select the most relevant identifier (entity ID). In IR-based KGQA systems, the linked entity IDs are used to guide ranking and reasoning, while in semantic parsing-based systems, they are essential to formulating executable queries. A typical KGQA pipeline is depicted in Figure 1.2.

NER and EL are foundational to the performance of KGQA systems. Despite their critical role, many systems continue to rely on out-of-the-box NER and EL components for practical and time-saving reasons (Luo et al., 2024; Wang and Qin, 2024). However, these general-purpose models suffer from limitations, such as in linking unseen entities, particularly when applied to complex or domain-specific questions. A key limitation of these models lies in their inability to effectively process domain-specific terminology, especially in the biomedical domain, where the meaning of terms often depends heavily on context. And those contexts are richly stored in the KGs in the form of synonyms and descriptions (Möller and Usbeck, 2025), therefore, to combine the information from the KG to enhance those modules is a focus of this dissertation. We review existing work and raise a new framework in this perspective in chapter 5 and 6.

Recent advancements in NER, EL, and KGQA increasingly leverage Large Language Models (LLMs) as foundational components (Klager and Polleres, 2023b). These models, based on transformer architectures (Vaswani et al., 2017), have markedly enhanced the language model's ability to comprehend and generate human language. In the current landscape of generative and reasoning-driven approaches (Liu et al., 2024; OpenAI, 2023), LLMs exhibit strong performance in semantic understanding, entity disambiguation, and contextual reasoning (Li et al., 2024). This thesis builds on these developments to design principled resources and systems grounded in these capabilities.

Despite the remarkable progress enabled by LLMs, their practical application to biomedical Knowledge Graph Question Answering (bioKGQA) still faces several notable challenges. The biomedical domain is characterized by a high demand for accurate, interpretable, and timely access to information—driven by needs in clinical decision support, biomedical research, and public health (Chandak et al., 2023a). At the same time, it benefits from a rich ecosystem of structured resources, including curated biomedical knowledge graphs such as UMLS (Bodenreider, 2004a), MeSH (Lipscomb, 2000a), DrugBank (Wishart et al., 2008), etc. However,

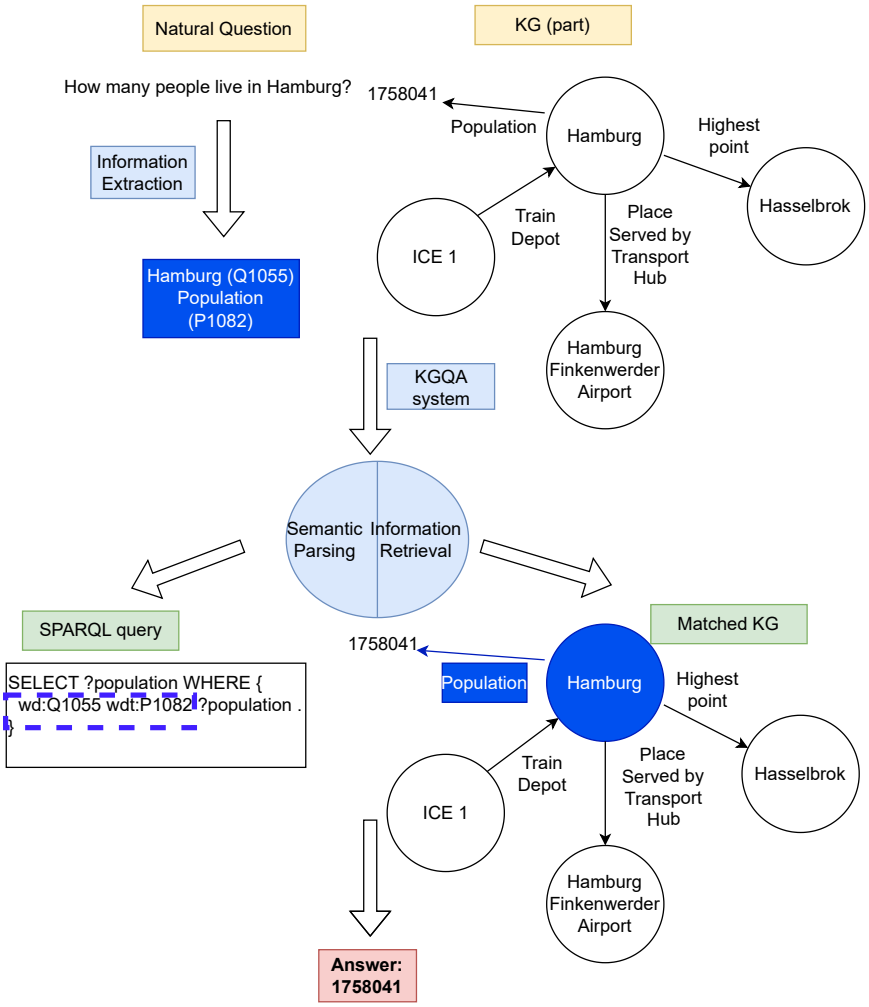



**Figure 1.2:** A typical KGQA pipeline, where information extraction modules (entity linking and relation extraction) form the basis for later query generation or retrieval. The linked entities and relations are marked with blue boxes. The KG is sampled from Wikidata.

effectively leveraging these structured resources in combination with LLMs remains a non-trivial task due to challenges such as domain-specific language, entity ambiguity, and the evolving nature of the biomedical knowledge graph. These factors make bioKGQA a compelling and timely area of research.

1. One major issue is *hallucination* (Ji et al., 2023), where LLMs may generate plausible-sounding statements or facts that are incorrect or not logical. Such behavior can significantly undermine the reliability of downstream tasks, particularly in EL and KGQA, where factual grounding is essential. Finding and tackling this problem requires systematic evaluation on curated and complex datasets. Unfortunately, in biomedical KGQA, the availability of such datasets remains limited (Yan et al., 2024), which complicates efforts to analyse the performance of KGQA systems. This is due to the high cost of creating such a dataset, since it requires professionals to annotate and validate biomedical questions and corresponding queries. Therefore, this

dissertation sets up a leaderboard based on past resources (chapter 3), builds a manually created dataset (chapter 4), and also proposes a framework for the automatic generation of KGQA datasets and applies it to the biomedical domain, establishing a benchmark for evaluating the factuality of LLMs and mitigating hallucinations (chapter 7.

2. LLMs often require *domain adaptation* to perform effectively in specialized tasks such as biomedical NER, EL, and KGQA (Saad-Falcon et al., 2023). Without fine-tuning on domain-specific corpora, their performance is insufficient, particularly in recognizing technical terms, rare entities, etc, with specific terminologies that are rarely seen in the LLM's training process. Unlike traditional symbolic systems, LLMs lack transparency and do not encode explicit knowledge—limitations that are especially critical in high-stakes domains like biomedicine. To address these challenges, this work reviews and explores existing and emerging methods for integrating LLMs with structured biomedical knowledge to support robust and interpretable KGQA (chapter 6).

Given the scarcity of high-quality biomedical KGQA datasets and the reliability concerns surrounding LLMs in this domain, this dissertation focuses on the following key objectives: reviewing current KGQA resources, creating KGQA datasets with an automatic and manual annotation method, as well as developing knowledge-enhanced KGQA systems.

## 1.2 Related Work

We categorize existing work on bioKGQA into two main directions: datasets and systems. This classification allows us to highlight key gaps in the research landscape—specifically, the lack of a comprehensive benchmark dataset and the limited ability of current systems, particularly general-purpose LLMs, to effectively retrieve information in the presence of complex biomedical terminology. These limitations motivate the central research question of this dissertation and form the basis for its main contribution: the development of a high-quality, domain-specific bioKGQA dataset and accompanying methods of extracting and linking the information from natural text to the KG to advance the field. On top of that, a systematic review and presentation of KGQA datasets and systems are important. Therefore, we also review the work on the KGQA leaderboard for reproducibility. In addition, we provide an in-depth review of neural-symbolic information extraction, which is an important part of KGQA systems.

### 1.2.1 bioKGQA Dataset

To date, research on biomedical KGQA remains limited, largely due to the absence of standardized benchmark datasets for training, evaluation, and experimentation. Effective use of knowledge graphs in practical applications—particularly for KGQA—depends heavily on the availability of high-quality, domain-specific datasets. This need is especially critical in the era of LLMs. Without robust

| Dataset | QALD-4 | Bgee-QA | OMA-QA | CORDIS-QA |
|---|---|---|---|---|
| # Q-A Pairs | 50 | 20 | 10 | 30 |
| Underlying KG | DrugBank | Bgee | OMA | CORDIS |

**Table 1.1**: Statistics of existing BioKGQA datasets. Adapted from (Yan et al., 2024).

biomedical KGQA benchmarks, it is difficult to assess the capabilities of LLMs or fine-tune them effectively for domain-specific tasks.

Currently, only a few public **biomedical KGQA datasets** are available, including BgeeQA, OMA QA, CORDIS-QA (Sima, de Farias, et al., 2021), and Task 2 of QALD-4 (Unger et al., 2014a). These datasets are typically small in scale and focus on narrow biomedical subdomains (Yan et al., 2024). For example, BgeeQA centers on gene expression across anatomical entities, while OMA QA and CORDIS-QA are limited to comparative genomics and EU-funded biomedical project data, respectively. Statistics and underlying KGs are listed in the Table 7.1.

Facing the data scarcity issue, there are two lines of methods for generating the KQGA dataset. The first involves collecting natural questions from users or corpora and manually annotating them with the corresponding SPARQL queries. While this method ensures high semantic alignment, it is time-consuming and labor-intensive—especially for questions involving multiple hops, filters, or complex constraints. This is especially true for bioKGQA, since it involves biological experts and, thus, costs more.

The second approach, known as the OVERNIGHT (ON) framework, is introduced by Wang et al. (2015) as a semantic parsing dataset generation method. It follows a three-step pipeline: (1) logical forms (e.g., SPARQL queries) are generated directly from the knowledge graph; (2) these logical forms are transformed into canonical questions, which are grammatically simplified but preserve the intended semantics; and (3) the canonical questions are paraphrased into natural language by crowd workers. This method may result in artificial question distributions that lack the linguistic variety and complexity of naturally occurring queries. Recent method also features finetuning an LLM over triples from KG to generate natural questions, which proves to enhance model capacity in query generation (Shu and Yu, 2024a).

This dissertation explores both perspectives of data generation by employing manual and automatic frameworks, resulting in a multilingual KGQA dataset and a bioKGQA dataset.

### 1.2.2 bioKGQA System

In the general domain, LLMs are increasingly leveraged for both **SPARQL query generation** (Banerjee et al., 2022a; Jiang, Yan, et al., 2023a) and **information retrieval over knowledge graphs** (Chen, Jiang, et al., 2024; Zhao et al., 2025), owing to their strong capabilities in understanding complex grammar and capturing semantic relationships. Recent advances also explore the use of **LLMs as autonomous agents** that can navigate and reason over KGs, driven by the growing

ability of these models to *reflect on failed attempts and incorporate feedback* into their reasoning process (Jiang, Zhou, et al., 2024; Su et al., 2024; Xu et al., 2024).

In biomedical and general-purpose settings, Sima, de Farias, et al. (2021) uses a graph-based strategy to translate user queries into candidate SPARQL queries, which extract, link, and rank the related entities as well as the related path, and turn them into SPARQL queries, which proves to be efficient over different KGs. Reyes et al. (2024) fine-tunes an LLM (OpenLLaMA_7B_v2[1]) using QLoRA and PEFT, first on a general-domain KGQA dataset over Wikidata—namely KQA Pro (Cao, Shi, et al., 2022)—and subsequently on a customized, small set of question-query pairs generated from domain-specific SPARQL queries. Huang et al. (2021) links a corpora with entities extracted to form a KG and retrieves the answer from it.

LLMs have been employed for various tasks within KGQA pipelines, including SPARQL query generation (Klager and Polleres, 2023a), retrieval of relevant entities (Ji et al., 2024), and even direct interaction with knowledge graphs (Xu et al., 2024). Since the development of powerful LLMs coincided with the early stages of this dissertation project, the research presented here is fundamentally built upon and inspired by recent advancements in LLM technologies. As can be seen, LLMs have become the backbone of modern KGQA systems, and this technology will be discussed in detail in the theoretical background Chapter 2.

### 1.2.3 KGQA Leaderboard

This section draws on content from chapter 3, which includes a systematic review of existing KGQA leaderboards and platforms. The advancement in this direction of research has remained largely unchanged from the time of its publication to the completion of this dissertation.

Tracking progress in machine learning and NLP can be done through benchmarking frameworks or reporting platforms. Benchmarking frameworks, such as GERBIL QA (Usbeck, Röder, Hoffmann, et al., 2018), provide FAIR (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, Blomberg, Boiten, Bonino da Silva Santos, et al., 2016) evaluation of KGQA systems and maintain integrated leaderboards, but their usefulness depends heavily on community adoption and developer support. Alternative command-line tools (e.g., QALDGen (Singh et al., 2016)) enable flexible benchmarking, yet their offline nature limits transparency and reproducibility. Reporting platforms like Huggingface Space [2] and Papers with Code [3] offer centralized access to results, but for KGQA, coverage remains sparse (Perevalov* et al., 2022). More structured initiatives, such as the Open Research Knowledge Graph (ORKG)(Auer et al., 2020), show promise in enabling persistent and semi-automatic reporting, though adoption is still limited. Surveys provide valuable overviews (Pereira et al., 2022; Song et al., 2023; Yani and Krisnadhi, 2021), but they become outdated quickly or focus on narrow subtopics. Altogether, these efforts highlight the need for a sustainable, centralized, and community-driven reporting platform for KGQA to ensure trustworthy and up-to-date insights.

1. https://huggingface.co/openlm-research/open_llama_7b_v2
2. https://huggingface.co/spaces
3. https://huggingface.co/papers/trending

### 1.2.4 Neuro-symbolic Information Extraction

This section draws on content from chapter 6, which is a systematic review of existing datasets and systems for neural symbolic relation extraction methods.

According to (Yan, Usmanova, et al., 2025), neuro-symbolic systems can be categorized by the type of information they exploit: KG-based knowledge, linguistic features, prior meta-information, and logical rules. The ways this information is utilized also differ: some studies leverage external sources to generate distantly supervised training data, while others integrate symbolic information directly into ML systems.

Only a few studies (Dai et al., n.d.; Le et al., 2023) have explored enhancing distant supervision. Existing approaches often rely on external sources such as subgraph paths and entity types, which require entity linking given the richness of KGs. After aligning text with the KG, integration strategies differ: some enrich the text with KG-derived information, while others employ KG embeddings or generate constraints. Importantly, incorporating ontological information has been shown to improve data quality.

More studies focus on how to combine existing NLP systems with symbolic information. Such information may include KG information, linguistic information, prior meta-information, and inferred logical rules.

**KG information** may consist of entities (Zhang, Zhu, et al., 2021), relations, attributes (Hogan et al., 2006), ontologies (Aghaebrahimian et al., n.d.), paths (Jain et al., 2023), and classes (Liu et al., 2023). This knowledge can be incorporated during inference or training to enhance the contextual understanding and reasoning capacity of LLMs.

**Linguistic information** has also been exploited, such as sememe knowledge from KGs (Zhao et al., 2023) and entity synonyms and types (Jain et al., 2023).

**Prior meta-information** refers to pre-existing heuristic or statistical knowledge about entities and their relations, such as entity co-occurrence (Zhang, Yu, et al., 2021) or precomputed entity similarity scores (Li and Qian, 2022). Such information can help disambiguate entities and reduce noise.

**Logical rules** are valuable in structured domains such as medical reports, financial statements, and legal documents, where recurring patterns and fixed expressions are common. They are particularly useful when labeled data is limited. Typically, these rules are learned or generated from existing data (Fan et al., 2022; Lu et al., 2023) and incorporated into the system.

## 1.3 Research Questions

Based on the related work, progress in the biomedical domain of KGQA remains limited, largely due to the absence of large-scale, standardized benchmark datasets. Existing resources such as BgeeQA, OMA-QA, and CORDIS-QA are small in size and narrowly focused, restricting their effectiveness for training and evaluating generalizable systems. On the systems side, LLMs have become central to modern KGQA pipelines, supporting tasks such as SPARQL query generation, entity retrieval, and multi-hop reasoning. However, biomedical KGQA systems often depend on domain-adapted LLMs fine-tuned on limited datasets, which constrains

their scalability and performance. Built upon the challenges and gaps identified above, this dissertation investigates two key aspects of bioKGQA: (1) methods for generating high-quality bioKGQA datasets that can effectively support the training and evaluation of LLM-based approaches, and (2) system design principles for building robust and interpretable biomedical KGQA systems. The following research questions serve as the foundation for the core contributions of this work.

**Research Question 1**

How to generate KGQA datasets for the biomedical domain that exhibit a natural linguistic style and reflect a realistic distribution of both simple and complex questions?

**Research Question 2**

How to enhance information extraction of bioKGQA systems?

## 1.4 Publications

I list below the accepted papers that compose my dissertation. I am the first author of all papers, with some being shared first authorship marked with a star sign *.

### 1.4.1 Accepted Papers Composing this Dissertation

Aleksandr Perevalov*, Xi Yan*, Liubov Kovriguina, Longquan Jiang, Andreas Both, and Ricardo Usbeck. 2022. Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, 2998–3007. * Shared first authorship. Marseille, France: European Language Resources Association, June. (Cited on pages 7, 10 sq., 28).

Ricardo Usbeck*, Xi Yan*, Aleksandr Perevalov*, Longquan Jiang*, Julius Schulz*, Angelie Kraft, Cedric Möller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, et al. 2024. Qald-10–the 10th challenge on question answering over linked data: Shifting from dbpedia to wikidata as a kg for kgqa. * Shared first authorship, *Semantic Web* 15 (6): 2193–2207. (Cited on pages 10 sq.).

Xi Yan, Cedric Möller, and Ricardo Usbeck. 2025. Biomedical Entity Linking with Triple-aware Pre-Training. In *Proceedings of the Third International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data (SemTech4STLD 2025), co-located with the Extended Semantic Web Conference (ESWC 2025),* edited by Rima Dessí, Joy Jeenu, Danilo Dessì, Francesco Osborne, and Hidir Aras. To appear. Portoroz, Slovenia: CEUR-WS.org, June. (Cited on pages 10, 12).

Xi Yan, Aida Usmanova, Cedric Möller, Patrick Westphal, and Ricardo Usbeck. 2025. Neuro-Symbolic Relation Extraction. In *Handbook on Neurosymbolic AI and Knowledge Graphs,* 400:550–576. Frontiers in Artificial Intelligence and Applications. IOS Press. (Cited on pages 8, 10, 12).

Xi Yan, Patrick Westphal, Jan Seliger, and Ricardo Usbeck. 2024. Bridging the Gap: Generating a Comprehensive Biomedical Knowledge Graph Question Answering Dataset. In *ECAI 2024,* 1198–1205. IOS Press. (Cited on pages 4, 6, 10 sq., 114).

### 1.4.2 Comments on the Degree of Authorship

This section provides a detailed explanation of the respective contributions made by me and the co-authors to the published chapters included in this thesis.

**Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis** (Perevalov* et al., 2022): I contribute significantly to *Data Curation* by reviewing the majority of cited chapters and manually extracting evaluation metrics, dataset statistics, and system characteristics. These efforts form the empirical backbone of the chapter. I also contribute to *Investigation*, analyzing dataset usage trends, and participated in *Writing – Original Draft* as well as *Writing – Review & Editing*. I share first authorship with Aleksandr Perevalov.

**QALD-10 – The 10th Challenge on Question Answering over Linked Data: Shifting from DBpedia to Wikidata as a KG for KGQA** (Usbeck* et al., 2024): The project is conceptualized by Prof. Dr. Ricardo Usbeck, who also provide annotation resources. I lead the *Data Curation*, managing and supervising the annotation process. I conduct *Investigation* into the validity and complexity of multilingual data, verify SPARQL usability, and develop the *Methodology* for dataset creation. I also coordinate the challenge as part of *Project Administration* and host the associated workshop. Post-project, I contribute to *Writing – Original Draft* and *Writing – Review & Editing*.

**Biomedical Entity Linking with Triple-aware Pre-Training** (Yan, Möller, et al., 2025): I am involved in all stages of the project including *Conceptualization*, *Methodology*, *Implementation*, *Evaluation*, and *Writing*. Cedric support conceptual development, and supervision is provided by Prof. Dr. Ricardo Usbeck.

**Neurosymbolic AI and Knowledge Graphs** (Yan, Usmanova, et al., 2025): Along with Aida Usmanova and Cedric Möller, I contribute to *Writing – Review & Editing* across three sections. I take the lead in *Project Coordination*, managing timelines and consolidating feedback.

**Bridging the Gap: Generating a Comprehensive Biomedical Knowledge Graph Question Answering Dataset** (Yan et al., 2024): I am responsible for the *Experimental Design*, *Implementation*, *Data Validation*, and *Writing*. The idea of leveraging network reasoning is suggested by Prof. Dr. Ricardo Usbeck, and Patrick Westphal handle RDF integration for PrimeKG.

Overall supervision and strategic guidance are provided by Prof. Dr. Ricardo Usbeck.

### 1.4.3 Other papers

In addition to the main contributions presented in this dissertation, there are several other publications arising from collaborative efforts. The KGQA paper (Jiang, Yan, et al., 2023b) is the result of joint work with colleagues, to which I contribute primarily in the form of writing and manuscript preparation. Another work on semantic drift (Krause et al., 2023) emerges from a summer school project undertaken with a group of exceptional PhD peers; however, the content of that study is only tangentially related to the core focus of this dissertation.

Longquan Jiang, Xi Yan, and Ricardo Usbeck. 2023b. A Structure and Content Prompt-based Method for Knowledge Graph Question Answering over Scholarly Data. In *Joint Proceedings of Scholarly QALD 2023 and SemREC 2023 co-located with 22nd International Semantic Web Conference ISWC 2023, Athens, Greece, November 6-10, 2023,* edited by Debayan Banerjee, Ricardo Usbeck, Nandana Mihindukulasooriya, Gunjan Singh, Raghava Mutharaju, and Pavan Kapanipathi, vol. 3592. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 11).

Franz Krause, Xi Yan, Baptiste Darnala, and Michel Dumontier. 2023. On the Combination of Event Calculus and Empirical Semantic Drifts. In *Joint Proceedings of the ESWC 2023 Workshops and Tutorials co-located with 20th European Semantic Web Conference (ESWC 2023), Hersonissos, Greece, May 28-29, 2023,* edited by Mehwish Alam, Cássia Trojahn, Sven Hertling, Catia Pesquita, Christian Aebeloe, Hidir Aras, Amr Azzam, Juan Cano, John Domingue, Simon Gottschalk, Olaf Hartig, Katja Hose, Sabrina Kirrane, Pasquale Lisena, Francesco Osborne, Philipp D. Rohde, Luc Steels, Ruben Taelman, Aisling Third, Ilaria Tiddi, and Rima Türker, vol. 3443. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 11).

## 1.5 Contributions

This section provides a high-level overview of how this dissertation advances the field of KGQA by systematically addressing the proposed research questions. This dissertation presents a visual summary in Figure 1.3, which outlines the research pipeline—from dataset creation to KGQA systems—highlighting the interdependencies between components and mapping our contributions (i.e., associated publications) to each phase of the research.

### RQ1: How to generate KGQA datasets that exhibit a natural linguistic style and reflect a realistic distribution of both simple and complex questions?

To address this question, we first construct a comprehensive KGQA leaderboard (Perevalov* et al., 2022) that aggregates existing KGQA datasets and systems to systematically analyze their features and limitations. Based on this analysis, we manually create a natural, multilingual KGQA dataset using expert annotation (Usbeck* et al., 2024). Building upon this foundation, we design a framework capable of automatically generating both complex and simple questions from knowledge graphs (Yan et al., 2024). This framework is subsequently applied to the biomedical domain to mitigate the lack of domain-specific KGQA datasets.

### RQ2: How to enhance information extraction of bioKGQA systems?

To tackle this challenge, we conduct a systematic review of information extraction techniques that integrate external knowledge into large language models (LLMs) (Yan, Usmanova, et al., 2025) and their underlying datasets. Our analysis reveals that most methods leverage structured knowledge—such as definitions, synonyms, and hierarchical relations from knowledge graphs—to improve LLM performance. Based on these insights, we proposes a knowledge-enhanced approach for improving information extraction, particularly entity linking, within biomedical KGQA systems (Yan, Möller, et al., 2025). Our method is evaluated on several benchmark biomedical entity linking datasets, demonstrating significant improvements in accuracy and contextual understanding.

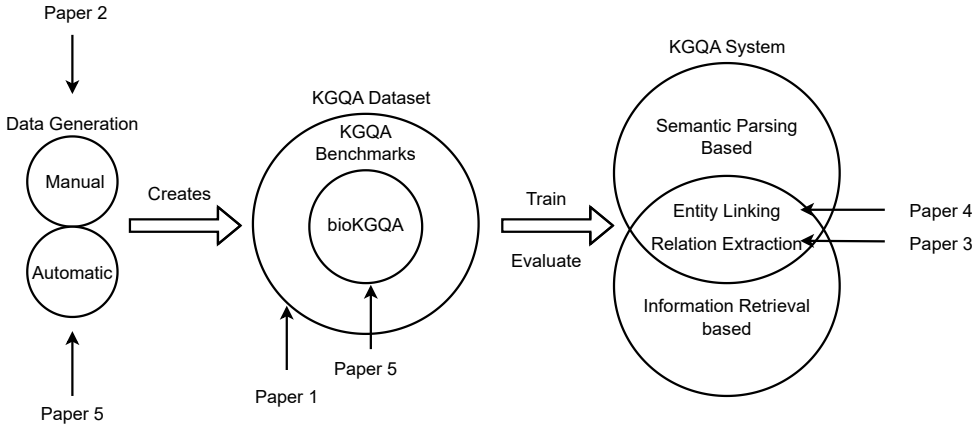



**Figure 1.3:** Situating the thesis topic and related publications in the KGQA landscape.

## 1.6 Thesis Outline

The first two chapters (Chapter 1 and Chapter 2) serve as introductory material for the main body of this dissertation. They present the motivation behind this work, outline the research questions to be addressed, and provide the necessary theoretical background for understanding the contributions. Chapters 3 to 8 consist of peer-reviewed and published research articles that directly tackle the core research questions posed in this dissertation. Each chapter is self-contained and contributes to different aspects of knowledge graph question answering KGQA. Finally, Chapter 8 presents the conclusion, summarizing the findings and outlining directions for future research.

# 2

# Theoretical Background

**This chapter provides the theoretical background necessary to understand the publications that follow. It is divided into two main parts: first, an overview of KGQA, and second, the technologies driving KGQA solutions—primarily LLMs. We begin with LLMs, tracing their evolution from non-neural to neural approaches, then transformer architectures, and finally to generative language models, explaining their development and underlying techniques. For KGQA, we introduce KGs and discuss the motivation and methods for performing question answering over them. Additionally, the chapter covers information extraction techniques, which are essential components in all KGQA systems.**

## Contents

## 2.1 Introduction

This work builds upon several years of foundational research across multiple areas. At the forefront of this evolution are the advancements in LLMs, which have fundamentally reshaped the landscape of KGQA. LLMs now serve as the backbone of many KGQA systems, powering tasks such as question understanding, entity linking, and query generation. Therefore, we begin by introducing the principles and technical details of LLMs, including their training processes. Special attention will be given to the transformer architecture, which underlies most modern LLMs, with an in-depth look into BART (Lewis et al., 2020) and Mixtral (Jiang, Sablayrolles, Mensch, Bamford, Chaplot, de Las Casas, et al., 2023), which we use in later publications. This section is directly related to chapter 5 and 7.

Building on this, we turn to the task of KGQA itself, which involves answering natural language questions by reasoning over structured data in KGs. To contextualize the integration of LLMs within KGQA systems, it is essential to introduce the principles, structures, and roles of KGs in this setting. We also introduce the pipeline of general KGQA systems. This section is fundamental to understanding chapter 3, 4 and 7.

The third section will focus on information extraction—an indispensable component of all KGQA systems and the central topic of two of the papers included in this dissertation. We will concentrate on core tasks such as relation extraction and entity linking, discussing their definitions and relevance to the broader KGQA pipeline. This section is related to chapter 5 and 6.

## 2.2 Language Model

For centuries, scholars have sought to understand the human language system. As early as the 3rd or 4th century, efforts began to uncover the relationships between different aspects of language—such as the connection between writing and pronunciation (Restall, 2003), as well as the structure of syntax (Matthews, 1981), semantics (Lyons, 1995), and morphology (Matthews, 1991). However, the task of modeling language has never gained as much prominence as it has in recent years, with the rise of *LLMs*. Today, LLMs are deeply embedded in many aspects of society, powering applications such as chatbots, search engines, and personal assistants. What has reignited the interest in this longstanding subject? What has made these recent approaches so successful? This section aims to provide an overview of the key developments that answer these questions.

We would divide the section by important language models in history and introduce their ideals as well as how they affect the development of language models, and essentially, we see the LLM that is popular today. This section is inspired by the book Speech and Language Processing by Jurafsky and Martin (2025).

Language models aim at predicting the probability of the next word given the current input sequence. Earlier modeling approaches like n-gram models predict the next word by calculating probabilities based on a limited context window of preceding words and fixed word representations calculated based on statistics of word combinations (Jurafsky and Martin, 2025)—readers can explore this further

in the referenced textbook.[1] Although statistically simple and computationally efficient, these methods face notable limitations: they perform poorly on unseen word sequences, require substantial memory to store probability distributions over all possible combinations, and, most importantly, are restricted to surface-level co-occurrence patterns without capturing the deeper semantic relationships between words.

### 2.2.1 Neural Networks

Neural networks (NNs) provide a robust framework for language modeling by learning continuous vector representations of words (embeddings) and capturing complex non-linear mappings that approximate underlying language patterns. Unlike traditional approaches that rely on creating a discrete probability table based on word combinations for fixed sequences, neural architectures project language into a continuous vector space. This projection is learned from training data by capturing semantic similarities and abstract linguistic features, enabling the model to generalize beyond observed word combinations. The resulting continuous representations facilitate high-dimensional geometric operations, such as distance computations and vector arithmetic, which are valuable for tasks including semantic similarity estimation, analogy detection, and other downstream natural language processing applications.

The early stage of neural networks is a class of computational methods inspired by the biological structure and function of interconnected neurons in the nervous system. These architectures represent a fundamental paradigm in machine learning, characterized by their ability to learn complex, non-linear mappings between input and output spaces through adaptive parameterization. Inspired by signals in the human brain, neural networks—such as the one shown in part (b) of Figure 2.1—are developed. There are two parts to this transformation: linear and non-linear. In the linear section, the input $\mathbf{x}$ is multiplied by a weight matrix $\mathbf{W}$ and added to a bias vector $\mathbf{b}$. This operation can be mathematically represented as:

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b} \tag{2.1}$$

where $\mathbf{z}$ is the result of the linear transformation, $\mathbf{W}$ represents the weights, and $\mathbf{b}$ the bias term. Weights and bias are both trainable parameters that are optimized in the process of calculation.

However, in many real-world scenarios, for instance, language understanding, the relationships between input features and target outputs are not linearly correlated. To model such complex patterns, a non-linear layer is introduced. This is achieved by sending the output of the linear transformation through a non-linear function, known as the activation function. The final output $\mathbf{a}$ of the neuron is thus computed as:

$$\mathbf{a} = \sigma(\mathbf{z}) = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{2.2}$$

Here, $\sigma(\cdot)$ represents a non-linear activation function such as Rectified Linear Unit (ReLU) (Agarap, 2018), Tanh, or Sigmoid (Rumelhart et al., 1986). This non-

1. https://web.stanford.edu/~jurafsky/slp3/3.pdf

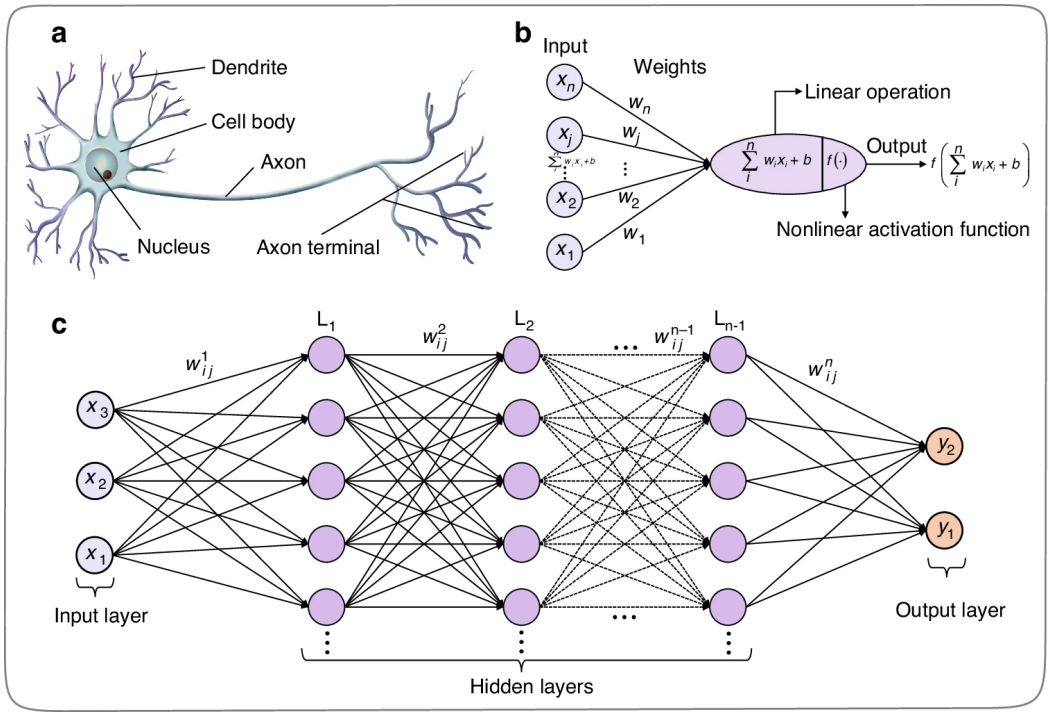



**Figure 2.1**: Illustrative figure for neural network with relation to human brain's nervous system, from work by Fu et al. (2024).

linearity allows neural networks to approximate complex, hierarchical functions beyond linear transformations. In the equation

$$\mathbf{a} = \sigma(\mathbf{z}) \tag{2.3}$$

, where $\sigma(\cdot)$ denotes an activation function applied element-wise to the input vector $\mathbf{z}$. An example is the ReLU, which is widely used in modern neural networks thanks to its computational efficiency and ability to mitigate the vanishing gradient problem. It is formally defined as,

$$\text{ReLU}(x) = \max(0, x) \tag{2.4}$$

This function outputs zero for all negative inputs and passes positive values unchanged, enabling sparse representations and improved gradient flow during training (Agarap, 2018). Interested readers could look further into this review (Rasamoelina et al., 2020) for the collection of modern activation functions used in NN.

To model complex patterns in data, neural networks commonly employ a structure known as a *multilayer perceptron* (MLP), which stacks multiple layers of transformations. As depicted in part (c) of Figure 2.1, the network consists of several layers of interconnected neurons. The neurons in the first hidden layer (L1) receive the input vector and apply a linear transformation followed by a non-linear activation, as defined in Equation 2.2. The resulting activations are then propagated to the next hidden layer (e.g., via weight matrix $\mathbf{W}_2$), which performs its own transformation. This process continues sequentially through the layers until the output layer is reached. The number of hidden layers and the number of neurons within each layer are hyperparameters that can be tuned based

on the task. This layered structure allows the network to progressively extract and represent increasingly abstract features from the input.

Such architectures are commonly referred to as *feed-forward neural networks* (FFNNs), since the input signals travel only forward through the network layers without any backward connections. In the illustration, the layer transforms three input features into two output features, typically by applying a non-linear activation function such as softmax (Goodfellow et al., 2016) to the last layer, which converts the raw output scores into probabilities that sum to one. This makes softmax especially useful for classification tasks where the output represents the likelihood of each class.

Mathematically, the softmax function for an output vector $\mathbf{z} = [z_1, z_2, \dots, z_n]$ is defined as:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{2.5}$$

Usually, there is a big difference between the output and the true value. And a **loss function**, formally defined as: $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ quantifies the discrepancy between predicted outputs $\hat{y}$ and ground truth targets $y$, where $\mathcal{Y}$ represents the output space. The loss function serves as an optimization objective, providing a differentiable measure of model performance that enables gradient-based learning algorithms.

For multi-class classification problems with $C$ classes, the **cross-entropy loss** is defined as:

$$\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{2.6}$$

where: $y = [y_1, y_2, \dots, y_C]^T$ is the one-hot encoded true label vector with $y_i \in \{0, 1\}$ and $\sum_{i=1}^{C} y_i = 1$ - $\hat{y} = [p_1, p_2, \dots, p_C]^T$ is the predicted probability distribution with $p_i \in [0, 1]$ and $\sum_{i=1}^{C} p_i = 1$

For binary classification, this reduces to the **binary cross-entropy loss**:

$$\mathcal{L}_{BCE}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{2.7}$$

where $y \in \{0, 1\}$ denotes the ground-truth label and $\hat{y} \in [0, 1]$ represents the predicted probability of the positive class.

The objective of the model is to reduce the loss, which enables the formulation of the problem:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(f_\theta(x_n), y_n) + \lambda R(\theta) \tag{2.8}$$

where

- $\theta$ represents the model parameters,

- $f_\theta(x_n)$ is the model's prediction for input $x_n$,
- $N$ is the number of training samples, and
- $R(\theta)$ is a regularization term with hyperparameter $\lambda$, which helps prevent overfitting by penalizing large or complex model parameters.

The gradient of the loss function $\nabla_\theta \mathcal{L}$ provides the direction for parameter updates during backpropagation (Rumelhart et al., 1986). Specifically, the gradient vector $\nabla_\theta \mathcal{L} = \left[\frac{\partial \mathcal{L}}{\partial \theta_1}, \frac{\partial \mathcal{L}}{\partial \theta_2}, \ldots, \frac{\partial \mathcal{L}}{\partial \theta_d}\right]^T$ represents the partial derivatives of the loss with respect to each parameter $\theta_j$, indicating the rate of change of the loss function in the parameter space. This gradient information enables the gradient descent optimization process, where the model parameters are updated iteratively according to

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t) \tag{2.9}$$

where $\eta$ is the *learning rate*, a hyperparameter that controls the step size of each update, and $t$ denotes the *iteration step*, indicating the progression of updates during the training process.

### 2.2.2 Introducing Neural Networks to Language Modeling

Language modeling can be defined as a model that predicts the next possible word from the vocabulary. Given this probabilistic nature, neural networks are natural choices for training language models, as they can be optimized to output such distributions. The introduction of neural networks into language modeling began with Bengio et al.'s proposal of a neural probabilistic language model (Bengio et al., 2000), marking a key milestone in natural language processing.

In the neural probabilistic language model, as illustrated in Figure 2.2, words are vectorized and processed through hidden layers to predict subsequent words. This architecture introduced learnable *word embeddings* as inputs to a feedforward network generating next-word probability distributions. The model is trained on a joint learning manner of embeddings and network parameters on the pre-training task of next word prediction. After the training round, there are two products that could be used, a model which can predict next word and word embeddings which could be used for downstream tasks such as similarity calculation, etc.

This foundational work catalyzes the development of more computationally efficient embedding methods, notably Word2Vec (Mikolov, Chen, et al., 2013) and GloVe (Pennington et al., 2014), which decouples representation learning from the language modeling objective to generate high-quality word vectors at scale. These approaches enhanced training efficiency through innovations such as hierarchical softmax, negative sampling, and global matrix factorization techniques. They also extend the pre-training task from next word prediction to contextual word prediction (guess the neighboring words), etc. While these methods fall outside the scope of this dissertation, given their limited applicability to contemporary transformer-based architectures, interested readers are referred to the corresponding publications for further reading.

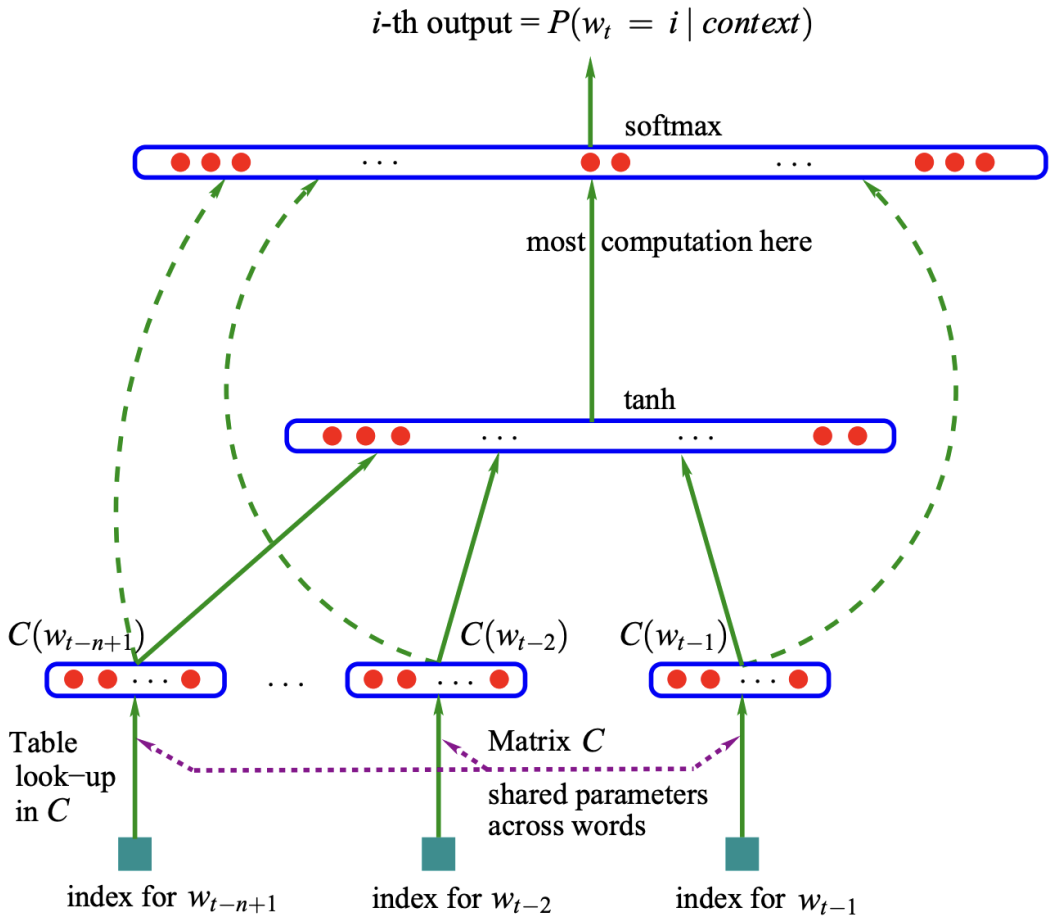



**Figure 2.2:** A multilayer perceptron-based language model, as proposed by Bengio et al. (2000). The figure is from their original article.

The evolution of neural architectures subsequently encompassed Recurrent Neural Networks (RNNs) (Rumelhart et al., 1986), Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), and Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; LeCun et al., 1989; Zhang et al., 1988). They are architecturally optimized to capture distinct structural regularities in data. RNNs demonstrate particular efficacy in modeling sequential dependencies inherent in textual and temporal data, while CNNs excel at extracting local feature patterns, proving highly effective for image processing and sentence-level classification tasks. The fundamental principles and paradigms - training neural networks on big corpora for learning word embedding and model parameters, established by these early neural language models, provided the theoretical and methodological foundation for subsequent architectural innovations. These advances encompasses not only traditional next-word prediction paradigms but also bidirectional context modeling and exploration of network architectures, ultimately culminating in the transformer architectures that underpin contemporary LLMs.

### 2.2.3 Transformer Architecture

The transformer architecture introduces a paradigmatic shift in language modeling through its novel combination of a **multi-head attention mechanism** and **encoder-**

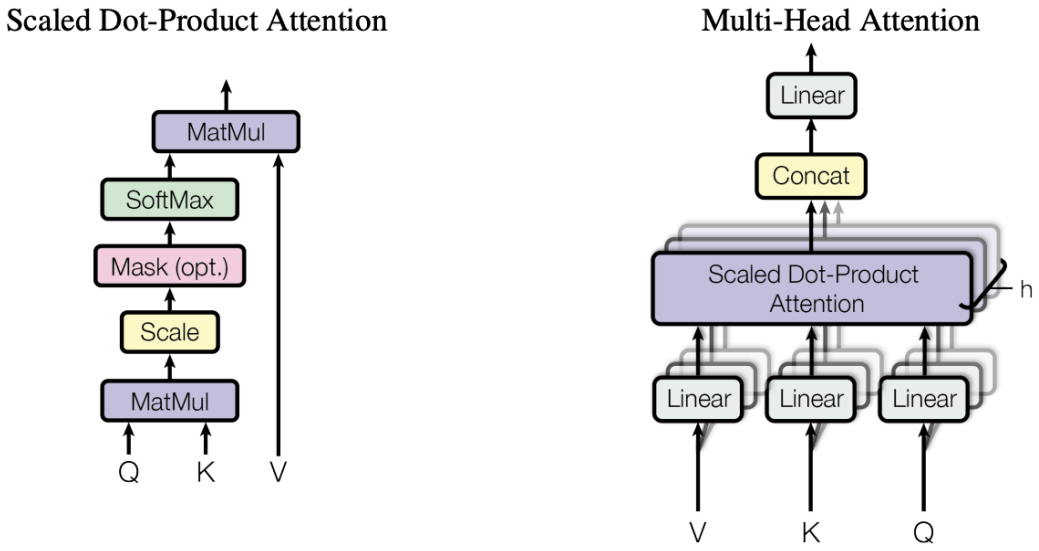



**Figure 2.3**: Scale Dot-Product Attention and Multi-Head Attention mechanism, figure is reproduced from original author (Vaswani et al., 2017).

**decoder** framework (Vaswani et al., 2017), fundamentally departing from recurrent and convolutional approaches. This architectural innovation establishes the foundation for subsequent breakthroughs, including BERT (Devlin et al., 2019a) and the GPT series (OpenAI, 2023; Radford et al., 2018), which strategically leverage distinct components of the Transformer for specialized training objectives.

The **attention mechanism** represents a shift in how neural networks process sequential information. Rather than compressing an entire input sequence into a single fixed representation as in FFNN, attention allows models to dynamically put weight on different parts of the input when making predictions (Bahdanau et al., 2015). More specifically, attention mechanisms compute relevance scores that determine how much weight to assign to each part of the input sequence. The scaled dot-product attention mechanism decomposes each input vector into three fundamental components through linear projection: queries, keys, and values of dimension, where the queries and keys are used to calculate the importance of each element. The computation proceeds by calculating dot products between the query vector of the input at the current time step and all key vectors from other input tokens, scaling each result by $\frac{1}{\sqrt{d_k}}$, and applying a softmax function to derive attention weights. These normalized weights are then used to compute a weighted sum of the corresponding value vectors, producing the final attended representation. This process is illustrated in Figure 2.3.

The complete attention operation is mathematically expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.10}$$

where $Q$ refers to the query matrix, $K$ represents the key matrix, and $V$ denotes the value matrix. The term $QK^T$ computes the dot product between all query-key pairs between all input with the input at current time step, which are then scaled by $\frac{1}{\sqrt{d_k}}$ and normalized through the softmax function to produce attention weights that are applied to the value matrix $V$. This matrix formulation enables

parallel computation of attention scores across all query-key pairs, with the resulting attention weights applied to aggregate information from the value vectors through matrix multiplication.

Multi-head attention refers to applying multiple learned linear projections of the queries, keys, and values. Each head is computed separately with the scaled dot-product attention. Mathematically, this is expressed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.11}$$

Then the output from different heads is concatenated with:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.12}$$

The projection matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, and $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ map the input representations into $h$ different subspaces, typically with $d_k = d_v = d_{model}/h$. The final output projection matrix $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ combines the concatenated head outputs back to the original model dimension, enabling the model to jointly attend to information from different representation subspaces at different positions.

The multi-head attention is a crucial composing module for the encoder-decoder architecture of transformer models, enabling parallel processing of different types of relationships within and across sequences, as depicted in the figure 2.4. The encoder (the left block) processes the input sequence through multiple layers of attention and FFNN to generate contextualized representations for each position in the sequence. The decoder (the right block) utilizes both the encoded representations from the encoder and previously generated tokens to produce a probability distribution over the output vocabulary at each time step. This encoder–decoder framework enables the model to first construct a contextualized representation of the input sequence in the encoder by leveraging the attention mechanism, which captures long-range dependencies and semantic relationships across tokens. The decoder then generates the output sequence autoregressively, using masked self-attention to ensure causality and cross-attention to selectively focus on relevant parts of the encoder's representation. This design allows the model to both comprehensively understand the input and produce coherent, contextually aligned outputs.

Given that both encoder and decoder modules demonstrate significant capability in capturing semantic representations of sequences, different architectural variants have emerged based on different combinations of these components. Contemporary language models can be categorized into three primary architectures: encoder-only models (such as BERT (Devlin et al., 2019a)), decoder-only models (such as GPT (Radford et al., 2018)), and encoder-decoder models (such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020a)). Each architecture is optimized for distinct classes of natural language processing tasks through specialized training objectives.

As illustrated in Figure 2.5, BERT employs a bidirectional encoder architecture, enabling the model to access contextual information from both preceding and succeeding tokens when processing each position in the sequence. In contrast, decoder-only models like GPT operate autoregressively, restricting access to only

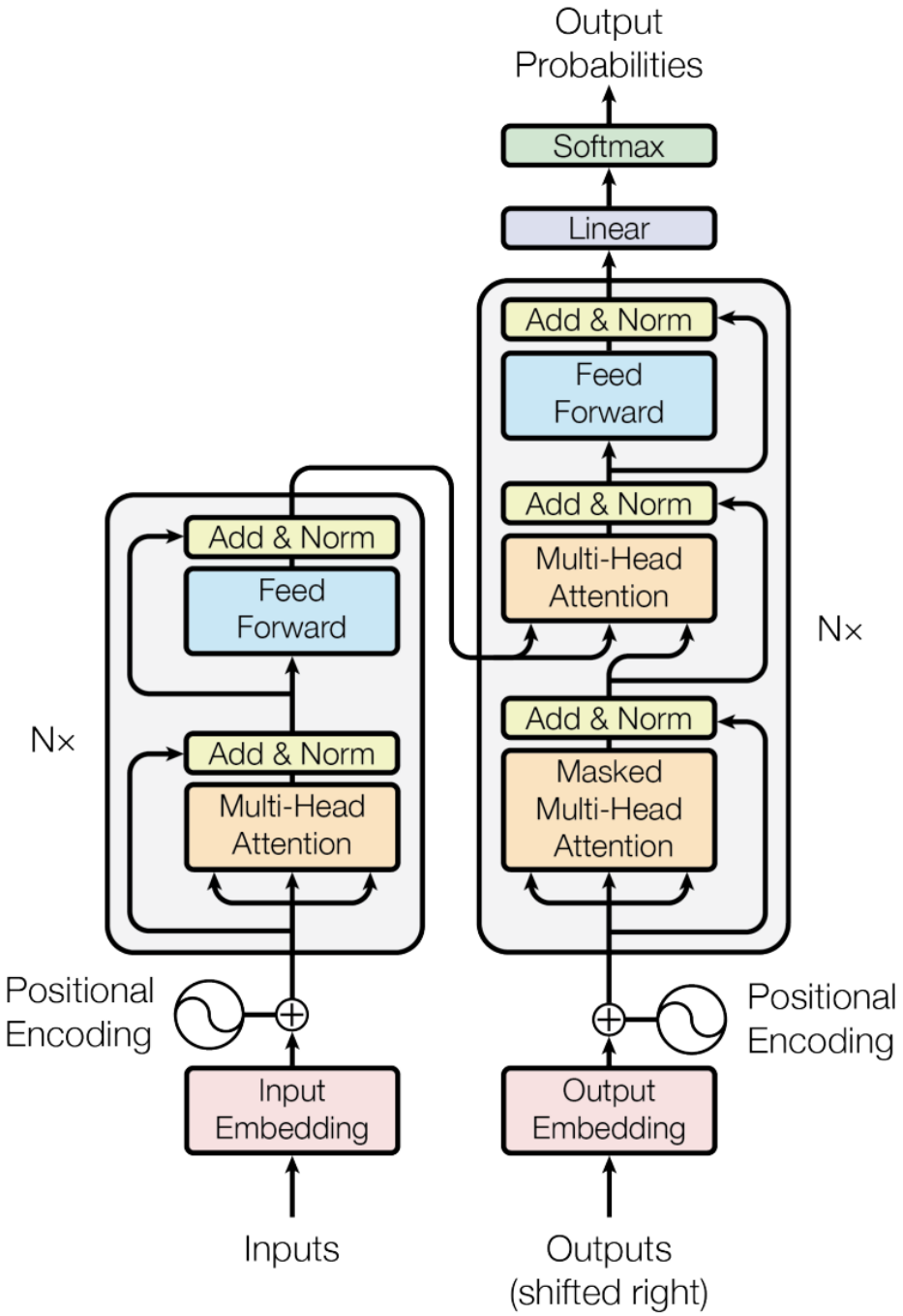
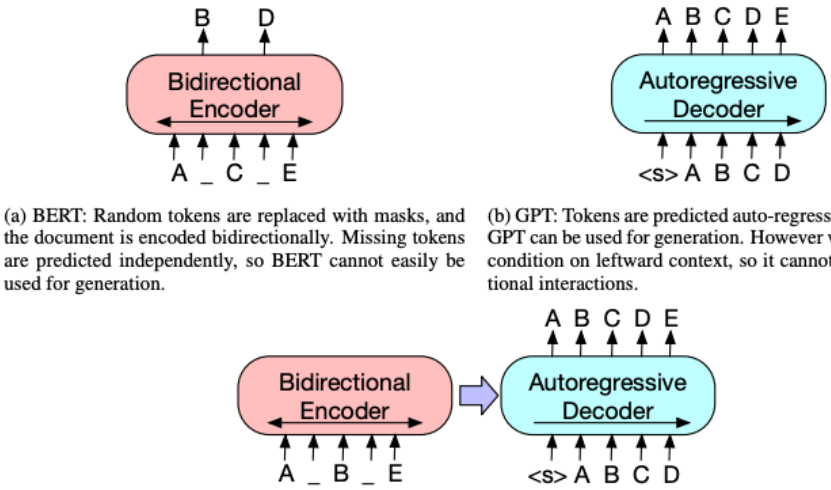



**Figure 2.4**: Transformer architecture, figure is reproduced from original author (Vaswani et al., 2017).




(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

**Figure 2.5:** A comparison of BERT (Devlin et al., 2019a), GPT (Radford et al., 2018), and BART (Lewis et al., 2020)'s training architecture, figure is reproduced from the original author of BART (Lewis et al., 2020).

previously generated tokens during inference, which naturally aligns with left-to-right text generation. BART combines both paradigms through its encoder-decoder architecture: the encoder processes input bidirectionally to create comprehensive contextual representations, while the decoder generates output autoregressively by attending to both the encoded representations and previously generated tokens.

These architectural distinctions directly influence task performance, with encoder-only models excelling at understanding and classification tasks, decoder-only models demonstrating superior text generation capabilities, and encoder-decoder models proving particularly effective for sequence-to-sequence tasks such as summarization, translation, and text infilling that require both comprehension and generation.

In the later chapter 5, we employ BART-Large and introduce a novel task that converts information from the knowledge graph into complete sentences, corrupts key information, and prompts the model to predict the missing content. This approach remains consistent with BART's original denoising training scheme.

**Bidirectional and Auto-Regressive Transformers (BART)** (Lewis et al., 2020) is a language model that combines the strengths of both encoder and decoder architectures within a Transformer framework. We discuss here in terms of its architectural characteristics and pre-training objective, which are among the reasons for its use in chapter 5. A distinctive feature of BART is its denoising autoencoder training objective, which trains the model to reconstruct the original text from a corrupted version. During pretraining, the input text is intentionally corrupted using strategies such as token masking, token deletion, sentence permutation, and document rotation. The corrupted sequence is passed through the encoder–decoder structure, and the model is trained to recover the uncorrupted text. This

denoising process equips BART with the ability to handle noisy or incomplete inputs, enhancing its adaptability to a wide range of downstream tasks.

BART has shown strong performance across generative and comprehension-based NLP tasks, including summarization, translation, question answering, and dialogue. Its flexibility allows fine-tuning as a text-to-text model or as encoder-only/decoder-only, depending on the task.

Contemporary language models represent iterative refinements of transformer architectures, primarily through architectural modifications—such as scaling depth, width, and attention mechanisms—and diversified training paradigms including instruction tuning (Ouyang et al., 2022), reinforcement learning from human feedback (Griffith et al., 2013), and multi-task learning (Chen, Zhang, et al., 2024). These evolutionary developments, rather than revolutionary changes, have culminated in the sophisticated language models deployed today, demonstrating the enduring influence and adaptability of the original transformer design principles.

### 2.2.4 Large Language Models

These developments, along with the discovery of scaling laws (Kaplan et al., 2020)—which shows that model performance improves predictably with increased parameter count, dataset size, and training computation—have fueled a surge in the development of foundation language models (Bommasani et al., 2021), also known as LLMs. As a result, modern LLMs are capable of performing a wide range of tasks and are now widely adopted across diverse domains and applications without pre-training private language models from scratch.

**Mixtral** is one example of this model. It is a family of large language models introduced by Mistral AI that adopts a mixture-of-experts (MoE) (Jiang, Sablayrolles, Mensch, Bamford, Chaplot, de Las Casas, et al., 2023) architecture to improve both computational efficiency and performance. Mixtral activates only a subset of its parameters during inference—specifically, two out of eight networks per forward pass—thereby significantly reducing computational requirements while maintaining high accuracy across a range of natural language understanding and generation tasks. See Figure 2.6 for an illustration. This sparse activation mechanism allows the model to scale more effectively without proportionally increasing inference cost. At the same time, it reuses the sliding multi-head attention to ensure the efficiency of training. Built on a decoder-only Transformer architecture, Mixtral has demonstrated strong performance on various NLP tasks and reasoning tasks. Therefore, we use Mixtral for question generation in the last chapter of this dissertation.

**GPT-3.5 Turbo** is a large-scale decoder-only Transformer model developed by OpenAI (OpenAI, 2023), optimized for conversational AI and multi-turn dialogue. It is trained using a combination of supervised fine-tuning and Reinforcement Learning from Human Feedback (RLHF) (Griffith et al., 2013), enabling strong instruction-following behavior. The model processes inputs as structured `system`, `user`, and `assistant` messages, which allows fine-grained control over tone and behavior.

The training procedure leverages the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017), a reinforcement learning method that adjusts the model's output probabilities to better align with human preferences while avoiding

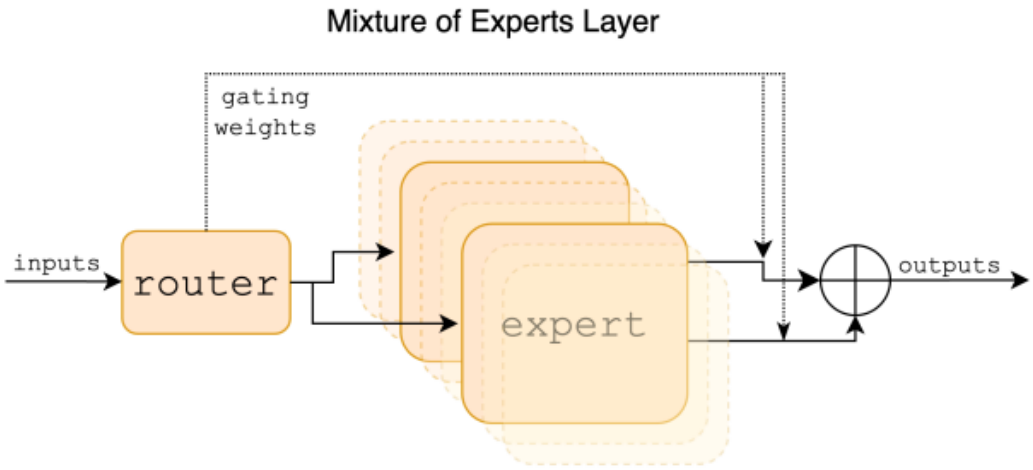



**Figure 2.6:** Mixture of Expert layer from Mixtral original paper (Jiang, Sablayrolles, et al., 2024).

large, destabilizing parameter updates. This, combined with high-quality human-annotated data, ensures more stable convergence and improved generation quality. Despite these advancements, GPT-3.5 Turbo has been criticized for generating information not directly supported by the input (Mohammed et al., 2024). This phenomenon, known as hallucination (Ji et al., 2023), challenges the reliability and factual accuracy of the model's outputs, posing significant concerns for applications requiring trustworthy information.

The combination of decoder-only architectures and scaling enables LLMs to achieve task-agnostic text generation performance through **few-shot learning** or zero-shot learning (Kojima et al., 2022; Wang, Yao, et al., 2020), eliminating the need for task-specific retraining. In this paradigm, the model leverages its pre-trained knowledge to infer the desired behavior directly from a handful of input–output examples provided in the prompt (few-shot) or from natural language task descriptions without any examples (zero-shot). This capability to achieve high performance via model interaction without parameter modification has established prompt engineering and prompt tuning as distinct research fields focused on optimizing input demonstrations for enhanced task performance (Lester et al., 2021; Wang et al., 2023).

Among LLMs, despite the existence of various architectural configurations, for instance, encoder-only, decoder-only, and encoder-decoder—the majority, including GPT series (Klager and Polleres, 2023b; OpenAI, 2023), LLaMA series (Grattafiori et al., 2024; Touvron, Lavril, Izacard, Martinet, Lachaux, Lacroix, Rozière, Goyal, Hambro, Azhar, et al., 2023a), Claude series (Anthropic, 2024), and Mixtral (Jiang, Sablayrolles, et al., 2024; Jiang, Sablayrolles, Mensch, Bamford, Chaplot, de las Casas, et al., 2023), have adopted the decoder-only architecture, , which stacks the decoder component of the Transformer as the core of the language model. While several studies have compared these architectures in specific downstream tasks, such as word understanding (Qorib et al., 2024), relation extraction (Sarrouti et al., 2022), text generation (Cai et al., 2022), and document analysis (Cong et al., 2019), a comprehensive theoretical or empirical evaluation of their relative effectiveness remains largely unexplored.

Due to the substantial computational resources and vast amounts of data required for training LLMs, it remains prohibitively expensive for many in the research community to independently develop such models. As a result, most state-of-the-art LLMs are either developed by, or in close collaboration with, large technology companies. In the case of closed-source models, companies often withhold crucial details such as the exact composition of training data, data cleaning methodologies, and architectural choices—making full replication and transparency difficult for the broader community (Kukreja et al., 2024). The models employed in this dissertation are primarily open-source, including BART, LLaMA, and Mixtral. However, we also leverage the GPT series in chapter 7, given its outstanding performance in language generation.

## 2.3 Knowledge Graph Question Answering

KGQA is the task of retrieving answers from a knowledge graph given a natural language question. To fully understand this task, we cover the definition of knowledge graphs, the KGQA problem, classifications of KGQA challenges, common solution approaches, and the role of information extraction modules essential to all KGQA systems.

### 2.3.1 Knowledge Graph

A **knowledge graph (KG)** is a structured representation of data that encodes real-world knowledge in a machine-interpretable form. It is typically organized around a manually defined *schema (***s***)*, which specifies the types of entities and relations, thereby constraining and giving precise meaning to the stored facts. Formally, the schema is defined as

$$S = (C, R_s, A)$$

where $C$ denotes the set of classes (e.g., *Symptom*, *Disease*, *Drug*), $R_s \subseteq C \times R \times C$ is the set of admissible relations between classes (e.g., *Drugs*, *treats*, *Disease*), and $A \subseteq C \times D$ represents the attributes associated with classes, with $D$ being the set of datatypes (e.g., *molecularWeight*).

On top of this *schema level*, a KG contains *instances*, which populate the abstract classes and relations. The instance level can be represented as

$$I = (V_i, E_i, L)$$

where $V_i \subseteq V$ is the set of entity instances of classes in $C$, $E_i \subseteq V_i \times R \times V_i$ is the set of factual relations between instances, and $L \subseteq V_i \times A \times \Delta$ denotes literal statements, with $\Delta$ being the domain of data values (e.g., strings, numbers, dates, booleans). For example, the triple (“Marie Curie”, *birthDate*, “1867-11-07”) $\in L$ links an entity to a literal value.

By integrating schema definitions $S$, instance data $I$, and literal values $\Delta$, knowledge graphs yield a semantically rich and interpretable representation of information. In many cases, these graphs are formalized as *RDF* (Resource Description Framework) graphs (Lassila and Swick, 1999), where knowledge

is represented as triples of the form (*subject*, *predicate*, *object*). Their structured design and inherent explainability make KG a powerful paradigm for organizing and accessing large-scale knowledge.

To retrieve knowledge stored in a knowledge graph, users commonly employ query languages such as *SPARQL*(Pérez et al., 2009), *Cypher* (Francis et al., 2018), or similar declarative query interfaces.

**SPARQL** (SPARQL Protocol and RDF Query Language) (Pérez et al., 2009) is the W3C-standardized query language for retrieving and manipulating data stored in *RDF* (Resource Description Framework) graphs, which are a common formalism for representing knowledge graphs. SPARQL is graph-oriented and allows users to specify query patterns over triples of the form (*subject*, *predicate*, *object*).

SPARQL supports a wide range of operations, including pattern matching, filtering, aggregation, and graph traversal, which allow fast access and edit of the KG. For example, the following query retrieves the birth date of *Marie Curie*:

```
SELECT ?birthDate
WHERE {
  ?person rdfs:label "Marie Curie"@en .
  ?person dbo:birthDate ?birthDate .
}
```

Here, the query searches for a resource labeled “Marie Curie” and returns the value associated with its *birthDate* property.

Due to its ability to express complex graph patterns in a declarative manner, SPARQL has become the standard query mechanism for large-scale knowledge graphs, including widely used resources such as DBpedia (Bizer et al., 2009) and Wikidata (Erxleben et al., 2014).

The graph-based structure of KGs enables logical reasoning, allowing the exploitation of implicit knowledge encoded within them. Furthermore, the use of vocabularies with well-defined semantics supports the integration of heterogeneous data sources. This facilitates the combination of even domain-specific KGs; for instance, in the biomedical domain, *PrimeKG* (Chandak et al., 2023a) integrates multiple biomedical knowledge graphs through mapping and linking, leveraging both domain-specific vocabularies and general semantic standards. Such interoperability enhances the applicability of KGs in downstream tasks, including personalized medicine.

### 2.3.2 Knowledge Graph Question Answering

KQGA models aim to retrieve answers to natural language queries directly from knowledge graphs. This enables users to access accurate and grounded information stored in databases without requiring prior knowledge of the underlying schema or query languages.

KGQA problems can be classified in various ways. One common approach is to distinguish between *simple* and *complex* KGQA, based on the number of triples involved in the reasoning chain. A *simple* question is one whose answer can be retrieved from a single factoid statement containing only one relation or predicate (Yani and Krisnadhi, 2021). In contrast, *complex* questions include those

with additional constraints or those requiring multiple hops across relations to arrive at the answer (Song et al., 2023).

Depending on the underlying knowledge graph, KGQA problems can be classified as either *open-domain* or *domain-specific*. Open-domain questions are typically answerable from large encyclopedic KGs such as Wikidata (Vrandečić and Krötzsch, 2014), DBpedia (Auer et al., 2007a), or Freebase (Bollacker et al., 2008a). In contrast, domain-specific datasets focus on specialized areas such as biomedicine, movies, science, or events. Since KGQA performance is heavily dependent on the availability and quality of the underlying KGs, well-established open-domain KGs have spawned numerous KGQA datasets; for instance, 17 KGQA datasets (Perevalov* et al., 2022) are dependent on Wikidata, which places particular emphasis on multilingual aspects, temporal information, etc. However, domain-specific KGs often have only a single available dataset, which is typically used solely for evaluation. In the biomedical domain, despite the existence of numerous structured KGs—thanks to the efforts of biologists in personalized medicine and FAIR data maintenance (Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, Blomberg, Boiten, da Silva Santos, et al., 2016)—the number of publicly available KGQA datasets remains limited. This gap in the availability of large-scale and complex biomedical KGQA datasets is one of the primary motivations of this dissertation, which aims to bridge this shortage by contributing new resources and methodologies tailored to the biomedical domain.

## 2.4 Information Extraction

As discussed in Section 1.1, solutions to the KGQA problem are generally categorized into two main paradigms: *query generation*-based and *information retrieval*-based approaches. Regardless of the paradigm, both fundamentally rely on *information extraction* as an essential first step to enable effective downstream processing. For instance, in Figure 1.2, the matched entity *Hamburg* and the relation *Population* serve as fundamental components for composing a query and aligning it with the knowledge graph. Therefore, in this section, we focus on key information extraction methods in KGQA, particularly *named entity recognition*, *entity linking*, and *relation extraction*.

### 2.4.1 Named Entity Recognition

Named Entity Recognition (NER) has a long-standing history in natural language processing. Its primary goal is to extract and classify mentions of rigid designators from text, such as person names, locations, organizations, and miscellaneous entities (Nadeau and Sekine, 2007). For example, in the question shown in Figure 1.2, the entity `Hamburg` should be recognized and classified as `LOC`. In some cases, an NER system may only detect the surface form `Hamburg` without assigning a label such as `LOC`; this process is also known as mention detection. NER is widely applied in various tasks, including question answering, machine translation, and text summarization.

In the biomedical domain, however, the entities of interest differ significantly from those in the general domain. Instead of `LOC` or `PERSON`, the common classes in

biomedical NER (BioNER) include `gene`, `cell`, `protein`, `disease`, and `chemical`, which cover almost all entity types addressed in BioNER tasks (Li et al., 2022). For a more detailed description of BioNER entity types, we refer the reader to Table 2 of the BioNER review in Song et al. (2021).

In both general and biomedical domains, NER is commonly formulated as a sequence tagging problem. This involves assigning a label to each token in a sentence to indicate whether it is part of an entity and what type of entity it belongs to. Popular tagging schemes include BIO, BIOES, and related variants (Chang et al., 2022).

The BIO scheme stands for **B**eginning, **I**nside, and **O**utside. Each token is labeled as:

- **B** if it is the beginning token of a named entity,
- **I** if it is inside a named entity but not the first token,
- **O** if it is outside any named entity.

For example, in the phrase "acute myocardial infarction," the tokens might be tagged as B-Disease, I-Disease, I-Disease.

The BIOES scheme further refines this approach by adding two more labels:

- **E** for the end token of an entity,
- **S** for a single-token entity.

This provides more precise boundary detection, especially useful for complex biomedical entities that often vary in length.

These tagging schemes enable sequence models such as Conditional Random Fields (CRFs) and Transformers to effectively learn entity boundaries and types, which is crucial for extracting structured information from unstructured text.

### 2.4.2 Entity Linking

Entity Linking (EL) refers to the process of identifying an entity mention in text and linking it to its corresponding entry in a KG. For example, the mention *Hamburg* could be linked to the Wikidata entity `Q1055`, distinguishing it from other possible meanings such as Hamburg University, Hamburg football club, etc. This process typically follows NER, where entity mentions are first detected; EL then disambiguates them and resolves each to a unique KG identifier. Note that some EL approaches incorporate the mention detection step directly, often referred to as *end-to-end* EL (Glasmachers, 2017). In the biomedical domain, EL can help to disambiguate medical terms and link to rich semantic information in the biomedical KG, which can act as an essential means for many downstream applications.

Due to the vast number of entities present in a KG, it is therefore computationally costly to directly compare an extracted mention against all possible entities. Consequently, the EL process typically involves two steps: *candidate generation* and *candidate ranking*. The candidate generation phase generates relevant candidate entities based on the entity mentions extracted from NER. The candidate ranking stage aims at giving relevancy scores based on the extracted

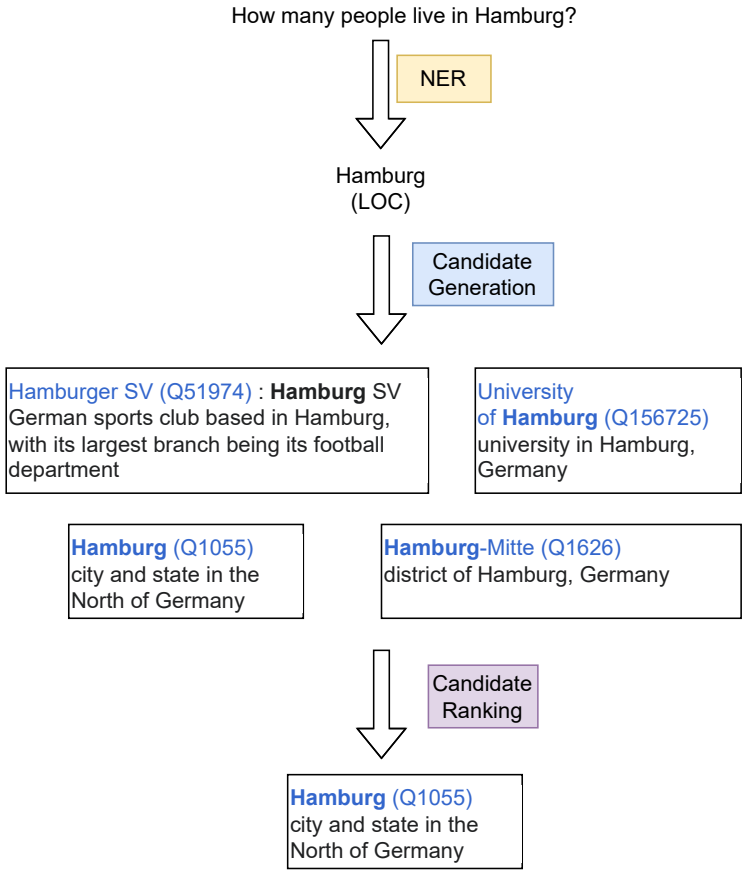



**Figure 2.7:** A pipeline for entity linking on the sentence "How many people live in Hamburg?" The candidates are from Wikidata.

mentions and candidates from the last step, and ranks the right one higher. This process is illustrated in Figure 2.7.

According to Shi et al. (2023), biomedical EL approaches can be broadly categorized into three types: rule-based, machine learning (ML)-based, and deep learning (DL)-based methods. Rule-based EL is primarily implemented using string-matching or dictionary look-up techniques to disambiguate entities. String-matching methods typically define templates that transform text inputs into features based on linguistic rules, such as spelling rules, word-formation rules, indicator words, and prefix/suffix patterns. In contrast, dictionary-based methods leverage curated lexical resources to enrich linguistic features with vocabulary abbreviations, variants, synonyms, and other lexical forms.

ML-based methods represent mentions and candidates as vectors through manual design or statistical features and formulate EL as a traditional machine learning problem, such as classification or learning-to-rank. Classification-based approaches train a model to determine whether a mention is positively or negatively associated with candidate entities in a KG. To mitigate the limitations of binary classification, many systems adopt learning-to-rank techniques, which rank candidate entities based on their likelihood of being the correct link.

More recent methods leverage DL thanks to its ability to automatically learn feature representations and generalize across domains. These approaches typically use bi-encoders to map both the input mention and candidate entities into a shared vector space. The resulting vectors are then fed into a neural similarity function, which assigns higher scores to correct mention–entity pairs while pushing apart unrelated ones. By jointly learning the representation and similarity scoring functions, DL-based EL models can capture complex semantic relationships beyond surface-level lexical matching, leading to substantial performance improvements in biomedical domains.

Biomedical EL faces challenges across all three methodological categories, primarily due to domain-specific linguistic characteristics and the scarcity of corpora capable of teaching LLMs deep semantic features (Shi et al., 2023). Consequently, disambiguating mentions and candidates—often similar on the surface but contextually distinct—remains difficult. One approach to address this is to enrich the LLM with structured knowledge by converting KGs into training corpora, enabling the model to capture deeper semantic relations. We explore this approach under a novel setting in our Chapter 5, where predicting the synonym, definition as well as the relation based on head and tail entity is part of the pre-training objective. In this way, we are able to teach the LLM semantics of biomedical terms and deep relations.

### 2.4.3 Relation Extraction

The process of extracting such relations between entities mentioned within or across sentences (or documents) is known as relation extraction (RE). Each relation is typically represented as a triplet `(head_entity, relation, tail_entity)`, consisting of two entities and the semantic relation linking them. RE enables the transformation of unstructured text into structured knowledge, making it closely related to tasks such as knowledge graph completion and KGQA. As shown in figure 1.2, the relation `Population (P1082)` is also essential for formulating search queries and identifying potential subgraphs.

Depending on whether the identified entities and relations are connected to an underlying KG, RE can be classified into open information extraction and closed information extraction. Open information extraction approaches extract related phrases as representations of relations and entities directly from the text. While close RE classifies the relation into pre-defined categories.

Also, relations do not only exist within a single sentence but can span across sentences, documents, and even modalities. In chapter 6, we define three types of RE: in-sentence, in-document, and cross-document. In-sentence relation extraction focuses on entities occurring within a single sentence. In-document relation extraction deals with a single document on a specific topic containing multiple sentences. Finally, cross-document relation extraction captures relations between entities that appear in separate documents. See Figure 2.8 for three types of RE. However, nowadays relations also exist beyond text. Multimodal relation extraction (Han et al., 2020) aims to extract facts that connect textual information with other modalities, such as images, which can also store factual knowledge.

Since RE relies on identifying relations between entities, NER is a prerequisite for this task. Depending on whether NER and RE are trained jointly, RE methods

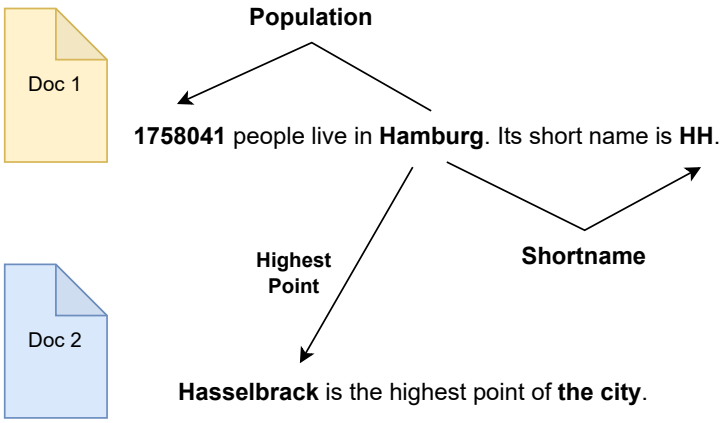



**Figure 2.8:** In-sentence, cross-sentence,d and cross-document relation extraction.

can be classified into two categories (Zhao et al., 2024): *pipeline-based* and *joint* approaches.

*Pipeline-based* methods perform entity recognition and relation extraction in two separate stages—first detecting entities, then predicting the relation between each entity pair. Some pipeline methods use off-the-shelf NER modules and train a relation classifier based on their output, while others use gold-standard entities from the dataset to train the classifier. Errors in NER often propagate and negatively affect the performance of RE.

While *joint* approaches aim to design frameworks that recognize and classify relations simultaneously, these models are optimized to accurately identify entities and relation pairs, making them less prone to error propagation during the Named Entity Recognition stage. This paradigm has gained popularity recently, largely due to advancements in LLMs, which foster multitask learning and sequence labelling (Hillebrand et al., 2022; Zeng et al., 2020).

Despite the progress in applying LLMs to RE, there remain significant challenges, such as the lack of sufficient data in specific domains. The main drawbacks of PLMs include their resource-intensive nature, requiring substantial computational power for both training and inference, and their tendency to overfit on smaller or domain-specific datasets. Therefore, in Chapter 6, we explore approaches to enhance PLMs with symbolic information, either by creating distant supervised data or by improving the model architecture.

# 3

# Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis

## Bibliographic Information

Aleksandr Perevalov*, Xi Yan*, Liubov Kovriguina, Longquan Jiang, Andreas Both, and Ricardo Usbeck. 2022. Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, 2998–3007. * Shared first authorship. Marseille, France: European Language Resources Association, June. (Cited on pages 7, 10 sq., 28).

## Abstract

Data-driven systems need to be evaluated to establish trust in the scientific approach and its applicability. In particular, this is true for Knowledge Graph (KG) Question Answering (QA), where complex data structures are made accessible via natural-language interfaces. Evaluating the capabilities of these systems has been a driver for the community for more than ten years while establishing different KGQA benchmark datasets. However, comparing different approaches is cumbersome. The lack of existing and curated leaderboards leads to a missing global view over the research field and could inject mistrust into the results. In particular, the latest and most-used datasets in the KGQA community, LC-QuAD and QALD, miss providing central and up-to-date points of trust. In this paper, we survey and analyze a wide range of evaluation results with significant coverage of 100 publications and 98 systems from the last decade. We provide a new central

and open leaderboard for any KGQA benchmark dataset as a focal point for the community - https://kgqa.github.io/leaderboard/. Our analysis highlights existing problems during the evaluation of KGQA systems. Thus, we will point to possible improvements for future evaluations.

## 3.1 Introduction

Question Answering (QA) is a rapidly growing field in research and industry[1]. QA systems already deliver their potential into many real-world problems, e.g., (Both et al., 2021; Diefenbach et al., 2021; Mutabazi et al., 2021). These systems can be divided into two main paradigms (Jurafsky and Martin, 2018): IR-based that works over unstructured data, closely related to Machine Reading Comprehension and Retriever-Reader architecture) and Knowledge-Based (KBQA) which works over structured data, such as relational tables, specific data APIs, knowledge graphs (KGs). In this regard, Question Answering over Knowledge Graphs (KGQA) is of particular interest to this work.

Many different benchmarking datasets are used for evaluating KGQA systems. These datasets differ in the underlying knowledge graph (e.g., DBpedia (Auer et al., 2007a) or Wikidata (Erxleben et al., 2014)), size order of magnitude (Fu et al., 2020), questions complexity (Saleem et al., 2017), multilingual support (Chandra et al., 2021), and many more dimensions. In the KGQA research community, several datasets have become a de facto standard for evaluation of such systems, such as the QALD (Usbeck, Gusmita, et al., 2018b) and LCQuAD (Dubey et al., 2019a) benchmark dataset series. As more and more researchers introduce new evaluation results using these well-known datasets, it becomes more challenging to follow the up-to-date state-of-the-art in the KGQA field. The related research fields such as IR-based QA and Knowledge Graph research community already have their own well-established and maintained leaderboards of the best solutions (SQuAD[2] (Rajpurkar et al., 2016), OGB[3] (Hu et al., 2020)). However, it is not the case for KGQA.

This lack - in particular of curated leaderboards - leads to a missing global view over the research field. In turn, this could inject mistrust into result tables within publications when they are incomplete or lack a comparison to certain systems, as often required by reviewers. In particular, the latest and most-used datasets in the Semantic Web community, LCQuAD and QALD, miss providing central points of trust such as leaderboards. In this paper, we analyze the publications of KGQA evaluations of the last decade. We evaluated 100 papers and 98 systems on 4 datasets focusing on the LC-QuAD and QALD series. Our results show that evaluation numbers are often consistent. Existing errors stem from minor differences in the data (e.g., gAnswer (Hu et al., 2018) on QALD-9 (Usbeck, Gusmita, et al., 2018b)) that seems to be rounding errors or inconclusive behavior. Finally, we

1. https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020 (September 28, 2020)

2. The Stanford Question Answering Dataset leaderboard at https://rajpurkar.github.io/SQuAD-explorer/

3. Open Graph Benchmark – is a collection of the benchmark datasets for machine learning over graphs: https://ogb.stanford.edu

discuss the consequences of our findings and will point to possible improvements for future evaluations.

Our contributions are as follows:

- We present the first, extensive evaluation analysis of the state of the research in KGQA.
- We provide a new central and open leaderboard for any KGQA benchmark dataset as a focal point for the community -

  https://github.com/KGQA/leaderboard. With pull requests, the community can easily enhance the leaderboard.
- We provide an up-to-date overview of all available demos or Web services for KGQA at the point of publication.

These contributions should help the scientific community to foster replication and cross-evaluation in the future.

In the following, we analyze related studies and approaches in Section 2. Afterward, we introduce the analyzed datasets and systems in Section 3. In Section 4, we describe our extensive state-of-the-art data and delve into its analysis. Next, we discuss possible interpretations and paths forward and end with a summary and outlook in Section 6.

## 3.2 Related Work

There are multiple approaches to tracking the progress of any research field. In machine learning and NLP, these approaches can be subdivided into *benchmarking frameworks* and *manual or semi-automatic reporting platforms*.

Today, benchmarking frameworks need to limit their scope to a subset of tasks to cover the necessary metrics and experiment types out-of-the-box. A general benchmarking framework, which works without writing code, does not exist. For KGQA, different *benchmarking* frameworks have been proposed. For example, GERBIL QA (Usbeck et al., 2019), can benchmark KGQA systems via their Web APIs in a FAIR way (Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, Blomberg, Boiten, da Silva Santos, et al., 2016). It also has an integrated leaderboard[4] which displays a summary of all experiments run via the platform. At the same time, this is the biggest downside – only experiments run via the platform are integrated. Thus, a realistic view depends on the adoption of the platform. This adoption seems to lack due to missing developer resources, which continuously update available systems and datasets. A different direction is followed by systems like https://github.com/AKSW/irbench or QALDGen (Singh et al., 2019), which provide command-line tools for benchmarking any KGQA system. However, the offline nature of these tools leads to offline results, i.e., the results might be used in papers but do not contribute to a trustworthy overview of the field of research.

Recently, *reporting platforms* gained popularity. They allow quick access to results, but either they are curated manually via a community or semi-automatically updated. https://nlpprogress.com/ is a famous community website launched by

4. http://gerbil-qa.aksw.org/gerbil/overview

Sebastian Ruder. Regarding KGQA, the website's most recent information is 3 years old, possibly displaying the disinterest of the NLP community in semantic tasks. https://paperswithcode.com/ is another reporting platform run by Facebook AI research allowing to openly edit papers, code, datasets, methods, and evaluation tables. While this is of tremendous help for reproducibility, its results for KGQA are sparse. There is only one result for LC-QuAD 2 and QALD-9, and both are for relation extraction rather than Question Answering.

A promising semi-automatic approach is the Open Research Knowledge Graph (ORKG) (Auer et al., 2020). It allows the community to persistently annotate papers via smart tools with meta and evaluation data, e.g.,

https://www.orkg.org/orkg/paper/R6386/R6393 for QALD-6 data. However, the current adoption in the community does not go beyond prototypes provided by the ORKG team. A change might come with the European Open Science Cloud (EOSC) and the Nationale Forschungsdateninfrastruktur für Data Science und Künstliche Intelligence (German National Data Infrastructure for Data Science and AI). Those publicly funded initiatives strive to foster ecosystems like ORKG in the long term.

Finally, surveys can be viewed as reporting platforms. Different surveys have been published in the past decade focusing on a variety of topics such as challenges in general KGQA (Höffner et al., 2017), challenges in complex KGQA (Fu et al., 2020), core techniques of KGQA (Diefenbach, López, et al., 2018) or neural network-based KGQA systems (Chakraborty et al., 2021). However, these are automatically outdated when published or focus only on a narrow subtopic.

Thus, there is the need for a central, dense, and open reporting platform focusing on KGQA, which provides trustworthy insights.

## 3.3 Benchmark Datasets and Systems

We surveyed 14 DBpedia-based KGQA benchmark datasets that were published in the last decade (Section 3.3.1). In this paper, we consider 4 KGQA datasets for an in-depth analysis. Requirements for selecting a dataset include usage for the evaluation of different systems, availability in English, relying on DBpedia (primarily) or Wikidata (knowledge bases, which are still maintained), and cited above 5 times. Our goal was to make sure that the chosen QA datasets are: up-to-date, close to a real-world setting, can be manually evaluated, and are vastly studied. Note, we use benchmark datasets and datasets synonymous.

We took 98 QA systems into the consideration. They are collected manually from articles that include evaluation results on the considered benchmark datasets. The article search was conducted in two ways. First, we retrieved articles using a keyword search on Google Scholar[5]. Specifically, the selection criteria were: published after 2019, and titles satisfy: `['question answering' AND ('semantic web' OR 'data web' OR 'web of data')]`. The second method is to extract all articles which cite the benchmark dataset from Google Scholar either as direct citation or as URL to the location of the dataset. We removed duplicates and manually extracted the QA systems evaluated or referred to in the articles. This resulted in 100 analyzed papers. Note, some systems are evaluated on a subset of

5. http://scholar.google.com

the dataset or a dataset where the benchmark dataset is just a subset. We indicated such a difference in the leaderboard accordingly.

### 3.3.1 KGQA Datasets

The first dataset is QALD which is multilingual dataset challenge series. In QALD-8, there were 219 training question-answer pairs and 42 test data points respectively. It was the first edition to use GERBIL QA as a benchmarking platform (Usbeck et al., 2019). The newest instance – QALD-9 (Usbeck, Gusmita, et al., 2018b) – contains 558 questions incorporating information from the DBpedia knowledge base[6] where for each question the following is given: a textual representation in multiple languages, the corresponding SPARQL query (over DBpedia), the answer entity URI, and the answer type. The QALD series has a growing number of questions per edition and thus grows continuously in its expressiveness. The dataset has become a staple for many research studies in QA (e.g., (Diefenbach, López, et al., 2018; Höffner et al., 2017)).

The second and third dataset is LCQuAD. LCQuAD (version 1) (Trivedi et al., 2017) is a Question Answering dataset with 5000 pairs of questions and its corresponding SPARQL query. LCQuAD v2 (Dubey et al., 2019a) is the follow-up dataset with 30.000 question-answer pairs to better suit novel machine learning approaches. The SPARQL queries are intended to be executed on DBpedia. LCQuAD is widely used in the process of QA systems development (Dubey et al., 2018; Singh et al., 2018).

Other KGQA datasets are Free917 (Cai and Yates, 2013), WebQuestions (Berant et al., 2013), ComplexQuestions (Bao et al., 2016), SimplesQuestions (Bordes et al., 2015), GraphQuestions (Su et al., 2016), WebQuestionsSP (Yih et al., 2016a), 30MFactoidQA (Serban et al., 2016), ComplexWebQuestions (Talmor and Berant, 2018), PathQuestion (Zhou et al., 2018), MetaQA (Zhang et al., 2018), TempQuestions (Jia et al., 2018), TimeQuestions (Jia et al., 2021), CronQuestions (Saxena et al., 2021), FreebaseQA (Jiang et al., 2019), Compositional Freebase Questions (CFQ) (Keysers et al., 2019), Compositional Wikidata Questions (CWQ) (Cui et al., 2021), RuBQ (Korablinov and Braslavski, 2020a; Rybin et al., 2021a), QALD-9-plus (Perevalov, Diefenbach, et al., 2022a), GrailQA (Gu et al., 2021), Event-QA (Souza Costa et al., 2020), SimpleDBpediaQA (Azmy et al., 2018), CLC-QuAD (Zou, Yang, Zhang, Xu, Pan, Jiang, Qin, Wang, He, Huang, et al., 2021), KQA Pro (Shi et al., 2020), SimpleQuestionsWikidata (Diefenbach, Tanon, et al., 2017a), DBNQA (Yin et al., 2019), etc.

These datasets do not fulfill our current criteria and thus are not part of the initial version of the KGQA leaderboard. However, we encourage the community to help us update the leaderboard also for these datasets to prevent a replication crisis before it starts.

### 3.3.2 QA systems

While there are decentral collections of KGQA systems and there are available as code or Web service, e.g.,

6. https://www.dbpedia.org/

https://github.com/semantic-systems/NLIWOD/tree/master/qa.systems, there is no up-to-date and systematically curated collection as of now. Our analysis shows that 24 provide a URL to a repository and 16 even to an online demo or Web API. However, after inspection, only 8 demos or Web APIs are still functional. This is the first hint toward an upcoming replication crisis. For a full list of systems, their descriptions, and pointers to their web services and demo, see https://github.com/KGQA/leaderboard/blob/gh-pages/systems.md#Systems

## 3.4 Dataset Analyses

We evaluated 100 papers and 98 systems focusing on 4 datasets, namely LC-QuAD version 1 and version 2 as well as the QALD-8 and 9 versions (all datasets released in 2017 or later). Figure 3.1 comprehensively summarizes the considered results of the leaderboard.

Based on the results, it became clear that *the evaluation values across the publications are often consistent.* The results contain multiple values for some of the system-dataset combinations (e.g., WDAqua-core0 over LCQuAD 1.0), reported by different publications. Figure 3.2 demonstrates the evaluation values given a particular benchmark dataset grouped by the KGQA systems. For system-dataset combinations with multiple values, we calculated the standard deviation (std.). The std. values for such systems as QAKiS, TeBaQA, Elon, QASystem, gAnswer, and QAmp are not higher than 1%. This non-null std. is probably caused by the rounding errors. The only outliers in the evaluation values were observed given the WDAqua-core systems. For example, the paper (Zheng and Zhang, 2019) reports Fscore of 38.7% for WDAqua-core0 over QALD-8, taking the results from the original publication of WDAqua-core0. Another paper (Orogat et al., 2021) reports Fscore of only 33.0% for the same system-dataset combination. The authors (Orogat et al., 2021) calculated this result. The std. of both WDAqua-core versions reaches 9% on LCQuAD 1.0 dataset and 3% on QALD-8. Note, the high std. values are not dependent on the datasets. Hence, the papers reporting significantly different results regarding WDAqua-core require further investigation. One of the assumptions is that WDAqua-core provides a publicly accessible demonstrator and API[7] which enables researchers to re-run the evaluation. This fact naturally implies possible differences in the evaluation results. However, there is no such systematic tendency for the other results as probably the majority of them were not reproduced but cited from the original publication. Despite the consistency of the results, the Fscore values of the systems have a wide variance range given a particular dataset (Figure 3.3).

Surprisingly, the number of papers from ArXiv (pre-prints) in our leaderboard appears to be higher than the number of peer-reviewed papers (54% vs 46%). It was observed that the peer-reviewed papers report significantly higher results w.r.t. Fscore which is 30.2% for preprints and 39.5% for peer-reviewed papers. The logical reason for this is that the peer-reviewed papers typically report state-of-the-art results, while preprints might contain preliminary work.

Given the considered results, it was observed that the authors of 72% papers did not include all the evaluation results from other publications in their comparison

7. https://qanswer-frontend.univ-st-etienne.fr

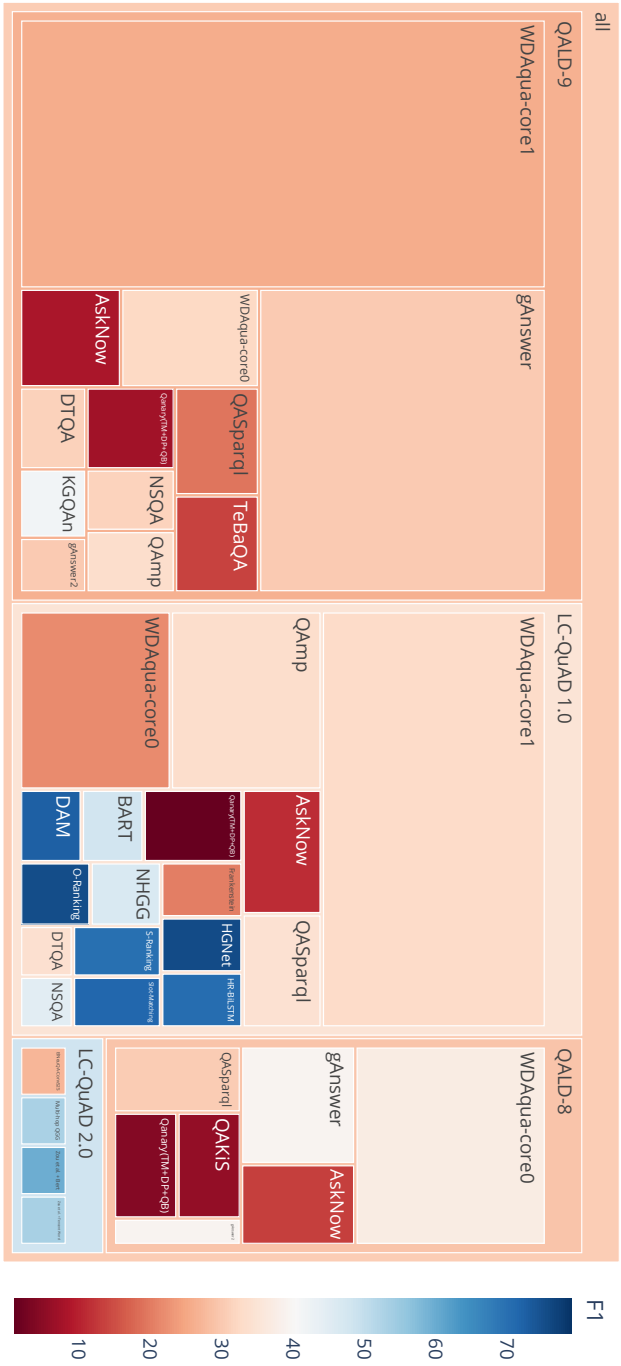



**Figure 3.1:** Treemap chart based on the collected results grouped by considered datasets (QALD-8, QALD-9, LCQuAD 1.0, LCQuAD 2.0). The KGQA systems are located within the dataset rectangles. The size of the rectangles is proportional to the number of mentions of a particular system in the whole leaderboard. The color of the rectangles denotes the average Fscore of the corresponding systems. Only systems with more than 2 mentions are included.

that were already available at a particular point in time. To find out this number, the set of systems from a publication reporting the values on particular datasets was compared to the set of systems released a year ago or earlier. For example,

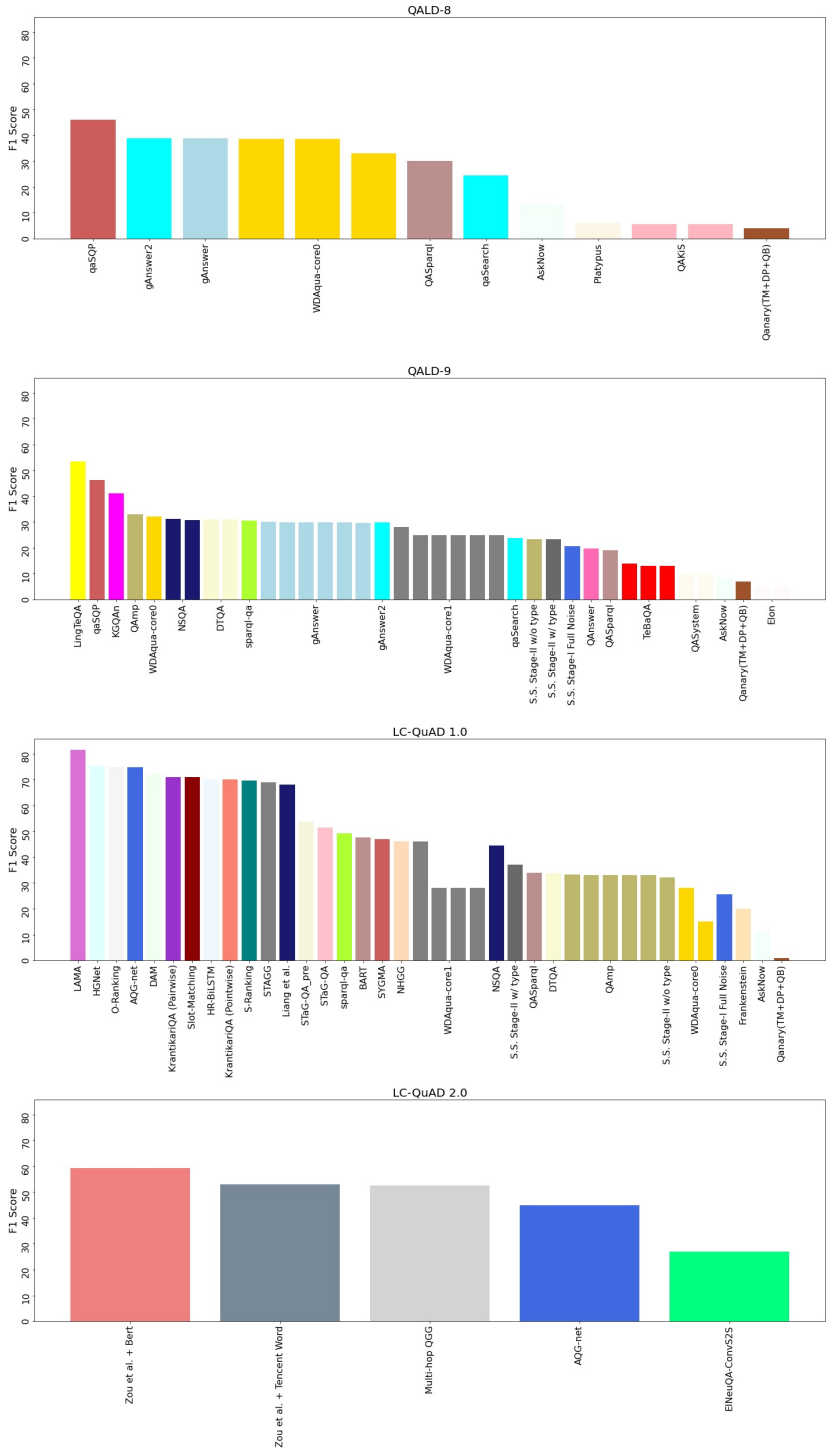



**Figure 3.2**: The chart demonstrates evaluation values (Fscore) grouped by KGQA systems (same color) given a dataset. Each bar corresponds to a particular publication.

the publication (Orogat et al., 2021) released in 2021 does not consider the results of the QAmp system (Vakulenko et al., 2019) that was published in 2019.

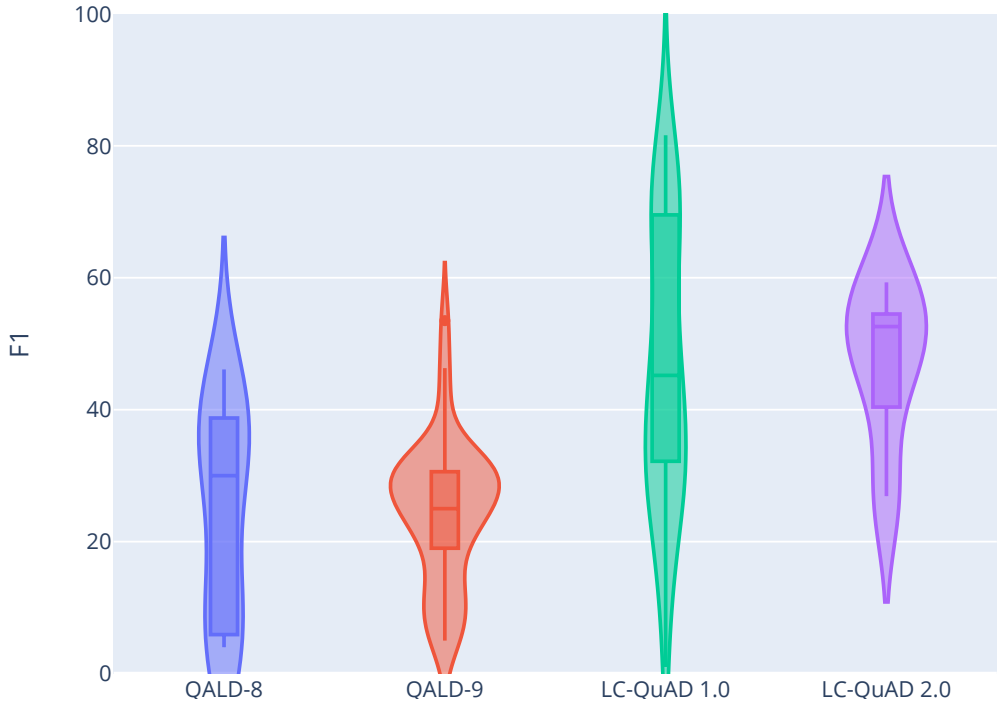



**Figure 3.3:** The figure demonstrates the distribution of the Fscore values and their statistics from different publications given a dataset.

## 3.5 Discussion

The trustworthiness of scientific results strongly depends on their comparability and replicability. In the field of KGQA, one could assume that the existence of a large and rising number QA datasets ensures comparability. Indeed, our analysis shows that the reported evaluations are *overwhelmingly coherent*. However, we observed several issues: first, the main reason why most numbers are identical is that people refer to results given in an original paper and its evaluation section. We could not find evidence that researchers actively tried to replicate results. A reason could be that only, 16 percent of the systems are available as source code (or web service/demo). However, even in the existence of an online demo, e.g., (Diefenbach et al., 2020; Diefenbach, Singh, et al., 2017a), the current state of the KGQA system seems not to be re-evaluated.

Second, our analysis indicates that researchers might have overlooked (best case) or omitted (worst case) relevant results that speak against their claims. For example, in (Wu et al., 2021) there are similar earlier works (Maheshwari et al., 2019; To and Reformat, 2020) which evaluated the same datasets and provided similar or even better results. However, we are well aware that researchers struggle with establishing an up-to-date overview of current research due to the time-consuming nature of the process without a central overview of KGQA systems.

Third, we see a strong need for improved evaluation methods. This demand can be covered by online evaluation methods, e.g., using platforms like Gerbil (Usbeck et al., 2019)). However, we also observed a decreasing amount of working online

demos suggesting that a new form of a platform where models as such can be uploaded[8] could be a future direction.

Fourth, while developing new platforms and systems, we should also consider the rising critique on leaderboards regarding their utility for the NLP community at large (Ethayarajh and Jurafsky, 2020). Thus, we concur that evaluation protocols need to be published to foster transparency on leaderboards.

Finally, the lack of open-source implementations could be a starting point for a replication crisis. While there is no replication crisis in the field of KGQA as of now, the community needs to leverage novel initiatives such as the European Open Science Cloud[9] or the National Research Data Infrastructure for Data Science and AI[10]. Otherwise, models and source code might be lost or results will become incomparable in the long term.

## 3.6 Summary and Future Work

In this paper, we presented a novel community resource to track advances in the field of KGQA research. We foresee the need to maintain a KGQA focused platform as long as approaches such as ORKG (Auer et al., 2020) are not widely used or developed far enough. Of course, we could have just added our findings to reporting platforms. However, we believe, that this publication provides a more valuable base for discussions and reaches a wider audience than a silent upload. Additionally, since the QALD-9 evaluation campaign has passed for more than 3 years now, we intend to establish a central leaderboard to keep people on the same page.

In the future, we are looking into automatic ways to synchronize various reporting platforms with the KGQA leaderboard. We plan to extend the evaluation of QA systems, $^{st}$replicable evaluations, and data collections are possible. Additionally, improved metrics (e.g., (Orogat and El-Roby, 2021; Siciliani et al., 2021)) should be evaluated over models, source code, or via platforms to allow in-depth analyses of the capabilities of QA systems.

We are aware of research on other KGQA datasets grounded in Wikidata, Freebase, WikiMovies, and EventKG and want to encourage the community to update the KGQA leaderboard with the corresponding numbers.

## 3.7 Acknowledgements

The authors also acknowledge the financial support by the Federal Ministry for Economic Affairs and Energy of Germany in the project CoyPu (project number 01MK21007G).

8. For example, https://project-hobbit.eu/outcomes/hobbit-platform/.
9. https://eosc-portal.eu/
10. https://www.nfdi4datascience.de/

## 3.8 KGQA Leaderboard

To ensure the replication of KGQA systems and the trustworthiness of their evaluation results, we provide a leaderboard. The leaderboard is available at https://kgqa.github.io/leaderboard/. It can be used to compare the capabilities of these KGQA systems over the latest and commonly used KGQA benchmark datasets by tracking the progress. It includes the datasets, links, papers, and SOTA results.

At the time of writing, the leaderboard includes a total of 34 KGQA datasets across 5 knowledge graphs (i.e., DBpedia, Wikidata, Freebase, WikiMovies, and EventKG). As shown in Fig. 3.4, these KGQA datasets are separated by the used target KGs. Fig. 3.5 shows an example of LCQuAD V1.0 Leaderboard. We will continuously add newly released datasets and their SOTA results, and invite other researchers to make their contributions by adding new results based on these KGQA dataset overviews.

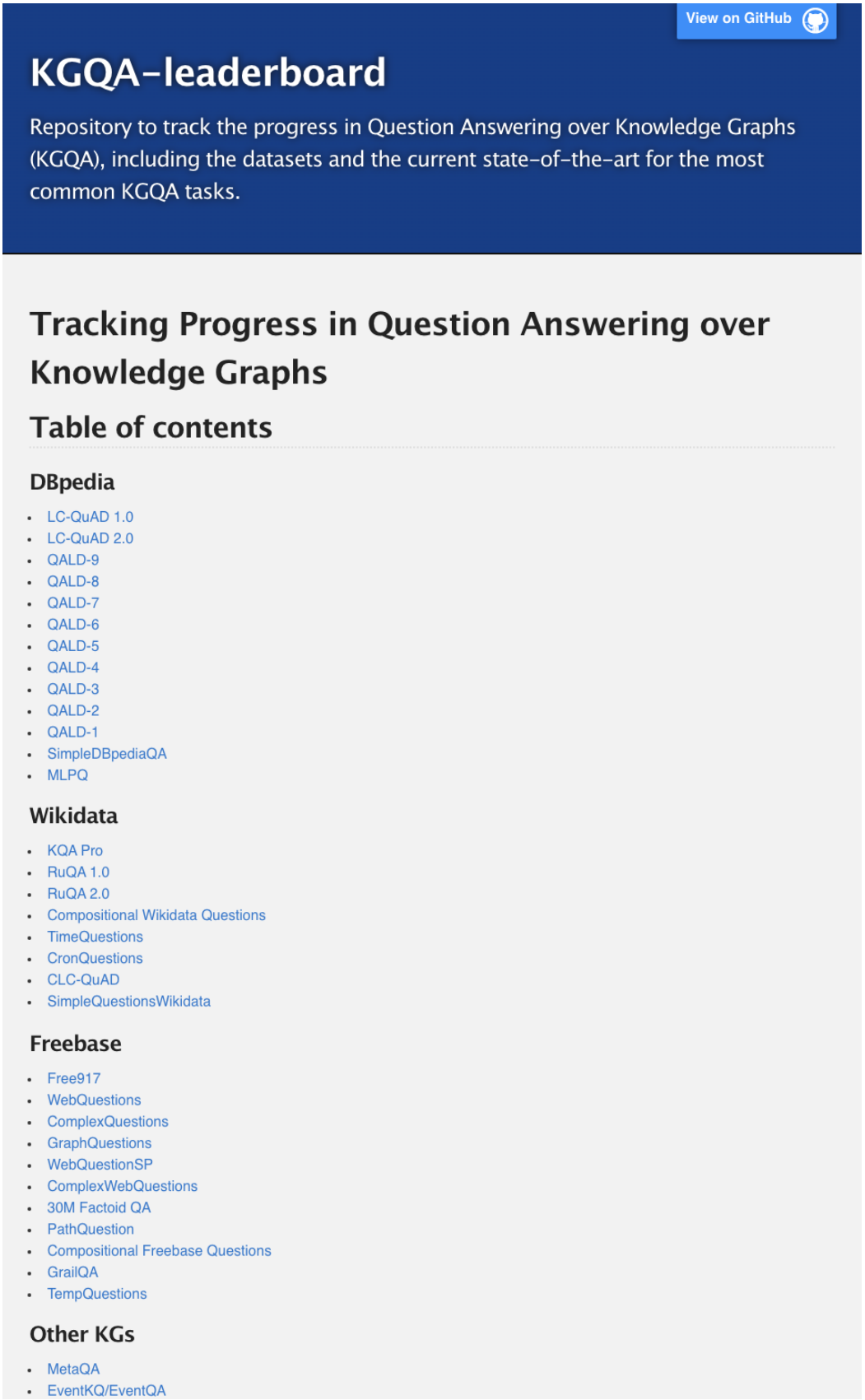



**Figure 3.4**: Interface of the KGQA leaderboard.

| Model / System | Year | Precision | Recall | F1 | Language | Reported by |
|---|---|---|---|---|---|---|
| mBERT | 2021 | 73 | - | 85.50 | EN | Zhou Y. et al |
| Stage-I No Noise | 2021 | 83.11 | 83.04 | 83.08 | EN | Purkayastha et al. |
| mBERT | 2021 | - | - | 82.40 | DE | Zhou Y. et al |
| LAMA | 2019 | - | - | 81.60 | EN | Radoev et. al. |
| mBERT | 2021 | - | - | 80.90 | NL | Zhou Y. et al |
| mBERT | 2021 | - | - | 76.10 | ES | Zhou Y. et al |
| HGNet | 2021 | 75.82 | 75.22 | 75.10 | EN | Chen et al. |
| O-Ranking | 2021 | 75.54 | 74.95 | 74.81 | EN | Chen et al. |
| AQG-net | 2021 | - | - | 74.80 | EN | Chen et al. |
| mBERT | 2021 | - | - | 74.50 | RU | Zhou Y. et al |
| mBERT | 2021 | - | - | 74 | PT | Zhou Y. et al |
| mBERT | 2021 | - | - | 73.20 | FR | Zhou Y. et al |
| mBERT | 2021 | - | - | 72.60 | RO | Zhou Y. et al |
| mBERT | 2021 | - | - | 72.30 | IT | Zhou Y. et al |

**Figure 3.5**: An example of LCQuAD V1.0 Leaderboard.

# 4

# QALD-10 — The 10th Challenge on Question Answering over Linked Data-Shifting from DBpedia to Wikidata as a KG for KGQA

## Bibliographic Information

Ricardo Usbeck*, Xi Yan*, Aleksandr Perevalov*, Longquan Jiang*, Julius Schulz*, Angelie Kraft, Cedric Möller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, et al. 2024. Qald-10–the 10th challenge on question answering over linked data: Shifting from dbpedia to wikidata as a kg for kgqa. * Shared first authorship, *Semantic Web* 15 (6): 2193–2207. (Cited on pages 10 sq.).

## Abstract

Knowledge Graph Question Answering (KGQA) has gained attention from both industry and academia over the past decade. Researchers proposed a substantial amount of benchmarking datasets with different properties, pushing the development in this field forward. Many of these benchmarks depend on Freebase, DBpedia, or Wikidata. However, KGQA benchmarks that depend on Freebase and DBpedia are gradually less studied and used, because Freebase is defunct and DBpedia lacks the structural validity of Wikidata. Therefore, research is gravitating toward Wikidata-based benchmarks. That is, new KGQA benchmarks are created on the basis of Wikidata and existing ones are migrated. We present a new, multilingual, complex KGQA benchmarking dataset as the 10th part of the Question Answering over Linked Data (QALD) benchmark series. This corpus formerly depended on DBpedia. Since QALD serves as a base for many machine-generated benchmarks, we increased the size and adjusted the benchmark to Wikidata and its ranking mechanism of properties. These measures foster novel

KGQA developments by more demanding benchmarks. Creating a benchmark from scratch or migrating it from DBpedia to Wikidata is non-trivial due to the complexity of the Wikidata knowledge graph, mapping issues between different languages, and the ranking mechanism of properties using qualifiers. We present our creation strategy and the challenges we faced that will assist other researchers in their future work. Our case study, in the form of a conference challenge, is accompanied by an in-depth analysis of the created benchmark.

## 4.1 Introduction

Research on Knowledge Graph Question Answering (KGQA) aims to facilitate an interaction paradigm that allows users to access vast amounts of knowledge stored in a graph model using natural language questions. KGQA systems are either designed as complex pipelines of multiple downstream task components (Both et al., 2016; Diefenbach, Singh, et al., 2018) or end-to-end solutions (primarily based on deep neural networks) (Wei et al., 2019) hidden behind an intuitive and easy-to-use interface. Developing high-performance KGQA systems has become more challenging as the data available on the Semantic Web, respectively in the Linked Open Data cloud, has proliferated and diversified. Newer systems must handle more volume and variety of knowledge. Finally, improved multilingual capabilities are urgently needed to increase the accessibility of KGQA systems to users around the world (Perevalov, Ngomo, et al., 2022). In the face of these requirements, we introduce QALD-10 as the newest successor of the *Question Answering over Linked Data* (QALD) benchmark series to facilitate the standardized evaluation of KGQA approaches.

### 4.1.1 The Rise of Wikidata in KGQA

Among the general-domain KGs (knowledge graph) like Freebase (Bollacker et al., 2008b), DBpedia (Lehmann et al., 2014), and Wikidata (Erxleben et al., 2014), the latter has become a focus of interest in the community. While Freebase was discontinued, DBpedia is still active and updated on a monthly basis. It contains information that is automatically extracted from Wikipedia infoboxes, causing an overlap. However, Wikidata is community-driven and continuously updated through user input. Moreover, the qualifier model[1] of Wikidata allows more specific annotations of relations, which in turn allow for more complex questions (i.e., more than a single triple pattern). Hence, we expect that new KGQA benchmarks will utilize Wikidata and existing benchmarks will migrate away from Freebase and DBpedia to Wikidata. This becomes evident in the distribution of benchmarks represented in the curated KGQA leaderboard (Perevalov, Yan, et al., 2022) and is supported by a few publications in which KGQA benchmarks have already been moved to Wikidata (Diefenbach, Tanon, et al., 2017b; Tanon et al., 2016; Zou, Yang, Zhang, Xu, Pan, Jiang, Qin, Wang, He, Huang, and Zhao, 2021). One approach is to map the Freebase topics to Wikidata items using automatically generated mappings, for subjects as well as objects of triples (Diefenbach, Tanon, et al., 2017b; Tanon et al., 2016). For properties, handmade mappings are used.

1. https://www.wikidata.org/wiki/Help:Qualifiers

**Table 4.1**: Infobox for QALD-10.

| | |
|---|---|
| Name | QALD-10 |
| URL | https://github.com/KGQA/QALD-10 |
| Version date and number | 1.0/May 29th, 2022 |
| Licensing and availability | MIT license, open available |
| Topic coverage | General domain |
| Source for the data | Real life questions |
| Purpose and method of creation and maintenance | Academic purpose, manual creation and maintenance |
| Reported usage | Academic purpose |
| Metrics | Precision, Recall, Macro F1 QALD |
| Use of established vocabularies | RDF, purl, geosparql, wikibase, Wikidata, XMLSchema |
| Language expressivity | English, German, Chinese, and Russian |
| Growth | Static |

It is worth mentioning, that due to the structural differences between Freebase and Wikidata, some of the gold standard SPARQL (SPARQL Protocol and RDF Query Language) queries cannot be transformed, and respective questions become unanswerable. The original SPARQL queries from the CFQ (Keysers et al., 2020) benchmark over Freebase were mapped to Wikidata using a multi-step approach including property mapping and entity substitution.

### 4.1.2 Multilinguality in KGQA

Several works have contributed to the extension of the multilingual coverage of KGQA benchmarks. The authors of (Zou, Yang, Zhang, Xu, Pan, Jiang, Qin, Wang, He, Huang, and Zhao, 2021) were the first who manually translated LC-QuAD 2.0 (Dubey et al., 2019b) (an English KGQA benchmarking dataset on DBpedia) to Chinese. However, languages besides English and Chinese are not covered and the work does not provide a deeper analysis of the issues with the SPARQL query generation process faced when working with Wikidata. The RuBQ benchmark series (Korablinov and Braslavski, 2020b; Rybin et al., 2021b) which was initially based on questions from Russian quizzes (totaling 2,910 questions) has also been translated to English via machine translation. The SPARQL queries over Wikidata were generated automatically and manually validated by the authors. The CWQ (Cui et al., 2022) benchmark provides questions in Hebrew, Kannada,

Chinese, and English, with the non-English questions translated by machine translation with manual adjustments. The QALD-9-plus benchmark (Perevalov, Diefenbach, et al., 2022b) introduced improvements and an extension of the multilingual translations in its previous version—QALD-9 (Usbeck, Gusmita, et al., 2018a)—by involving crowd-workers with native-level language skills for high-quality translations from English to their native languages as well as validation. In addition, the authors manually transformed gold standard queries from DBpedia to Wikidata.

### 4.1.3 Introducing QALD-10

In this paper, we present the latest version of the QALD benchmark series—QALD-10—a novel Wikidata-based benchmarking dataset. It was piloted as a test set for the 10th QALD challenge within the 7th Workshop on Natural Language Interfaces for the Web of Data (NLIWoD) (Xi Yan and Usbeck, 2022). This challenge uses QALD-9-plus as training set and the QALD-10 benchmark dataset as test set. QALD-10 is publicly available in our GitHub repository.[2] The infobox in Table 4.1 contains more details on the benchmark. We also provide a Wikidata dump and a long-term maintained SPARQL endpoint[3] for this benchmark to foster replicable and reproducible research. In summary, we make the following contributions:

- A new complex, multilingual KGQA benchmark over Wikidata—QALD-10—and a detailed description of its creation process;
- An overview of the KGQA systems evaluated on QALD-10 and analysis of the corresponding results;
- A concise benchmark analysis in terms of query complexity;
- An overview and challenge analysis for the query creation process on Wikidata.

## 4.2 QALD-10 Challenge Description and Benchmark Introduction

The QALD-10 benchmarking dataset is a part of the QALD challenge series, which has a long history of publishing KGQA benchmarks. The benchmark was released as part of the 10th QALD challenge within the 7th Workshop on Natural Language Interfaces for the Web of Data at European Semantic Web Conference (ESWC) 2022 (Xi Yan and Usbeck, 2022).[4] While looking at past benchmarks (Cimiano et al., 2013; Lopez et al., 2013; Perevalov, Diefenbach, et al., 2022b; Unger et al., 2012; Unger et al., 2014b, 2015; Unger et al., 2016; Usbeck, Gusmita, et al., 2018a; Usbeck, Ngomo, et al., 2018; Usbeck et al., 2017), we identified several challenges.

2. Benchmark repository: https://github.com/KGQA/QALD-10. Note, that the QALD-10 is relatively stable in terms of opened community issues over time, compare to https://github.com/ag-sc/QALD/issues which hosts previous versions of this challenge.

3. https://skynet.coypu.org/wikidata/

4. https://www.nliwod.org/challenge

First, the **poor translation quality** for languages other than English (Perevalov, Diefenbach, et al., 2022b). For instance, the question Which subsidiary of TUI Travel serves both Glasgow and Dublin? translates to Italian as Quale società sussidiaria di TUI Travel serve sia Dortmund che Dublino? (QALD-9 dataset). It is evident, despite not being a native Italian speaker, that two different cities are used in the original question (Glasgow) and its translated version (Dortmund). The corresponding SPARQL query uses Glasgow, consistent with the original English question. Second, the **low complexity of the gold standard SPARQL queries** since most of the questions are resulting in SPARQL queries with fewer significant patterns which are essential to test the robustness of KGQA systems. Finally, the **weak replicability** of the KGQA experiments caused by divergence between the SPARQL query results and constantly updating versions of the used knowledge graphs (e.g., DBpedia and Wikidata). For instance, the same systems show big difference when evaluated at different time points, making it harder to compare the capacity of the KGQA systems. With QALD-10 benchmark dataset, we remedy all the aforementioned flaws.

All the data for the challenge can be found in our project repository.[5] The QALD-10 uses the well-established QALD-JSON format[6] (Usbeck, Röder, Hoffmann, et al., 2018), which is adopted by other KGQA benchmarks (Perevalov, Diefenbach, et al., 2022b; Siciliani et al., 2022; Usbeck, Gusmita, et al., 2018a; Usbeck, Ngomo, et al., 2018; Usbeck et al., 2017). The overall multilingual KGQA challenge contained 806 human-curated questions, including for 412 training and for 394 test. In the following, we explain the creation process in detail.

The *QALD-10 challenge training set* includes 412 questions and the corresponding queries, which are runnable against our stable SPARQL endpoint. The SPARQL query transformation from DBpedia to Wikidata was done manually by a group of computer scientists who were the authors of the QALD-9-plus paper. As some of the queries were not transformable, the total number of questions decreased from 558 to 507 between QALD-9 and QALD-9-plus. The number of KGQA pairs further decreased to 412 with the introduction of our stable Wikidata endpoint. These 412 question-query pairs form the QALD-10 challenge train set.

### 4.2.1 Collection of English Natural Language Questions for the QALD-10 challenge test set

The *QALD-10 challenge test set* (with 394 question pairs), in contrast, was created from scratch. In the first step, we collected 500 natural language questions in English from speakers with at least a C1-level language proficiency in accordance with the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). We collected equal amounts of questions from each participant to ensure that the questions are unbiased and express real-world information needs. Questions vary with respect to their complexity type, including questions with counts (e.g., *How many children does Eddie Murphy have?*), superlatives (e.g., *Which museum in New York has the most visitors?*), comparatives (e.g., *Is Lake*

5. https://github.com/KGQA/QALD-10/tree/main/data
6. https://github.com/AKSW/gerbil/wiki/Question-Answering

*Baikal bigger than the Great Bear Lake?*), and temporal aggregators (e.g., *How many companies were founded in the same year as Google?*).

### 4.2.2 Multilingual Translations

To tackle the *first challenge*, the translations from English to Chinese, German, and Russian were created by crowd-workers in two steps: (1) each English question was translated into the target language by two native speakers, (2) the translations from the previous step were validated by another native speaker. To reduce ambiguity, the named entities in the questions were manually annotated with their Wikidata URIs (Uniform Resource Identifier) before translation. The crowd-workers were asked to follow the Wikidata label of a particular entity in their native language during the translation process. Note that 12 of the Wikidata items did not have Chinese versions so the respective questions could not be labeled accordingly. For instance, the entity T̈he Vanishing Half" (Wikidata ID: `wd:Q98476957`) in the question *In which year was the author of T̈he Vanishing Half* born?" did not have a Chinese entry.

### 4.2.3 From Natural Language Question to SPARQL Query

To tackle the **second challenge**, the final set of questions was manually transformed into complex SPARQL queries over Wikidata by a group of computer scientists with complementary skills and knowledge. The incompleteness of Wikidata regarding some of the multilingual labels and the lack of an ontology caused challenges in the SPARQL query generation process. These are listed and discussed in detail in Section 4.5. The gold standard answers to the questions were retrieved by querying over our own Wikidata endpoint which is also available online.

### 4.2.4 Stable SPARQL Endpoint

Due to the constant updates of KGs like Wikidata, outdated SPARQL queries or changed answers can commonly cause problems for KGQA benchmarks. KG updates usually concern (1) structural changes, e.g., renaming of properties, or (2) alignment with changes in the real world, e.g., when a state has appointed a new president. According to our preliminary analysis on the LC-QuAD 2 benchmark (Dubey et al., 2019b), a large number of queries are no longer answerable on the current version of Wikidata. The original dump used to create the benchmark is no longer available online.[7] As a result, the authors (Banerjee et al., 2022b) set up an endpoint with a Wikidata dump dated 13 October 2021[8] and filtered the test set of 6046 questions down to 4211 questions for which the gold query produced a valid response. We decided to follow a similar methodology by setting up a stable Wikidata endpoint that was used to execute the SPARQL queries. Hence, we provide a long-term stable endpoint to ensure reproducibility to tackle the **third challenge**. This endpoint was also provided to the participants of the QALD-10 challenge and is archived via Zenodo (Usbeck et al., 2022).[9]

7. https://databus.dbpedia.org/dbpedia/wikidata/debug/2020.07.01
8. https://dumps.wikimedia.org/wikidatawiki/entities/
9. https://zenodo.org/record/7496690#.Y7QfI-zMK3I

Note, since the QALD-10 challenge training set was created before we set up this endpoint, some answers and queries had to be changed from the original release of QALD-9-plus. That is, the QALD-9-plus original data and the one used as QALD-10 challenge training data are different. Different versions are recorded as releases in the GitHub repository.[10]

## 4.3 QALD-10 Challenge Evaluation

To promote the FAIR principles (Findable, Accessible, Interoperable, and Reusable) (Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, Blomberg, Boiten, Bonino da Silva Santos, et al., 2016) with respect to our experimental results, we utilize GERBIL QA (Usbeck, Röder, Hoffmann, et al., 2018)[11], an open-source and publicly available online evaluation tool. We adopt the established metrics for KGQA evaluation, more specifically Precision, Recall, and the Macro F1 QALD measure.

### 4.3.1 Evaluation Metric

The F-measure is one of the most commonly used metrics to evaluate KGQA systems, according to an up-to-date leaderboard (Perevalov, Yan, et al., 2022). It is calculated based on Precision and Recall and, thus, indicates a system's capacity to retrieve the right answer in terms of quality and quantity (Manning, 2008). However, the KGQA evaluation has certain special cases due to empty answers (see also GERBIL QA (Usbeck, Röder, Hoffmann, et al., 2018)). Therefore, a modification was made to the standard F-measure to better indicate a system's performance on KGQA benchmarks. More specifically, when the golden answer is empty, Precision, Recall, and F-measure of this question pair receive a value of 1 only if an empty answer is returned by the system. Otherwise, it is counted as a mismatch and the metrics are set to 0. Reversely, if the system gives an empty answer to a question for which the golden answer is not empty, this will also be counted as a mismatch. Our analyses consider both micro- and macro-averaging strategies. These two averaging strategies are automatically inferred by the GERBIL system.

During the challenge, *only the Macro F1 QALD measure was used to rank the systems.* This resulted from community requests[12] and to achieve compatibility with older QALD challenges. This metric uses the previously mentioned additional semantic information with the following exception: If the golden answer set is not empty but the QA system responds with an empty answer set, it is assumed that the system determined that it cannot answer the question. Here we set the Precision to 1 and the Recall and F-measure to 0.

### 4.3.2 GERBIL QA Benchmarking Platform

The GERBIL system was originally created as a benchmarking system for named entity recognition and linking; it also follows the FAIR principles. It has been

10. https://github.com/KGQA/QALD_9_plus
11. https://gerbil-qa.aksw.org/gerbil
12. https://github.com/dice-group/gerbil/issues/211

widely used for evaluation and shared tasks for its fast processing speed and availability. Later, it was extended to support the evaluation of the KGQA systems. While adopting the GERBIL framework, the evaluation can simply be done by uploading the answers produced by a system via web interface or RESTful API.[13] Each experiment has a citable, time-stable, and accessible URI that is both human- and machine-readable. The uploaded file should follow the QALD-JSON format. The GERBIL system was set up with the QALD-10 benchmark and provides an easy-to-use configuration, which allows one to choose a language for the multilingual evaluation.[14] After choosing the language and uploading the file containing a system's predictions, the evaluation is done automatically.

To promote the reproducibility of the KGQA systems and open information access, we uploaded the test result of all systems to a curated leaderboard (Perevalov, Yan, et al., 2022).[15] The leaderboard includes the system descriptions, standardized evaluation scores, references, and other details. Therefore, it presents state-of-the-art scores for comparison purposes.

### 4.3.3 Participating Systems

After six registrations, five teams were able to join the final evaluation. Allowing file-based submissions rather than requiring web service-based submissions led to a higher number of submissions and fewer complaints by the participants compared to previous years. Thus, unfortunately, the goal of FAIR and replicable experiments is still unreached for KGQA. Among the participating systems, three systems papers were accepted to the workshop hosting the challenge.

*QAnswer* (Diefenbach, Singh, et al., 2017b) is a rule-based system using a combinatorial approach to generate SPARQL queries from natural language questions, leveraging the semantics encoded in the underlying knowledge graph. It can answer questions on both DBpedia and Wikidata supporting English, French, German, Italian, Russian, Spanish, Portuguese, Arabic, and Chinese. This system, which does not require training, is run as a baseline system for our challenge due to its capacity to tackle multilingual data.

*SPARQL-QA (Santana et al., 2022)* is a QA system that exploits Neural Machine Translation (NMT) and Named Entity Recognition (NER) modules to create SPARQL queries from natural language questions. The MNT module translates the question into a SPARQL query template in which the KG resources are replaced by placeholders, while the NER module identifies and classifies the entities present in the question. The outputs of two modules are merged to produce a new equivalent of the original SPARQL query to be executed over Wikidata, by replacing the placeholders in the template with the corresponding named entities. An uniform input format, namely QQT (Question, Query Template and Tagging), is introduced to ensure training two modules together and reduce the impact of out-of-vocabulary (OOV) words.

*Shivashankar et al. (Shivashankar et al., 2022)* presented a graph-to-graph transformation-based QA system using an Abstract Meaning Representation (AMR) graph to generate SPARQL queries, leveraging its ability to represent

13. https://pypi.org/project/gerbil-api-wrapper/
14. See https://gerbil-qa.aksw.org/gerbil/config-qald for the QALD configurations on GERBIL.
15. https://kgqa.github.io/leaderboard/wikidata/qald.html#qald-10

the semantics of a natural language. For a given question, its AMR graph is generated using a pre-trained multilingual AMR parser and simplified by removing unnecessary nodes and information. All possible executable SPARQL graphs are extracted from its simplified AMR graph. The system supports English and German questions.

*Baramiia et al. (Baramiia et al., 2022)* developed a QA system that first learns to predict representations of entities and properties which are close to correct queries and far from the others. It then finds the top-$k$ nearest to the correct query via Scalable Nearest Neighbors method with the dot product similarity measure. It natively supports English but can be extended to the multilingual case using Transformers trained in other languages.

*Suraj Singh and Dmitrii Gavrilev*[16] presented a multilingual KGQA model that first translates the questions from low-resourced languages into English using a pre-trained T5 (Raffel et al., 2020b) model and then searches for the answer using the DeepPavlov-based (Burtsev et al., 2018) ensemble. The pipeline consists of query template type classification, entity detection, entity linking, relation ranking, and query generation. It can support answering questions written in English, German, Chinese, and Russian. More detailed information regarding all of the systems is provided in the proceedings (Xi Yan and Usbeck, 2022).

### 4.3.4 Results

All systems were evaluated on the test set of QALD-10 challenge before the challenge. Participants had to upload a file their QA system generates, upload it to the GERBIL system. which would output an URL based on submission. Participants submit the GERBIL URL with their final results. Table 4.2 shows the systems and their performances with links to GERBIL QA. In 2022, *SPARQL-QA* (Santana et al., 2022) won the QALD-10 challenge.

**Table 4.2:** Evaluation results of the challenge participants' systems.

| Author (System) | Language | Macro F1 QALD | GERBIL QA Link |
|---|---|---|---|
| Borroto et al. (SPARQL-QA) (Santana et al., 2022) | EN | 0.595 | https://gerbil-qa.aksw.org/gerbil/experiment?id=202205200035 |
| Baseline (QAnswer) (Diefenbach, Singh, et al., 2017b) | EN | 0.578 | https://gerbil-qa.aksw.org/gerbil/experiment?id=202205120000 |
| Steinmetz et al. (Shivashankar et al., 2022) | EN | 0.491 | https://gerbil-qa.aksw.org/gerbil/experiment?id=202205260012 |
| Baramiia et al. (Baramiia et al., 2022) | EN | 0.428 | https://gerbil-qa.aksw.org/gerbil/experiment?id=202205210032 |
| Singh & Gavrilev (no publication) | EN | 0.195 | https://gerbil-qa.aksw.org/gerbil/experiment?id=202205210017 |

## 4.4 QALD-10 Test Set Analysis

KGQA benchmarks should be complex enough to properly stress the underlying KGQA systems and hence not biased towards a specific system. Previous studies (Saleem et al., 2017) have shown that various SPARQL features of the golden SPARQL queries, i.e, the corresponding SPARQL queries to QALD natural language questions, significantly affect the performance of the KGQA systems. These features include the number of triple patterns, the number of joins between triple patterns, the joint vertex degree, and various SPARQL modifiers such as LIMIT, ORDER BY, GROUP BY etc. At the same time, the KGQA research

16. Their paper is not published in the proceedings.

community defines "complex" questions based on the number of facts that a question is connected to. Specifically, complex questions contain multiple subjects, express compound relations and include numerical operations, according to (Lan et al., 2021). Those questions are difficult to answer since they require systems to cope with multi-hop reasoning, constrained relations, numerical operations, or a combination of the above. As a result, KGQA systems need to perform aggregation operations and choose from more entity and relation candidates while dealing with those questions than on questions with one-hop relations.

In this section, we compare the complexity of the QALD-10 benchmark with QALD-9-plus with respect to the aforementioned SPARQL features. The statistics of the number of questions in different QALD series datasets is shown in Table 4.3. We use the Linked SPARQL Queries (LSQ) (Stadler et al., 2022) framework to create the LSQ RDF (Resource Description Framework) datasets of both of the selected benchmarks for comparison. The LSQ framework converts the given SPARQL queries into RDF and attaches query features. The resulting RDF datasets can be used for the complexity analysis of SPARQL queries (Saleem et al., 2015). The resulting Python notebooks and the LSQ datasets can be found in our GitHub repository.[17] The complexity analysis of the selected benchmarks is presented in the next subsections.

**Table 4.3:** Statistics of the number of questions in different QALD series datasets

| Q10-WD Test | Q9-Plus-DB Train | Q9-Plus-DB Test | Q9-Plus-WD Train | Q9-Plus-WD Test |
|---|---|---|---|---|
| 394 | 408 | 150 | 371 | 136 |

### 4.4.1 Frequency of Modifiers

When answering complex questions, KGQA systems are required to generate corresponding complex formal (SPARQL) queries, e.g., with multiple hops or constraint filters, that can represent and allow to answer them correctly. One way to represent the complexity of queries can be represented by the frequencies of modifiers (e.g., `LIMIT` or `COUNT`) that occurred in the queries. Table 4.4 shows the frequencies of each modifier represented in the QALD-10 test set and the different subsets of the QALD-9-plus benchmark, respectively. For QALD-9-plus, the Wikidata-based datasets use less modifiers than their DBpedia counterparts for the same set of questions. However, the results suggest that the proposed benchmark is way more complex than QALD-9-plus in terms of various important modifiers. A more detailed analysis of the complexity is done in the following text. In terms of significant modifiers such as COUNT, FILTER, ASK, GROUP BY, OFFSET, and YEAR, our benchmark has a significantly higher number than others. As to the modifiers such as LIMIT, ORDER BY, UNION, HAVING and NOW, this benchmark is close to the frequencies in the other datasets.

17. https://github.com/KGQA/QALD-10/tree/main/notebooks

**Table 4.4**: Frequencies of each modifier in different QALD series. Note that frequencies of modifiers with the * character are computed using keyword matching from SPARQL queries, while the others use the LSQ framework.

| Modifier | Q10-WD Test | Q9-Plus-DB Train | Q9-Plus-DB Test | Q9-Plus-WD Train | Q9-Plus-WD Test |
|---|---|---|---|---|---|
| **COUNT*** | **126** | 57 | 33 | 32 | 18 |
| **LIMIT** | 17 | 39 | 11 | **43** | 12 |
| **ORDER BY** | 17 | 36 | 11 | **43** | 12 |
| **FILTER** | **74** | 31 | 17 | 31 | 13 |
| **ASK** | **60** | 37 | 4 | 36 | 3 |
| **UNION** | 5 | **29** | 17 | 10 | 6 |
| **OFFSET** | **3** | 1 | 0 | 2 | 0 |
| **GROUP BY** | **95** | 19 | 11 | 12 | 12 |
| **HAVING*** | 1 | **3** | 2 | 1 | 2 |
| **YEAR*** | **43** | 6 | 10 | 20 | 4 |
| **NOW*** | 1 | **3** | 2 | 1 | 1 |

### 4.4.2 Query Feature Distribution

To measure structural query complexity, we calculate the Mean and Standard Deviation (SD) values for the distributions of three query features respectively: number of triple patterns, number of join operators, as well as joint vertex degree (see the definitions in (Saleem et al., 2019)). These features are often considered as a measure of structural query complexity when designing new SPARQL benchmarks (Saleem et al., 2015; Saleem et al., 2017; Saleem et al., 2019). The corresponding results are presented in Table 4.5. Again, the Wikidata-based datasets of QALD-9-plus have a lower distribution mean than their DBpedia counterparts and, thus, a lower complexity in general. Compared to the QALD-9-plus test sets, the QALD-10 test has a higher variation for the number of triple patterns and the number of join operators, as well as the second largest SD for joint vertex degree. This can be interpreted as getting the correct answers for the QALD-10 benchmark might be more difficult due to a wider range of possible SPARQL queries as compared to QALD-9-Plus

### 4.4.3 Query Diversity Score

From the previous results, it is still difficult to establish the final complexity of the complete benchmark. To this end, we calculate the diversity score (DS) of the complete benchmark $B$, formally defined as follows (Saleem et al., 2019).

$$DS = \frac{1}{k} \sum_{i=1}^{k} \frac{\sigma_i(B)}{\mu_i(B)} \tag{4.1}$$

where, $\mu$ and $\sigma$ are the Mean and Standard Deviation of a given distribution with respect to the $i$-th feature, respectively. The $k$ is the total number of query features analyzed in $B$. In this work, the query features chosen are the number of triple patterns, the number of joins, and the joint vertex degree. Table 4.6

**Table 4.5:** Structural complexity measured via the distribution of the number of triple patterns, the number of joins, and vertex degrees.

| Query Feature | | Q10-WD Test | Q9-Plus-DB Train | Q9-Plus-DB Test | Q9-Plus-WD Train | Q9-Plus-WD Test |
|---|---|---|---|---|---|---|
| Number of Triple Patterns | Mean | 1.605 | 1.728 | **1.993** | 1.685 | 1.640 |
| | SD | 1.199 | 0.944 | 1.167 | **1.215** | 0.998 |
| Number of Join Operators | Mean | 0.622 | 0.509 | **0.711** | 0.577 | 0.507 |
| | SD | **1.123** | 0.662 | 0.869 | 0.929 | 0.686 |
| Joint Vertex Degree | Mean | 0.889 | 0.941 | **1.089** | 0.953 | 0.929 |
| | SD | 1.133 | 1.113 | 1.116 | 1.148 | **1.176** |

shows the diversity scores (the higher the score the more complex the queries) of different QALD benchmarks. We observe that the QALD-10 test set has the highest diversity score ($DS = 1.275$) compared to the other benchmarks, hence it is a good starting point for template-based KGQA benchmark generation approaches aiming at diverse, complex, large-scale characteristics.

**Table 4.6:** Query diversity score of different QALD benchmarks.

| Q10-WD Test | Q9-Plus-DB Train | Q9-Plus-DB Test | Q9-Plus-WD Train | Q9-Plus-WD Test |
|---|---|---|---|---|
| 1.275 | 1.010 | 0.944 | 1.178 | 1.075 |

## 4.5 Challenging Translation of Natural Language Question to Wikidata SPARQL queries

During the creation of QALD-10 test SPARQL queries for given natural language questions, we identified several challenges. Therefore in this section, we formulate our challenges and solutions during the SPARQL generation process to aid further research in KGQA dataset creation as well as Wikidata schema research. Below, we systematically classify the problems into seven categories: (1) the ambiguity of the questions' intention, (2) incompleteness of Wikidata, (3) ambiguity of SPARQL queries, (4) limit on returned answers, (5) special vocabulary, (6) calculation limitation of SPARQL, and (7) endpoint version change. We discuss the cause of these issues and present our solutions.

### 4.5.1 Ambiguity of the Natural Language Question

The question *What is the biggest city in the world?* could be asking for the most populous city or the geographically largest city. This is an example for an ambiguous natural language question. We tried to circumvent this type of questions by specifically, asking crowd-workers to phrase their questions precisely.

After data collection, we chose the most reasonable interpretation based on real world experience, where necessary. Thus, in the example above, the question was changed to *What is the most populous city in the world?*. Some questions, however, remain vague and consequently correspond to multiple SPARQL queries. For instance, the question: *How many spouses do head of states have on average?* translates to:

where we used a *head of state*-property. However, using a *head of state*-class would also be feasible, leading to the SPARQL query below. There is no good way to make the question clearer except if one specifies the actual Wikidata elements, which is not realistic for real-world questions.

### 4.5.2 Incompleteness of Wikidata

Due to the incompleteness of Wikidata, some of the entity labels do not translate to all languages of interest (see Section 4.2.2). To avoid unanswerable questions, we supplemented the online Wikidata by manually adding a translation approved by a linguist. However, this update is not shown in our stable endpoint which has a fixed Wikidata dump. Consequently, the labels are still missing but the questions-query-answer tuples are inserted into QALD-10 benchmark dataset.

Another issue that can arise from Wikidata's incompleteness is that data necessary to answer specific SPARQL queries can be missing. In this case, the answer can not be retrieved by a correct SPARQL query since there are no suitable triples available directly. In our benchmark, questions of this category were deleted.

### 4.5.3 Ambiguity of SPARQL Queries in Wikidata due to Ranking of Properties

Ambiguities of SPARQL queries in Wikidata are connected to the ranking mechanism of Wikidata. This mechanism allows to annotate statements with preference information.[18] Such ranks decide how relevant different values of a statement are, which becomes a source of inaccuracy when there are multiple intuitively correct ways of writing a query. In some cases, only results with high ranking will be kept while the result discards the low-ranking but correct ones. Here, we show the differences between using `p/ps`[19] and `wdt`[20] in the queries:

- The `wdt` prefix for properties only returns values of properties with preferred rank, if one exists. If no ranking exists, it returns every property.
- The `p/ps` combination always returns all properties and their values, without respecting ranks.

This makes `wdt` the go-to choice to find the most recent value of well-maintained properties, like head of state, which (most of the time) has the preferred rank reserved for the active head of state. The ranking mechanism also introduces

18. https://www.wikidata.org/wiki/Help:Ranking
19. `PREFIX p: <http://www.wikidata.org/prop/>, PREFIX ps: <http://www.wikidata.org/prop/statement/>`
20. `PREFIX wdt: <http://www.wikidata.org/prop/direct/>`

semantic errors if the ranks get modified after creating the query. The ranking mechanism as a speciality of Wikidata challenges KGQA systems even more.

When a question gets complex, it is problematic to guarantee the correctness and completeness of its query e.g., *Which businesses are founded by the person in charge of Tesla?*. Here, our intuitive solution would be:

Despite the exact term business being signified, in QALD-10 benchmark dataset, we use a property path in the last BGP [21] that generalizes its denotation to include more general instances. For example, the entity `wd:Q28222602 Zip2` is also a company founded by the Tesla founder but is not linked to the Wikidata entity business (`wd:Q4830453 business`). Another solution would be using `UNIONS` to find every eligible entity, in the understanding of how this task looks for an average person. In general, creating query seems impossible due to no adherence to strict ontologies in Wikidata. However, that would increase the burden on KGQA systems which learn from the annotated SPARQL queries. This problem is more severe in writing-, music- and book-related queries.

### 4.5.4 Limit on Returned Answers

The limit on returned answers is a significant problem in creating this dataset. A substantial amount of questions have a limited result set due to Wikidata's factual base. For instance, the question: *List the novels that won the Modern Library 100 Best Novels (*`wd:Q671613`*)?* has one answer although one would expect 100 results with common sense. As a result, we discarded such questions from QALD-10 test set to maintain a righteous benchmark.

### 4.5.5 Special Characters

A number of questions are based on special characters . For instance questions, *Find all Turkish verbs ending with "uş" in their lemma.* and *When did the district of Höxter come into existence?* have special characters or ask for special Wikidata properties (see example below). We tried to keep those kinds of questions with their corresponding queries as much as possible to foster multilingual KGQA research.

### 4.5.6 Computational Limitations in SPARQL

SPARQL has limited capabilities to deal with numbers. For instance, there is a *lack of native normalization* for numeric values in Wikidata. For instance, a comparison or SPARQL result modifier like below does not take units into account: Therefore, 100 centimeters could be bigger than 10 meters. The following query takes units into account but requires a more complicated structure: Hence, simple comparison queries may require a high level of expertise of the Wikidata schema, which makes the KGQA task on these questions more challenging. Also, there are *rounding errors in calculations.* For instance, corresponding to question: *How many years did Steve Jobs take the role of Apple CEO?*, the SPARQL would be: The answer based on common sense is 14, however, the query execution produces the following: `?st = 01-09-1997`, `?et = 23-08-2011`, resulting in only

21. https://www.w3.org/TR/sparql11-property-paths/

13 full years. Thus, KGQA systems without common-sense reasoning capabilities fail in parts of the QALD-10 test set.

### 4.5.7 Endpoint Version Changes

Finally, version changes in the endpoint and endpoint technology, especially when switching between the graph stores HDT (Fernández et al., 2013) and Fuseki [22], can result in different answer sets. This is due to syntax errors and execution timeouts in their internal optimizations such as basic graph pattern[23] reordering or `rdf:type` indexing. Providing a stable endpoint in connection with a stable dump and dataset, see Section 4.2, helps to alleviate this challenge.

## 4.6 Summary

The QALD-10 benchmarking dataset is the latest version of the QALD benchmark series that introduces a complex, multilingual and replicable KGQA benchmark over Wikidata. We increased the size and complexity over existing QALD datasets in terms of query complexity, SPARQL solution modifiers, and functions. Also, we presented the issues and possible solutions while creating SPARQL queries from natural language. We have shown how QALD-10 has solved three major challenges of KGQA datasets, namely poor translation quality for languages other than English, low complexity of the gold standard SPARQL queries, and weak replicability. We deem solving the migration to Wikidata issue an important puzzle piece to providing high-quality, multilingual KGQA datasets in the future.

We were able to prove the appropriateness and robustness of the dataset by means of an ESWC challenge. Due to a pull request from the participant group,[24] we published two releases: the original QALD-10 challenge dataset in version 1.0 and an open upstream branch. Overall, the feedback of the participants on the dataset was positive.

In the future, we will focus on generating and using existing complex, diverse KGQA datasets to develop large-scale KGQA datasets with advanced properties such as generalizability testing (Gu et al., n.d.; Jiang and Usbeck, 2022) to foster KGQA research.

**Acknowledgements**

This work has been partially supported by grants for the DFG project NFDI4DataScience project (DFG project no. 460234259) and by the Federal Ministry for Economics and Climate Action in the project CoyPu (project number 01MK21007G). We thank also Michael Röder for supporting the GERBIL extension https://gerbil-qa.aksw.org/gerbil/config-qald.

22. https://jena.apache.org/
23. https://www.w3.org/TR/sparql11-query/#BasicGraphPatterns
24. https://github.com/KGQA/QALD-10/pull/6

# 5

# Biomedical Entity Linking with Triple-aware Pre-Training

## Bibliographic Information

Xi Yan, Cedric Möller, and Ricardo Usbeck. 2025. Biomedical Entity Linking with Triple-aware Pre-Training. In *Proceedings of the Third International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data (SemTech4STLD 2025), co-located with the Extended Semantic Web Conference (ESWC 2025),* edited by Rima Dessí, Joy Jeenu, Danilo Dessì, Francesco Osborne, and Hidir Aras. To appear. Portoroz, Slovenia: CEUR-WS.org, June. (Cited on pages 10, 12).

## Abstract

The large-scale analysis of scientific and technical documents is crucial for extracting structured knowledge from unstructured text. A key challenge in this process is linking biomedical entities, as these entities are sparsely distributed and often underrepresented in the training data of large language models (LLMs). At the same time, those LLMs are not aware of high-level semantic connections between different biomedical entities, which are useful in identifying similar concepts in different textual contexts. To cope with aforementioned problems, some recent works focused on injecting knowledge graph information into LLMs. However, former methods either ignore the relational knowledge of the entities or lead to catastrophic forgetting. Therefore, we propose a novel framework to pre-train the powerful generative LLM by a corpus synthesized from a KG. In the evaluations we are unable to confirm the benefit of including synonym, description or relational information. This work-in-progress highlights key challenges and invites further discussion on leveraging semantic information for LLm performance and on scientific document processing.

## 5.1 Introduction

Biomedical entity linking (EL) is a critical process in biomedical text mining that seeks to identify and associate relevant biological and medical entities mentioned in unstructured text with their corresponding identifiers in knowledge bases. EL systems have also been combined to promote the knowledge acquisition task(Noullet et al., 2023). Accurate recognition and linking of these entities are pivotal in promoting biomedical research, drug discovery, and personalized medicine (Chandak et al., 2023b). Although substantial progress has been made in recent years, there is an ongoing need for refining methods and techniques employed for entity linking in the biomedical domain.

In this report, we present a novel approach that integrates linearized (in which a graph is traversed and encoded when producing the linearized representationHoyle et al. (2021).) triples into the biomedical entity linking process while reevaluating the inclusion of synonym information. Our proposed method linearizes triples and considers them during the pre-training step. In past studies, synonym information, which involves using alternative names or terminologies for the same biomedical entity, has been proven to enhance entity linking when used during pre-training (Xu, Chen, and Hu, 2023; Yuan, Yuan, and Yu, 2022). Our study aims to build upon this existing knowledge by integrating both strategies and assessing their impact on performance.

Despite the reported benefits of synonym information in prior studies, our analysis of this approach, combined with the introduction of linearized triples (Li et al., 2021), yielded different results. We find that incorporating linearized triples only lead to minimal improvements in our entity linking model's performance. Moreover, we are unable to confirm the purported advantages of including synonym information in our experiments, which stands in contrast to the findings of previous literature.

We highlight the limitations of our study and suggest possible avenues for future research to further advance biomedical entity linking techniques by building on our work with linearized triples and reevaluating synonym information. The code is available at our GitHub repo [1].

## 5.2 Related work

Entity Linking has a long history of research. Recent methods can be categorized into two types. First, discriminative methods that are based on the bi-encoder / cross-encoder pairing (Ayoola et al., 2022; Logeswaran et al., 2019; Wu, Petroni, et al., 2020). Both encoders are commonly BERT-like models. The bi-encoder encodes the description of each entity and matches it to the text by using an approximate nearest neighbor search. This is important as the next step, the cross-encoding, is expensive. Here, those neighbors are reranked by applying a cross-encoder to the concatenation of both, the input text and the entity description. The highest-ranked entity is then the final linked one. In the biomedical domain, the works by (Angell et al., 2021), (Varma et al., 2021), (Agarwal et al., 2021) and (Bhowmik et al., 2021) fall into this category.

1. https://github.com/xixi019/bio-EL

Another type of entity linker is based on generative models (Cao et al., 2021; Cao, Wu, et al., 2022; Xu, Chen, and Hu, 2023). Here, instead of using some external description of an entity, the whole model memorizes the KG during training. The linked entity is then directly generated by the model. Such methods skip the problem of mining negatives which are crucial for a good performance of bi-encoder-based methods. BioLinkerAI (Sakor et al., 2024) and Gallego et al (Gallego et al., 2025) use the entity definitions and therasaurs (i.e., UMLS) to enhance the performance of LLM. Only the work by Yuan et al. (Yuan, Yuan, and Yu, 2022) is based on such methods in the biomedical domain. As generative models lack the ability to incorporate external information, they alleviate this problem by introducing a pre-training stage where syntactical information from a knowledge graph is learned. This is especially important in the biomedical domain as entities often own a large variety of synonyms. We build upon their work by extending the pre-training regime to the inclusion of triple information.

## 5.3 Method

### 5.3.1 Task definition

Given are a text $t$, a set of marked mentions $M_t$ in the text and a KG $\mathcal{G} = (\mathcal{E}, \mathcal{R}, E)$. The KG consists of a set of entities $\mathcal{E}$, a set of relations $\mathcal{R}$ and a set of edges composed of head entity, relation and tail entity $E \subseteq (\mathcal{E} \times \mathcal{R} \times \mathcal{E})$. The task is to identify the subset of entities $E_t \subseteq \mathcal{E}$ which the mentions $M_t$ are referring to.

### 5.3.2 Model

In the vein of the work by (Cao et al., 2021), we model the problem as a sequence-to-sequence generation task. The input to the generative model is text and the output are the generated entity identifiers in the corresponding KGs. Similar to other works (Cao et al., 2021; Cao, Wu, et al., 2022; Yuan, Yuan, and Yu, 2022), we consider the definition of the concepts in the corresponding KGs as the unique textual representation of each concept. The definition and synonyms are short and unique, and will not introduce the problem of ambiguation of entities.

### 5.3.3 Pre-training

We linearize the information from synonym and triples in the pre-training stage. An overview of the pre-training and an example is give in Figure 5.1 . They are linearized into a synthesized corpora before feeding into the BART. We have tested 2 different settings for converting the triples, namely **line-by-line** and **all-in-one**. We add triple pre-training step on which add the triples information to the LLM, on top of synonym, which is used by (Yuan, Yuan, and Yu, 2022).

In terms of the **synonym information**, we follow the setting by (Yuan, Yuan, and Yu, 2022). We first extract the description of the entity and convert it to a text of the following form:

$$\text{[BOS][ST] } s_e^a \text{ [ET] is defined as } c_e \text{ [EOS]} \tag{5.1}$$

Here, $s_e^a$ stands for the synonym $a$ and $c_e$ for the description of entity $e$. This would be the input to the encoder of the generative model.

As an output the model has to generate:

$$[\text{BOS}]\ s_e^a \text{ is } s_e^b\ [\text{EOS}] \tag{5.2}$$

This lets the model learn the connection between the different synonyms of the same entity.

Based on that, we introduce an additional pre-training step to incorporate more semantic information by utilising **triple information** from the underlying knowledge graph. A triple is of the form $< e, r, e' >$ which describes that a relationship $r$ holds between entity $e$ and $e'$. The input is here the same as for the synonym information. The output is of the form:

$$[\text{BOS}]\ s_e^a\ l_r\ s_{e'}^b\ [\text{EOS}] \tag{5.3}$$

$l_r$ is here the label of relation $r$. We denote this **line-by-line**. Furthermore, we experimented with an **all-in-one** pre-training approach of the form:

$$[\text{BOS}]\ s_e^a\ l_{r_1}\ s_{e_1}^{b_1}\ \ldots\ l_{r_n}\ s_{e_n}^{b_n}\ [\text{EOS}] \tag{5.4}$$

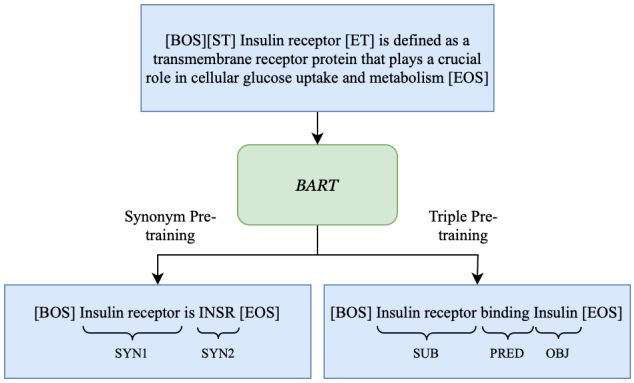



**Figure 5.1:** An overall workflow of our framework. We adopt different textualization formats for synonym information and triples. Both are included in the pre-training stage.

### 5.3.4 Fine-tuning

During fine-tuning, the model is trained for the actual entity linking task. The input to the generative model is the unlabelled biomedical text. To generate the linked entities, each mention is included in a template as follows:

$$[\text{BOS}]\ m_i \text{ is } s_e^a\ [\text{EOS}] \tag{5.5}$$

The model then generates the entity identifier after the token `"is"`. Similar to the work by Yuan et al. (Yuan, Yuan, and Yu, 2022), we choose the synonym which is syntactically close to the corresponding mention in the text as the target entity identifier during fine-tuning.

The generated entity identifier is mapped back to the concrete entity in the final step via a lookup table. During inference, we restrict the possible output space by limiting it to the available entity names and synonyms. See Figure 5.2 for an overview of the pre-training with an example.

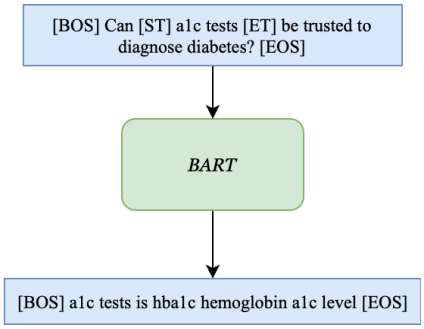



**Figure 5.2:** An overview of the fine-tuning stage.

## 5.4 Evaluation

### 5.4.1 Pre-training Strategy

We use a synthesized corpus composed of triples, synonyms and descriptions from UMLS. More specifically, we decide to use a subset of UMLS, st21pv (Mohan and Li, 2019). It is a well-connected KG with information about concept definitions and synonyms. Specifically, 160K out of 2.37M concepts have definitions, 1.11M concepts have several synonyms and 68K concepts are connected to on average 8 triples as a subject in a single hop. During the pre-training step, we construct samples by iterating through each concept's synonyms and triples. Each concept is densely connected and the distribution of the number of triples a concept is connected to is skewed. For instance, some "popular" concepts are connected to over 1000 triples, while some are connected to only 1 triple. To avoid the class imbalance, we sample the included triples based on the relation frequencies.

To train the model with KG information, we linearize triples. Linearization refers to a special type of technique on converting graph to text, i.e., converting triples to one/more sentences which serve as input of the LLM.

We sample the included triples based on the relation frequencies. First, we gather the occurrence frequency of all relations in the KB by counting the number of triples this relation is connected to.

Both settings are trained under the same experiment setting with a batch size of 128. We save the best model within 12 training epochs. We experiment with BART-base, bioBART-Large, and bioBART-Base. We choose BART to align to the work of (Yuan, Yuan, and Yu, 2022) so that we can make comparison about whether relational information is beneficial to the model. Note that we define the probability ($P_r$) of a relation $r$ to be negatively related to the frequency. Then, for each concept in the KG, we collect its connected triples and segment the triples into different groups based on their relation $r$.

#### Fine-tuning

The model is fine-tuned on two established datasets, namely BC5CDR (Li et al., 2016b) and NCBI (Dogan et al., 2014). Those entity linking datasets are constructed on subsets of UMLS, making them perfect choices to test our model's performance on. Among the datasets, NCBI and BC5CDR are generated by annotating PubMed papers. On the other hand, NCBI and BC5CDR are annotated against Medical

Subject Headings (MeSH) - a terminology knowledge graph for indexing and cataloging of biomedical information.

The statistics of the four datasets are exhibited in Table 5.1 below. As we can see, NCBI and BC5CDR (annotated on academic text) are smaller in size. Also NCBI and BC5CDR are dense in terms of the target entities they contain (14,967 and 268,162).

**Table 5.1**: Numbers of the samples in the training, development and test set

| Nums | NCBI | BC5CDR |
|---|---|---|
| Train | 5,784 | 9,285 |
| Dev | 787 | 9,515 |
| Test | 960 | 9,654 |
| Entities | 14,967 | 268,162 |

BART-large (Lewis et al., 2020) is chosen as the generative model as it has been an established benchmark model for such tasks.

### 5.4.2 Results

We assess the performance of four distinct models in the entity linking task, including two of our own models, each pre-trained via either a line-by-line or all-in-one strategy, a synonym pre-trained model from (Yuan, Yuan, and Yu, 2022) (denoted Syn-Only), and a basic BART model. We also include the recent papers which pretrains BART on biomedical domain (Yuan, Yuan, Gan, et al., 2022) before finetuned on biomedical entity linking datasets and ResCNN (Lai et al., 2021) which achieves state-of-the-art results on various biomedical EL datasets. Each model undergoes fine-tuning specific to the entity linking task. Recall@1 for each model are presented in the Table 5.2. We limit ourselves to Recall@1 to follow the common practice when measuring entity linking performance without named entity recognition. The best-performing metrics are emphasized in bold.

**Table 5.2**: Recall@1 on BC5CDR and NCBI, which are PubMed articles annotated against MESH.

| | BC5CDR | NCBI |
|---|---|---|
| Syn-Only | 93.3% | 91.9% |
| Syn-Only | 92.68% | 89.45% |
| All-in-one | 92.86% | 88.43% |
| Line-by-line | 92.66% | 90.00% |
| BART | 92.58% | 89.06% |
| BioBART-Large | 93.01% | 89.27% |
| BioBART-Base | **93.26**% | 89.40% |
| ResCNN | 91.7 % | **92.4**% |

### 5.4.3 Analysis

Based on the table 5.2, our triple injection framework exceeds the BART baseline on the 2 benchmarks datasets. On BC5CDR and NCBI, the gain compared to BART is around 0.2% and 0.5%.

**Does triple injection enhance model's capacity to link to the correct entity?** The answer is yes, since over 2 datasets, the All-in-one or Line-by-line variants outperform the variant that was not trained on the linearized corpora for around 1% (Recall@1).

## 5.5 Conclusion

Our study seek to improve biomedical entity linking through the integration of linearized triples and synonym information. However, contrary to expectation, the incorporation of these elements leads to only minimal improvements in our EL model performance.

In conclusion, our study underscores the complexities of biomedical EL and prompts the need for more sophisticated approaches to improve its accuracy. A possible future extension of this work could be to explore more sophisticated methods to instruct the LLMs to learn external knowledge, such that the knowledge is injected in an efficient way which benefits the models in downstream tasks. For instance, by incorporating the KG information not just in a linearized manner but by exploiting the graph-structure with Graph Neural Networks (Wu, Pan, et al., 2020), mutliple methods could be further developed.

## 5.6 Acknowledgments

This work has been partially supported by the Ministry of Research and Education within the project 'RESCUE-MATE: Dynamische Lageerstellung und Unterstützung für Rettungskräfte in komplexen Krisensituationen mittels Datenfusion und intelligenten Drohnenschwärmen' (FKZ 13N16844), by the Federal Ministry for Economic Affairs and Climate Action of Germany in the project CoyPu (project number 01MK21007[G]). We utilized 2 x NVIDIA RTX A5000 24GB kindly provided by the NVIDIA Academic Hardware Grant Program. The authors have no competing interests to declare that are relevant to the content of this article.

## Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-4 and Gramby in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

# 6

# Neuro-symbolic Relation Extraction

## Bibliographic Information

Xi Yan, Aida Usmanova, Cedric Möller, Patrick Westphal, and Ricardo Usbeck. 2025. Neuro-Symbolic Relation Extraction. In *Handbook on Neurosymbolic AI and Knowledge Graphs,* 400:550–576. Frontiers in Artificial Intelligence and Applications. IOS Press. (Cited on pages 8, 10, 12).

## Abstract

Neuro-symbolic relation extraction lies at the intersection of neural networks and symbolic reasoning, presenting promising opportunities to enhance the capabilities of natural language processing (NLP) systems. Despite its potential, a comprehensive review of how these systems are developed and applied to the task of relation extraction has been lacking. This chapter addresses this gap by offering an in-depth overview of the current landscape in neuro-symbolic relation extraction, focusing on key methodologies and the datasets utilized in this field. We systematically categorize existing approaches, emphasizing how they integrate neural and symbolic components to tackle various challenges and the types of information they incorporate. Additionally, we review the datasets used to evaluate neuro-symbolic relation extraction systems, detailing their statistics, creation processes, and underlying domains. Furthermore, we discuss future research directions and challenges, such as the analysis of symbolic information and the integration of datasets with existing knowledge graphs. By synthesizing these findings, this chapter aims to provide researchers and practitioners with a clear understanding of the state of neuro-symbolic relation extraction and to inspire further innovations in this rapidly evolving field.

## 6.1 Introduction

RE is a fundamental problem in the area of Natural Language Processing (NLP). The goal is to classify the *relation* expressed between two *entities* in unstructured texts. An entity represents a notion of a “thing" and a relation is an association between two entities. Entities could be words, phrases or other syntactic units. Together, two entities and one relation form a *triple* (Chen et al., 2020), formally defined as $(e_1, r, e_2)$. For example, in the following sentence, we can identify two entities “Barack Obama" and “Michele Obama" and multiple potential relations to be extracted: Barack Obama married Michele in 1992.

One could extract that the relation `spouse` holds between Barack Obama and Michele. Here, we denote `spouse` as an identifier for the actual relation defined as “significant other in a marriage”. Other human-readable identifiers such as `married_to` or non-human-readable identifiers such as `P26` (the identifier given to the relation in Wikidata (Erxleben et al., 2014)) are possible as well.

Another potential relation occurring in the sentence is that the `wedding_year` of the marriage was 1992. In this case, the two entities between the relation holds are actually the statement “Barack Obama married Michele” on the one hand and the year “1992” on the other. Such a statement is also called a hyper-relational statement, essentially a statement on a statement (Galkin et al., 2020). These can be arbitrarily complex.

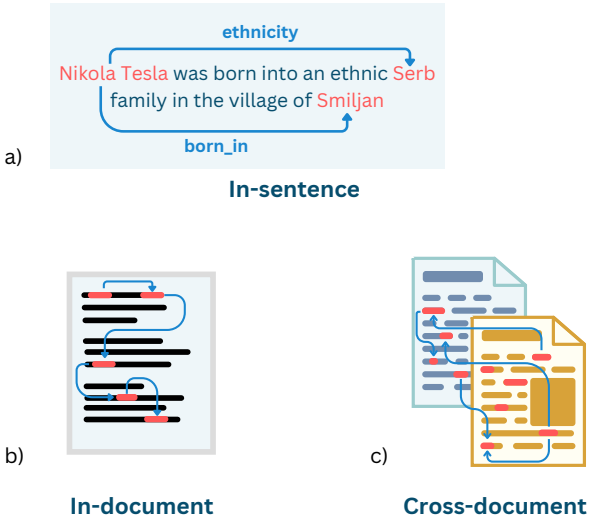



**Figure 6.1**: Different Relation Extraction types: a) in-sentence, b) in-document, c) cross-document.

Relation Extraction for such arbitrary complex relations can be applied to different types of texts: in-sentence, in-document, or cross-document (Han et al., 2020) as illustrated in a, b and c in Figure 6.1. In-sentence relation extraction focuses on entities occurring only in a single sentence. In-document relation extraction is occupied with a single document on a certain topic containing multiple sentences. Especially, are of interest relations holding between entities occurring in different sentences. This usually makes some type of more complex reasoning necessary. Finally, cross-document relation extraction extends this not only to a single

document but also to several documents with relations holding between entities that are occurring in separate documents. Additionally, many works focus on only identifying which relations hold between two entities, ignoring to cases where no relation holds. Predicting the non-existence of a relation is especially important when confronted with an in-document or cross-document relation, as the number of potential relations increases quadratically with an increasing number of occurring entities in the text.

Notably, some other tasks in NLP, such as event extraction (Ahn, 2006) and syntactic parsing (Gorrell et al., 1995), fall within this definition. Both tasks aim to extract relationships (syntactic or event-related) between entities in unstructured texts. However, we do not include them here for two reasons. First, research in syntactic parsing and event extraction often does not label itself as relation extraction, making it difficult to gather relevant studies. Second, the extensive body of work in these areas makes it impractical to cover all the research within this chapter. For the same reason, we do not consider work which contains multimodal relation extraction, but rather focus on the text. Multimodal relation extraction is a process that involves extracting meaningful relationships from multiple types of data sources, such as text, images, and videos.

Also, the task of Relation Extraction is closely related to other tasks such as Named Entity Recognition (NER), knowledge graph population or knowledge graph question answering. NER (Yadav and Bethard, 2018) extracts entity mentions from text for further processing and is a prerequisite to relation extraction. In contrast to NER, Knowledge graph population and knowledge graph question answering (Pereira et al., 2022) are tasks where Relation Extraction is necessary. In knowledge graph population (Ye et al., 2022), new triples need to be extracted from given text, where each triple usually consists of two entities connected by a relation. In knowledge graph question answering, especially when focusing on semantic parsing, extracting relations is essential for constructing a correct query (Galstyan, 2022).

While machine learning models have achieved high performance on Relation Extraction from pure text, they face robustness problems. These methods rely on large-scale training data and complex feature engineering, which limit the model's capacity to generalize to new entities and domain-specific texts. Introducing symbolic information is a crucial step in addressing these problems. Compared to machine learning models, which produce unstable results, symbolic information, such as Knowledge Graphs or logic, offers clear reasoning paths and robustness but lacks the flexibility and scalability of neural networks. By integrating neural and symbolic techniques, neuro-symbolic relation extraction methods aim to create more interpretable and generalizable models that can effectively understand and process complex relationships within text. In the sections below, we categorize and define common relation extraction methods, which are based on machine learning models, and neuro-symbolic relation extraction methods, which add symbolic information on top.

### 6.1.1 Common Relation Extraction

Today, Relation Extraction is commonly solved by applying a Pre-trained Language Model (PLM) (Devlin et al., 2019b) to the input. This means either encoding the

input via an encoder-only model such as BERT (Devlin et al., 2019b) and then applying a classification head on top of it, or, alternatively, (encoder-decoder or decoder-only) generative models (Raffel et al., 2020b; Touvron, Lavril, Izacard, Martinet, Lachaux, Lacroix, Rozière, Goyal, Hambro, Azhar, et al., 2023b) are used to directly generate the relation label. Generative models are especially suitable if the goal is not only to classify a relation, but to extract full triples (Josifoski et al., 2022).

Beyond the standard definition and methodologies, Relation Extraction can also be approached in few-shot or zero-shot settings (Qu et al., 2020). In these scenarios, the relations encountered during training are different from those encountered during inference. Thus, the model must generalize to new, unseen relations with either a few examples (few-shot) or no examples (zero-shot) provided during training. These settings challenge the model to leverage its learned knowledge to make accurate predictions for novel relations.

### 6.1.2 Neuro-symbolic Relation Extraction

While PLM-based Relation Extraction methods are powerful and efficient in extracting information, they suffer from problems such as hallucination due to instability and weak reasoning capacity (Rawte et al., 2023). Therefore, researchers attempt to add further concrete and verified information to address the lack of contextual information and to deepen the reasoning capacity of these models.

We define Neuro-Symbolic Relation Extraction as any relation extraction method that relies not only on the textual information at hand, but also on additional symbolic information related to the entities or relations in context. This information can either be explicitly defined for the task at hand or gathered from sources such as a KG. Among various types of symbolic information added to existing neuro-symbolic RE models, we categorize them into KG, linguistic, prior meta-information and logical rules. A more detailed definition can be found in section 6.2.2. In the next chapters, we systemically categorize and list all neuro-symbolic methods from published papers after 2021 and the datasets used by them.

## 6.2 Methods

Neuro-symbolic relation extraction represents a trend in the field of natural language processing of combining the strengths of neural networks and symbolic reasoning to enhance the generalizability and interpretability of information extraction. This section enumerates an exhaustive list of recent neuro-symbolic relation extraction, categorizes them systemically, and summarizes the methods with a focus on how incorporation of symbolic data improves the former methods in terms of reasoning capabilities and advantages.

Existing methods can be separated into two categories, each representing a different way of incorporating symbolic information. First, symbolic information can be utilised to **improve distant supervision** by using ontological information to mine or refine labels of textual data. Second, symbolic information is used during training and inference as additional information to **improve the relation extraction** performance directly. Through a detailed examination of recent ad-

vancements, we will illustrate how neuro-symbolic techniques are paving the way for more robust and reliable relation extraction systems. The papers are listed in chronological order in the improving distant supervision section and we categorize the works in the improving relation extraction section by the types of the symbolic information.

### 6.2.1 Improving Distant Supervision

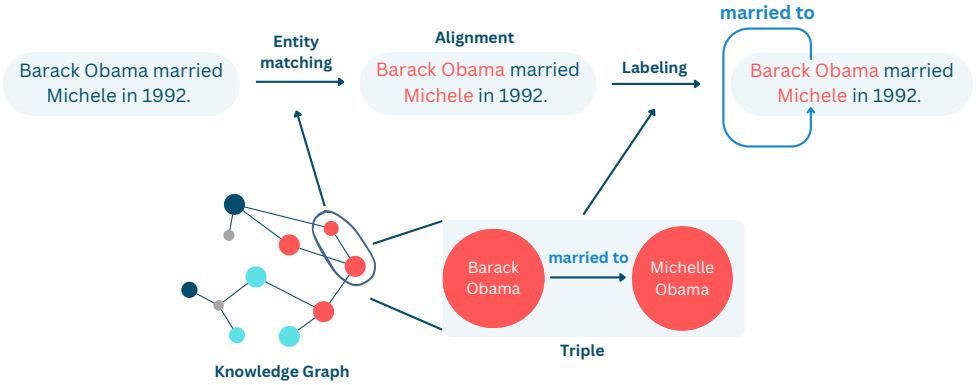



**Figure 6.2:** Alignment of entities from the sentence with the triple information from the KG. First, entities are matched and aligned with nodes from KG. Then, the relation between entities is labeled based on the edge between the matched nodes in KG.

Distant supervision is the idea of taking unlabeled data and labeling it via heuristics or an existing method. In the realm of Relation Extraction, this means identifying *entities* and the *relations* between them.

The *distant supervision assumption* implies that any sentence, containing two entities that participate in a relation, will express that relation (Mintz et al., 2009). Such an approach is a common way to identify entities and the relations between them by using the wealth of information available in a Knowledge Graph. Given an unlabeled sentence/document(s), it is checked whether two mentioned entities in the sentence/document(s) are connected by a triple (see Figure 6.2). If they are, it is assumed that the relation expressed by the triple is also expressed in the sentence.

This approach presents several difficulties. First, the entities need to be connected to the actual correct entities in the Knowledge Graph, this is a challenging research problem known as entity linking (Sevgili et al., 2022). Second, the mere co-occurrence of two entities within the same sentence does not necessarily indicate that the relation of a connecting triple is being expressed. Third, in connection to that, there might exist multiple relations between entities, which are usually not expressed at the same time in a single sentence. Lastly, most Knowledge Graphs are extremely sparse and incomplete, not covering everything expressed in an actual sentence. This leads to a lot of false negatives (Tan et al., 2022). Nevertheless, pre-training on distantly supervised data leads to an improvement in performance when combined with fine-tuning on manually annotated data (Yao et al., 2019). While there exist some approaches like bagging (Ye and Ling, 2019) to reduce the influence of the noisy distantly supervised labels, utilising symbolic information

such as n-hop paths in a Knowledge Graph or actual ontological information can improve the quality of distantly supervised labels as well.

The work by Dai et al. (Dai et al., n.d.) utilises an universal graph, here denoted as the combination of a KG and a large-scale text collection. The idea is that the sparse information in the KG is compensated by the additional textual documents, and together they act as distantly-supervised information. For each of the sources, textual and KG, multi-hop paths are computed and encoded. This gives textual, KG and hybrid (combining both text and KG) paths. An attention mechanism is used to combine the different sources of information. The authors suggest that pre-training the model sequentially on the different sources of information is crucial. Otherwise, the information in the KG has an excessive impact and the model is biased towards the information in it. After pre-training the model sequentially, they finally train on all sources together. This procedure is demonstrated to alleviate the problem of the bias. They show that the combination of the two sources leads to an increase in performance.

The *ASP-enhanced Entity-Relation extraction (ASPER)* framework, developed by Le et al. (Le et al., 2023) refines distantly-supervised labels by applying Answer Set Programming (ASP). They start with labeling an unlabeled dataset using an already trained model. Then they translate the underlying rules of a KG into ASP atoms. Those encompass type declarations targeting the domain and range of a relation, inference rules targeting relationships between relations, and optional rules that are dataset-specific. Given the atoms, the answer set with the highest score is computed using an ASP solver. Then, the answer sets which contain new pseudo-labels are used for retraining. Importantly, the retraining starts with a high confidence answer and progressively proceeds to train on lower-confidence answers as well. The inclusion of labels gathered in such a way significantly increases the performance. The downside is the additional effort for identifying and formulating the underlying rules describing a KG. However, in comparison to manually labeling such a large amount of data, this is often the preferred option. The quality of the distantly-labeled data increases and leads to superior performance of the RE model.

**Table 6.1**: A table for different types of symbolic information used in distant supervision model.

| Source | Paper | Info |
|---|---|---|
| KG | Dai et al. (Dai et al., n.d.) | KG path |
| KG | ASPER (Le et al., 2023) | Type declaration |

To sum up, there exist only a few works using symbolic information to improve distant-supervision, marking a significant research gap. The existing methodologies leverage external sources, such as subgraph paths and entity types, necessitating entity linking due to the extensive information available in KGs. Following the alignment of text to the KG, the methods for integrating KG information vary. Some approaches directly enhance the text with KG-derived information, while others utilize KG embeddings or generate constraints to incorporate the knowledge. However, none of the papers investigate whether the additional information introduced is noisy or biased, given the overwhelming

amount of data typically contained in KGs. This issue is addressed by incorporating an attention mechanism, allowing the model to maintain a strong focus on the text. Nevertheless, it was shown by the existing works that refining distantly-supervised data using ontological information can lead to an increase in quality. Examining the impact of ontological information together with the generative power of PLM might be a direction to pursue in the future. Additionally, the optimal methods for aligning and filtering information from KGs with PLMs are still to be explored and examined. An overview of all systems in this category with their source and the detailed information can be found in Table 6.1.

### 6.2.2 Improving Relation Extraction

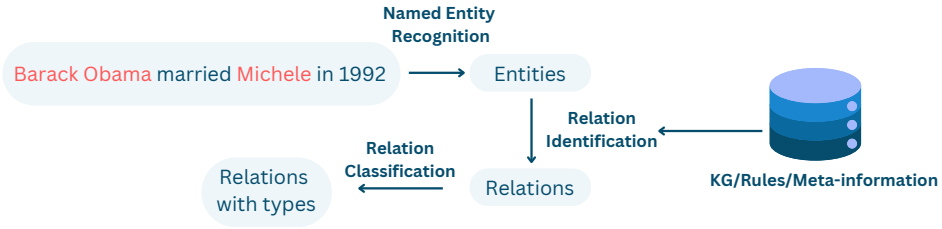



**Figure 6.3:** RE enhanced with external information from KG/logical rules/meta-information.

Symbolic information is believed to improve the performance of RE during inference as well. The key concept is that providing additional symbolic information on the included relations or entities will help identify the correct relation for the statement at hand (see Figure 6.3). In this section, we focus on methods in which the integration of neural and symbolic information extends beyond distant supervision. External information is not only concatenated with text, but is also used to generate constraints, which are applied during both the training and inference stages. Among the papers we reviewed, symbolic information can be categorized into several types: KGs, prior meta-information, inferred logical rules, linguistic information, temporal information and commonsense rules.

#### Knowledge Graphs

KGs are naturally discrete symbolic representations of information. Symbolic information in KG includes entities, relations, attributes, labels, hierarchies, temporal information, classes, etc. Such information originating from a KG can be used during inference or training to enhance the contextual information and reasoning capacity of the Pre-trained Language Model (Bastos et al., 2021).

Entity attributes, like labels, descriptions, and type information, can be included as a first step of RE. For instance, the *Relation Extraction using Knowledge Graph Context in a Graph Neural Network (RECON)* (Bastos et al., 2021) approach automatically identifies relations in a sentence by encoding the sentence and incorporating additional entity-related information from a KG. Information from

the KG is included in two ways. First, **Entity Attribute Context** are encoded via a word embedding method and combined using a Convolutional Neural Network (CNN). Second, **Triple Context** is incorporated by encoding each entity and relation by an initial vector. Each triple is represented by a projected concatenation of subject, object and relation. To gather a contextual representation, neighbouring triples are aggregated using an architecture resembling a Graph Attention Network (GAN). Finally, for a pair of entities where a relation is to be predicted, the entities are mapped to a relation-aware representation and a score is calculated using the translational equation $e_1 + r = e_2$. The Triple Context representation is combined with the Entity Attribute Context representation and a text representation to predict the final relation. Especially, the Entity Attribute Context has a large impact.

*Knowledge Enhanced Few-shot RC model for the Domain Adaptation* (KEFDA) (Zhang, Zhu, et al., 2021) uses the same entity information as additional features for few-shot RE. By incorporating general and domain-specific Knowledge Graphss they address the problem of few-shot RE. The method consists of two elements, first, a **knowledge-enhanced prototypical network** that combines the representations of the support set. The enhanced representations are computed from the contextual representation of the input sentence, the entity descriptions and a learned representation of the concept. The matching degree between instances is calculated by a distance based scoring function. Second, they use a **relation-meta learning network**, to identify meta-relations between the involved concepts. For that, they solely rely on the learned concept representation and the contextual representation of the concept description. They create each possible concept pair, calculate a pair-wise representation and finally a weighted average over all such pair-wise representations to get the plausibility that such relation exists. The entity description and concept representations had the largest impact on performance and led to superior performance compared to past methods, while the influence of the meta-relations was less significant.

The *Discriminative Rule-based Knowledge (DRK)* approach (Wang et al., 2022) introduces a external KG and logic rules in the encoding and inference stage of relation extraction model. This work adopts a few shot learning setting, in which each sample includes a support set (labeled sentences) and the query set (unlabeled sentences). Both sets are linked to an external KG to extract the related information of the entity and relation (e.g., entity types and relation description). The entities, text and relation representation would then be encoded separately before fed to the logic aware inference module. Based on the relation information, some logic rules can be inferred, for instance, the relation "Mother" implies that the type of the head and tail entity should be "Person" and the relation description is semantically equivalent to "Mother". Based on the inferred rules, the authors propose a hierarchical contrastive learning optimized against the probability distribution of the relations. The model is tested on one RE dataset and outperform former baselines.

The work by Liu et al. (Liu et al., 2023) focuses on few-shot relation extraction. They combine a contrastive-learning-based fine-tuning approach with knowledge enhancement. To enhance the stability and learning ability of contrastive learning-based fine-tuning, the authors use a data augmentation mechanism and type-aware networks to enrich the instances and incorporate class-sensitive features.

Knowledge enhancement is achieved by incorporating type information and by including pre-trained entity node representations. Especially the knowledge enhancement leads to a significant improvement in performance.

ReOnto (Jain et al., 2023) is an approach incorporating a KG to improve RE in the biomedical domain. Here, KG paths between mentioned entities are extracted and used in parallel to the text-only method to improve the accuracy of the relation extractor. Furthermore, they not only extract $n$-hop paths but also include ontological information expressing, for example, class subsumption. However, the ontological information is included by encoding them through a PLM, no formal logic is actually used. The additional information has a significant impact and surpasses the performance of the same model which only relies on the textual data.

The work by Aghaebrahimian et al. (Aghaebrahimian et al., n.d.) introduces ontological information into the task of biomedical relation extraction. They accomplish this through two perspectives: 1) they enrich the encoded input text with type embedding of the entity, and 2) they include ontology graph embedding. The ontology graph embedding are here node embedding of specific concepts, such as genes or diseases. The node embedding is learned by first generating paths using random walks and encoding them using Skip-Gram (Mikolov, Sutskever, et al., 2013). They show that the inclusion of those two types of information has a positive effect on performance. They either use an ontology available with the dataset or the Unified Medical Language System (UMLS) (Bodenreider, 2004b) (an unified biomedical database for terminology).

### Linguistics

Pre-trained language models (PLMs) typically project words into embeddings. Beyond this, the integration of linguistic and lexical knowledge associated with entities allows PLMs to effectively tackle fine-grained tasks such as named entity recognition (Li et al., 2020), relation extraction (Tuo and Yang, 2023), etc.

The *Sememe Knowledge-enhanced Abstract Meaning Representation and Reasoning (SKAMRR)* (Zhao et al., 2023) approach introduces sememe knowledge from HowNet (Fan et al., 2022) (a sememe Knowledge Graph) to enhance entity representation. A sememe is the minimum semantic unit in linguistics, and HowNet proposes that annotated sememes can represent senses and words well in a real-world scenario. Therefore, the authors extract the word senses and sememes from the OpenHowNet API (Qi et al., 2019) and combine this linguistic information with the textual feature of the input text for extracting an Abstract Meaning Representation (AMR) graph. An AMR graph, on the other hand, models the relations between entity pairs. The AMR graph is then sent to a reasoning module that builds an entity-pair graph, which is used for relation classification with a Global Adaptive Loss (GAL) to alleviate imbalances of the data. SKAMRR is tested on four document-level relation extraction datasets, including texts from Wikipedia and biomedical documents. The model shows competitive performance across all datasets which demonstrates its capacity to explore feature information within and across sentences, and infer classes of relations between entities.

Jain et al. (Jain, Mutharaju, Kavuluru, et al., 2024) focuses on the task of document-relation-extraction. The model consists of two submodules, first a document based encoding, combining two elements: 1) a graph consisting of mention

and sentence nodes, and 2) attention-based pooling of the contextual encoding of the input document. Second, external KG information is incorporated by learning entity node representations through a relational graph convolutional network (RGCN). Both types of information are combined and used in the final prediction. The method uses information about existing triples as well as WordNet(an English lexical database) (Fellbaum, 1998) entity synonyms and types. They use Wikidata as the background KG. Especially, the inclusion of the Wikidata information leads to a large increase in performance on three different datasets.

Jain et al. (Jain, Mutharaju, Singh, et al., 2024) develop a cross-document relation extraction method, relying on two types of information: the type information of the source and target entities, and the paths connecting both entities in the KG, specifically Wikidata. These paths are transformed into textual paths and concatenated with the textualized entity types to create a context, which is used to filter relevant sentences in the documents. The selected sentences are then concatenated with the context and further filtered based on their relevance to the entity path. Finally, the sentences and context are encoded using a specialized attention module generated from the text path for the final classification of the relation between entity pairs. This framework demonstrates improvement over baseline models on one benchmark dataset and also provides explainability for deep learning models.

**Prior Meta-information**

Prior meta-information refers to any pre-existing, heuristic or statistical, information about the entities and relations between them. Systematic incorporation of prior meta-information is also among popular approaches within RE. Using meta-information provides a context to the text, guiding the RE process. Examples include, entity co-occurrence, precomputed entity similarity scores, etc. Such information can help to disambiguate entities and reduce noise.

The *RAtionale Graph* (RAG) (Zhang, Yu, et al., 2021) assumes that global co-occurrence statistics among relations, entity types and trigger words are known. Using those, they build a *rationale graph.* This graph contains nodes corresponding to relations, trigger words or entity-type-pairs. Then, directed edges are introduced between trigger words and entity-type-pairs (in both directions) and from triggers words/entity-type-pairs to the relations. The edge weights correspond to the co-occurrence probability between the respective elements. During each inference step, the type-pair is predicted, and its representation is merged with the type-pair representation in the graph. The same happens for the trigger words. The whole graph is updated using a graph neural network. Lastly, using the co-occurrence probabilities, the relation representations are updated each and combined with the textual encoding to predict the final relation. This leads to a strong improvement over a model not using the co-occurrence statistics.

Specifically for extracting the relation from news articles, Zhang et al. (Zhang, Lyu, et al., 2023) model the inter-document casual relations by generating the storytrees. The authors firstly extract the keywords from the articles, to feed to EventX (Liu et al., 2020), a clustering algorithm which outputs events by gathering the stories. Those event clusters are converted into storytrees with nodes being the articles with high semantic similarity scores. Based on the generated trees,

two basic constraints (discouragement of causal relations between two news articles occurring at the same time and with very low similarity) and five types of constraints generated accordingly are generated based on the information of nodes and links. The constraints are utilized by Integer Linear Programming (ILP) (Gao, Choubey, et al., 2019) methods to perform global inference for classification of the relations with added information based on the articles' textual features. The model is tested on three datasets crawled from news articles and outperformed the extensively used classifiers and a state-of-the-art deep learning models.

Xu et al. (Xu, Chen, and Zhao, 2023) focus on document relation extraction. They encode the input text via an encoder-only model, then construct a graph consisting of mention and entity nodes, and compute a structural embedding using a Graph Neural Network (GNN). Furthermore, they construct the shortest possible paths between two entities in a document and encode them using a Long Short-term Memory (LSTM) network. All such path representations are pooled while using the attention mechanism. The pooled representation is used together with the structural embedding to classify the final relation. Incorporating the path-based representations helps to interpret long-range dependencies between entities and leads to an increase in the model performance.

The *Graph-based Model Generation (GM_GEN)* framework (Li and Qian, 2022) includes relation description as part of the few shot examples for the few shot relation extraction task. A PLM-based encoder firstly takes query samples (10 unlabeled sentences), relation description and support samples ($10 \times 5$ labeled sentences) as input. The encoded information is then sent to a graph-based generation module, which aims to combine the topology information with the attributes of samples and the relation descriptions. This module forms a graph matrix, with the nodes being the samples from query, relations or support samples and calculate the edges between them using a bilinear transformation based on the embedding from the PLM encoder. This matrix is then passed to each few shot learning relation extraction task for offering knowledge background. This framework is then applied to a CNN and a Bidirectional Encoder Representations from Transformers (BERT) model on two benchmark datasets while being compared to other CNN and BERT-based models. GM_GEN performs the best among all methods using the same encoder.

### Logical Rules

In some cases, employing a pre-defined set of rules could enhance the performance of RE, particularly within well-defined domains, like medical reports or financial statements. Rule-based approaches are exceptionally effective when dealing with legal documents, templates and forms, where patterns and specific phrases are prevailing. Such domains benefit from consistency and precision that rule-based approach methods offer, ensuring reliable and accurate results. Many studies also favour this approach when dealing with limited or small labeled datasets, as it does not require large amounts of training data.

The Knowledge-evolving Framework by Iterative Consolidation and Expansion (KICE) model, developed by Lu et al. (Lu et al., 2023), employs the Masked Language Modeling (MLM) paradigm to generate rules for each relation. Initially, a rule-generator is trained with a few training examples. The generator generates an

entity pattern, assigning to each entity a type, based on the input. Secondly, it generates a relation pattern, assigning to each entity-pattern combination a relation. Both such patterns together give a rule. Using such a trained rule-generator, rules are created for a larger set of unlabeled data. From this point on, the model is trained repeatedly on the same data using the soft-labels from previous iterations. At each iteration, human-annotators are included, for example where the most-matching rule has high ambiguity. A rule matches if the cosine similarity of the encodings surpasses some threshold. The approach achieves competitive results while needing a much smaller number of training examples than traditional approaches.

A study by Fan et al. (Fan et al., 2022) deals with documents, it presents a logic enhanced framework that boosts *Document-level Relation Extraction (DocRE)* by *Mining and Injecting Logical Rules (MILR)*. Upon the other DocRE models which optimize the logit towards prediction of correct relation type, MILR has modules which optimize the model towards logical consistency during training and inference. The logical consistency is referring to the logical rules, which are mined based on frequency from the annotated relations, with a confidence score estimated by the conditional probability of the co-occurrence of the relations. In the training phase, the authors combine the loss function from relation classification with the consistency regularization loss for predicting the logical rules. The trained model is then used to perform inference which aligns with the logical rules. The predicted logit over relation classification is taken as silver label, combined with rules and become a new objective to optimize against, which result in a set of logical constraints. The MILR is then applied on top of four established DocRE systems and tested on two datasets. It turns out that MILR is model agnostic and consistently improves the performance of other models.

The *weighted multi-channel transformer* (WMCT) (Haotian et al., 2024) focuses on the task of document-relation-extraction. It is noted that useful information about an entity in the context of a document, like its pronoun and keywords, are neglected. Authors advocate that such information should be involved in the graph as new types of nodes. The entity-level graphs are constructed based on several rules, and eventually consists of mention, pronoun, evidence word and dependency nodes of a given entity. The GAN, applied on each graph, aggregates the node representations of each mention node in the end. Then the pairwise concatenated representation of the mention nodes predicts the relation. Interestingly, the multi-view processing via the different graphs have a positive impact on the overall performance.

In conclusion, numerous neural-symbolic methods enhancing relation extraction models during both the training and inference stages. Unlike distant supervision, these methods incorporate a greater variety of symbolic information types. Some methods use the information directly, such as adding entity descriptions to textual representations, while others use encoded information, such as integrating structural KG ontology embedding into sentence representations. Additionally, some approaches transform this information into constraints for training, like contrastive learning. These advancements highlight the potential and effectiveness of combining statistical PLMs with more definitive symbolic information, thereby improving the robustness and generalizability of the models. We summarize the systems discussed in this section in Table 6.2.

**Table 6.2**: A table for different types of symbolic information used in improving relation extraction.

| Source | Paper | Info |
|---|---|---|
| KG | RECON (Bastos et al., 2021) | KG path |
| KG | KEFDA (Zhang, Zhu, et al., 2021) | Entity description, concept description |
| KG | DRK (Wang et al., 2022) | Entity types, relation description |
| KG | Liu et al. (Liu et al., 2023) | Type information |
| KG | ReOnto (Jain et al., 2023) | Class subsumption, KG paths |
| KG | Aghaebrahimian (Aghaebrahimian et al., n.d.) | Entity type, ontology graph |
| Linguistics | SKAMRR (Zhao et al., 2023) | Word sense, sememe |
| KG + Linguistics | Jain et al. (Jain, Mutharaju, Kavuluru, et al., 2024) | WordNet entity synonyms and types |
| Prior meta-information | RAG (Zhang, Yu, et al., 2021) | Co-occurrence statistics |
| Prior meta-information | Zhang et al. (Zhang, Lyu, et al., 2023) | Generated constraints |
| Prior meta-information | Xu et al. (Xu, Chen, and Zhao, 2023) | Shortest possible paths |
| Prior meta-information | GM_GEN (Li and Qian, 2022) | Relation description |
| Logical rules | KICE (Lu et al., 2023) | Entity and relation pattern |
| Logical rules | MILR (Fan et al., 2022) | Logical consistency |
| Logical rules | WMCT (Haotian et al., 2024) | Entity-level graph with pronoun and mention |

The integration of neural-symbolic methods in Relation Extraction has shown significant promise, leveraging symbolic reasoning to enhance the performance of deep learning-based language models. Despite progress in data-efficient few-shot and zero-shot learning, these methods still rely heavily on diverse and robust datasets that can be effectively combined with external information. Therefore, the choice and quality of datasets are pivotal in advancing neural-symbolic models, enabling them to learn intricate patterns and relationships more effectively.

## 6.3 Datasets

There exist various datasets focusing on the problem of Relation Extraction. These datasets exhibit considerable variation in the type of input texts across multiple dimensions.

Firstly, they can be differentiated by the specific relation extraction task they are designed to address, which could be at the sentence level, document level, or even across multiple documents.

Secondly, the datasets can be categorized based on their domains. These domains include, but are not limited to, legal texts, biomedical literature, biological studies, wikis, news articles, dialogues, and more. Each domain presents its unique challenges and nuances, which can significantly influence the complexity and nature of the relation extraction task. For instance, biomedical datasets often contain highly specialized terminology and require an understanding of complex scientific concepts, whereas news datasets might involve more general language but require handling a broader range of topics and events.

Lastly, datasets can also be distinguished by the number of relations they encompass. Datasets with a higher number of relations generally pose a more challenging relation extraction problem. This is because an increase in the number of possible relations typically leads to greater ambiguity and complexity. The more relations there are, the harder it becomes to accurately identify and classify the correct relation, as there are more possibilities to consider and distinguish between.

As far as we are aware, there exists no dataset with a dedicated focus on neuro-symbolic relation extraction. Despite this, to thoroughly investigate the impact of incorporating symbolic information into relation extraction tasks, it is crucial to determine whether such symbolic information is available within the existing datasets.

In the following sections, we will provide a comprehensive overview of all the different Relation Extraction datasets used in the aforementioned works. We will specify the type of each dataset, detailing the nature of the texts and the specific relation extraction tasks they are designed to address. Furthermore, we will examine whether symbolic information, which can include structured data, ontologies, or other forms of symbolic representations, is readily available within these datasets. This information is essential for researchers who aim to leverage symbolic approaches in their work, as it can guide the selection of appropriate datasets and inform the design of experiments and models.

### 6.3.1 Sentence Relation Extraction

As sentence-based Relation Extraction is the earliest existing type of Relation Extraction, the majority of existing datasets belong to this category. The majority of the datasets either belong to the news (Doddington et al., 2004; Riedel et al., 2010a; Roth and Yih, 2004; Yan et al., 2019) or biomedical domain (Gurulingappa et al., 2012; Hailu et al., 2013; Kruiper et al., 2020; Kulkarni et al., 2018; Pyysalo et al., 2007; Zhao et al., 2016) with some recent datasets focusing more on wikis (Ellis et al., 2013; Gao, Han, et al., 2019) or even dialogs (Yu et al., 2020). Possible reasons for that are on one hand the large interest in an working relation extraction in biomedicine on the one hand the abundance of news articles on the other.

In the following, we shortly summarise each sentence-based relation extraction dataset used in the previously presented works. Furthermore, an overview of the datasets can be found in Table 6.3.

**ADE (Gurulingappa et al., 2012)** is a biomedical relation extraction dataset, usually used by either classifying whether an adverse drug effect exists or whether none is expressed. Therefore, it only contains two relations. The entities in the dataset are not explicitly linked to an external KG, making separate linking necessary to introduce symbolic knowledge into the task.

**AGAC (Wang et al., 2019)** is a relation extraction dataset with a focus on relations between genes. It contains two different relations and the mentioned entities are matched to DrugBank entries (Knox et al., 2024a).

**BioRel (Xing et al., 2020)** is a large-scale biomedical relation extraction dataset encompassing over 500k sentences with overall 125 different relations. The advantage of this dataset is the availability of the concept unique identifier (CUI) for each entity occurring in the text. Concept unique identifiers are assigned to different concepts in the UMLS which is used throughout several different available ontologies. This allows the examination of the impact of various ontologies on the relation extraction task.

**ChemProt (Hailu et al., 2013)** is a relation extraction dataset with a focus on Chemical-Protein-interactions. The mentioned entities are not linked to a background KG or ontology.

**DDI (Herrero-Zazo et al., 2013)** is a dataset used to investigate the extraction of drug-drug-interactions. It contains entity annotations to the MedLine[1] and DrugBank ontologies. This simplifies research conducted in the realm of neuro-symbolic research. The only relation expressed in the dataset is whether there is an interaction.

**i2b2 (Uzuner et al., 2011)** is a relation extraction dataset with a focus on the relations between medical problems and their treatments. It contains eight different relations. It is unclear whether the mentioned entities are linked to a background KG or ontology.

1. https://www.nlm.nih.gov/medline/medline_overview.html

**Table 6.3**: All sentence relation extraction datasets. #R stands for the number of relations, #D for the number of documents, #S for the number of sentences, Linked denotes whether the entities are linked to a KG or ontology.

| Corpus | Desc. | #R | #D | #S | Text Type | Linked |
|---|---|---|---|---|---|---|
| ADE (Gurulingappa et al., 2012) | Biomedical Data (Adverse Drug Event) | 1 | 3000 | 20,967 | Biological | ✗ |
| AGAC (Wang et al., 2019) | Biomedical data (Genes) | 2 | - | 5,080 | Biological Abstracts | ✗ |
| BioRel (Xing et al., 2020) | Biomedical Data (Adverse Drug Event) | 125 | - | 533,560 | Biological | ✓ |
| ChemProt (Hailu et al., 2013) | BioCreative VI chemical-protein interaction Track 5 | 14 | - | 36,400 | Biological Abstracts | ✗ |
| DDI (Herrero-Zazo et al., 2013) | Drug-drug interactions from biomedical texts | 1 | - | 33,553 | Biological | ✓ |
| i2b2 (Uzuner et al., 2011) | Biomedical data (Medical problems and treatment relations) | 8 | 1,371 | - | Biological Abstracts | ✗ |
| CoNLL04 (Roth and Yih, 2004) | Newspaper | 5 | - | 1,441 | News | ✗ |
| NYT10 (Riedel et al., 2010a) | New York Times newspaper | 52 | - | 627,827 | News | ✓ |
| Re-TACRED (Stoica et al., 2021) | Revised version of TACRED | 40 | - | 91,467 | News | ✗ |
| FewRel 2.0 (Gao, Han, et al., 2019) | Wikipedia articles | 125 | - | 70,000 | Wiki | ✓ |

**Table 6.3:** All sentence relation extraction datasets (continued).

| Corpus | Desc. | #R | #D | #S | Text Type | Linked |
|---|---|---|---|---|---|---|
| Wikidata dataset (Sorokin and Gurevych, 2017) | Wikipedia articles | 353 | - | 732,393 | Wiki | ✓ |
| DialogRE (Yu et al., 2020) | Dialog-based datasets focusing on the Friends TV-Show | 37 | - | 7,900 | Dialog | ✗ |

**CoNLL04 (Roth and Yih, 2004)** is a relation extraction dataset annotated on the news domain. It is not directly connected to any background KG or ontology, making an additional linking step necessary. It contains only five different relations.

**NYT10 (Riedel et al., 2010b)** is a relation extraction dataset containing annotated New York Times articles. The entity mentions in the dataset were matched against entities in the Freebase KG. Furthermore, they use distant-supervision to assign Freebase (Bollacker et al., 2008c) relations to the entity pairs mentioned in the dataset. The authors note that the dataset suffers from false negatives which means that some actual expressed relations are not annotated as such. The Freebase identifiers of the entities are available for the dataset.

**Re-TACRED (Stoica et al., 2021)** is an improved version of the TACRED (Zhang et al., 2017) dataset. It resolves several labelling errors by analysing the original dataset and re-annotating it using crowd-sourcing. The entity mentions are not linked to a background KG or ontology.

**FewRel 2.0 (Gao, Han, et al., 2019)** is a relation extraction dataset with a focus on few-shot relation extraction. it is based on the first FewRel dataset (Han et al., 2018), which manually annotates a distantly-supervised subset of Wikipedia articles. Additionally, FewRel 2.0 provides a new test set, which consists of PubMed[2] articles aligned with the UMLS KG. This allows to study the generalization abilities of few-shot methods between different domains. It is a valuable resource as it contains examples from two sources aligned with two different KGs.

**Wikidata dataset (Sorokin and Gurevych, 2017)** is a distantly-supervised annotated dataset over Wikipedia using the Wikidata. It contains a large number of 353 different relations. Each entity mentioned is linked to Wikidata, facilitating the research on the impact of symbolic information.

**DialogRE (Yu et al., 2020)** is a human-annotated dialog-focused relation extraction dataset. It annotates the original script of the TV-show Friends with 36 relations and is available in Chinese and English. It is not connected to an ontology or KG.

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

In summary, the availability of background KG greatly influences the ease of integrating neuro-symbolic methods into relation extraction tasks. Biomedical and Wiki datasets often benefit from existing links to KG, such as UMLS for biomedical data and Wikidata for Wiki-based data. In contrast, the news datasets and the dialog dataset lack such direct connections, necessitating additional linking steps to incorporate symbolic information effectively.

### 6.3.2 Document Relation Extraction

Document-wide relation extraction (DocRE) emerged as a distinct task with datasets introduced between 2016 and 2017 (Li et al., 2016c; Peng et al., 2017; Quirk and Poon, 2017), primarily focusing on the biomedical domain or suffering from limited scope and quality. The release of the DocRED dataset (Yao et al., 2019) in 2019 marked a significant advancement by offering a large-scale, open-domain dataset with comprehensive manually annotated train, development, and test sets, thereby driving increased research into this domain of RE. An overview of all datasets used in the methods gathered in the previous section can be found in Table 6.4.

**SciERC (Luan et al., 2018)** is a document-wide Relation Extraction dataset containing 500 fully annotated paper abstracts. The papers encompass the topics of general AI, speech, machine learning and computer vision. The dataset is manually annotated and contains seven different relations. The entity mentions are not linked to a background KG.

**DocRED (Yao et al., 2019)** is large-scale document-wide RE dataset. It contains 96 different relations, making it one of the dataset with most-diverse relation set. It consists of a manually-annotated training, development and test set. Additionally, another large distantly-supervised dataset training dataset is offered as well. While the dataset was created by first linking the entity mentions to Wikipedia entities, the actual dataset does not provide the links. Furthermore, a later work (Tan et al., 2022) points out that the manually-annotated part contains a larger number of false-negatives.

**Re-DocRED (Tan et al., 2022)** revised the previously mentioned DocRED dataset to improve upon three problems: False Negatives, logical inconsistencies and coreferential errors. They accomplished this by iteratively improving the quality through a human-in-the-loop method. This leads to nearly two to three times as many triples annotated in the revised dataset. Similar to DocRED, Re-DocRED is not linked to an existing KG.

**CDR (Zhang, Lin, et al., 2021)** is a document-wide Relation Extraction dataset consisting of annotated PubMed abstracts created for the Fifth BioCreative Challenge Evaluation Workshop (Kim et al., 2015). It contains only a single relation, whether a chemical-disease interaction holds. Each entity is equipped with Medical Subject Headings (MeSH) (Lipscomb, 2000b) allowing the connection to background KGs or ontologies.

**GDA (Li et al., 2016a)** is a gene-disease association dataset containing distantly-labeled MEDLINE abstracts. It contains UniProt (Consortium, 2015) or CTD (Mattingly et al., 2003) gene ids for each gene and MeSH ids or CUIs for each mentioned disease allowing the study of the impact of an external KG. It again contains a single relation, whether an association holds or not.

**Table 6.4:** All document relation extraction datasets. #R stands for the number of relations, #D for the number of documents, #S for the number of sentences, Linked denotes whether the entities are linked to a KG or ontology.

| Corpus | Desc. | #R | #D | #S | Text Type | Linked |
|---|---|---|---|---|---|---|
| SciERC (Luan et al., 2018) | Abstract of computer science research papers | 7 | 500 | 2,700 | Paper abstract | ✗ |
| DocRED (Yao et al., 2019) | Wikipedia and Wikidata | 96 | 5,053 | - | Wiki | ✗ |
| Re-DocRED (Tan et al., 2022) | Revised DocRED by correcting false negatives | 96 | 5,053 | - | Wiki | ✗ |
| CDR (Zhang, Lin, et al., 2021) | PubMed biomedical documents | 1 | 1500 | - | Biomedical papers | ✓ |
| GDA (Li et al., 2016a) | MEDLINE abstracts | 1 | 30192 | - | Biomedical Abstracts | ✓ |
| DWIE (Zaporojets et al., 2021) | News Articles | 65 | 802 | - | News | ✓ |
| HacRED (Cheng, Liu, et al., 2021) | Documents from Chinese DBpedia | 26 | 9231 | - | Wiki | ✗ |
| HiEve (Glavas et al., 2014) | Event-Event Hierarchies | 2 | 100 | - | News articles | ✗ |
| MATRES (Ning et al., 2018) | Event-Event temporal relations | 4 | - | - | News articles | ✗ |
| Event StoryLine (Caselli and Vossen, 2017) | Event-Event temporal and causal relations | 4 | 258 | - | News articles | ✗ |
| APTERDOC (Wang, Xiong, et al., 2020) | Threat-intelligence field | 9 | 124 | 12,906 | Reports | ✗ |

**DWIE (Zaporojets et al., 2021)** is a corpus of annotated English "Deutsche Welle" [3] news articles. It contains annotations of entity mentions, entity links, coreferences and

3. https://www.dw.com/

relations. Due to the inclusion of entity linking, this makes this dataset very suitable to explore neuro-symbolic relation extraction. The annotated entity links point to Wikidata.

**HacRED (Cheng, Liu, et al., 2021)** is a dataset similar to DocRED focusing on Chinese documents while also constructing the dataset in such a way to include harder examples. Plain texts in the Chinese DBpedia (Auer et al., 2007b) are annotated using entity reocgnition and entity linking. This leads to a distantly-supervised dataset from which hard cases are sampled based on some heuristics. These are then manually-annotated and a model is trained to identify further hard cases in the full dataset. Finally, crowdsourcing was used to exhaustively annotate all identified hard cases. The actual dataset does not contain the original links to DBpedia.

**HiEve (Glavas et al., 2014)** is a Relation Extraction dataset focusing on events and their spatio-temporal relations. Of interest are here two relations, whether an event spatio-temporally contains another event or whether it they are coreferences of each other. The event mentions are not linked to any background KG.

**MATRES (Ning et al., 2018)** is a temporal Relation Extraction dataset. It contains four relations: "before", "after", "vague" and "equal" and focuses on news articles. Events are expressed verb instead of a noun. The events are not linked to a background KG or ontology.

**Event StoryLine (Caselli and Vossen, 2017)** is a Relation Extraction dataset with a focus temporal or causal relations between events. Four different relations “contains", “before", “after" and “overlap" are annotated. The focus is on news articles. The mentioned events are not linked to a background KG or ontology.

**APTER-DOC (Wang, Xiong, et al., 2020)** is a document-wide Relation Extraction dataset in the threat-intelligence field. It consists of annotated advanced persistent threats reports. Nine different relations are annotated specific to the threat-intelligence field. The annotated entities are not connected to a background Knowledge Graph or ontology.

Many datasets, such as SciERC and Re-DocRED, despite their substantial annotations, do not link entities to background KGs, necessitating additional linking efforts. Conversely, datasets like CDR and GDA benefit from built-in connections to biomedical ontologies, simplifying the integration of external knowledge. Datasets from other domains, including news articles (MATRES, Event StoryLine) and threat intelligence (APTER-DOC), lack such connections as well, reflecting the ongoing challenge of integrating symbolic knowledge across different fields.

### 6.3.3 Cross-document Relation Extraction

Cross-document relation extraction datasets are a fairly recent introduction to the problem area. Here, one is confronted with multiple documents each containing multiple sentences with again multiple entities inside. Relations can hold between entities inside the documents but also between them. Constructing datasets for this relation extraction category is more complex, hence why much fewer datasets are published for that.

See Table 6.5 for an overview of all the cross-document relation extraction datasets.

**CodRED (Yao et al., 2021)** is a cross-document relation extraction dataset annotated over Wikipedia while using Wikidata. It was annotated in three stages: 1) distant supervision, 2) manual annotation and 3) generating hard negative examples. It contains a large number of 276 different relations. As each entity mention is linked to Wikidata, studying the impact of symbolic data is significantly simplified.

**Yahoo/Reuters/CNN (Zhang, Lyu, et al., 2023)** are cross-document relation extraction datasets focusing on causal relations between events. It contains only a single relation. The mentioned events are not linked to a background KG or ontology.

CodRED stands out with its comprehensive approach, utilizing Wikidata to simplify the integration of symbolic data and covering an extensive range of 276 relations. In contrast, the Yahoo/Reuters/CNN datasets focus on causal relations between events but lack connections to background KGs, posing additional challenges for incorporating external knowledge.

**Table 6.5:** All cross-document relation extraction datasets. #R stands for the number of relations, #D for the number of documents, #S for the number of sentences, Linked denotes whether the entities are linked to a KG or ontology.

| Corpus | Desc. | #R | #D | #S | Text Type | Linked |
|---|---|---|---|---|---|---|
| CodRED (Yao et al., 2021) | Wikipedia articles | 276 | - | 30,504 [4] | Paper abstract | ✓ |
| Yahoo (Zhang, Lyu, et al., 2023) | News Articles | 1 | 1046 | - | News | ✗ |
| Reuters (Zhang, Lyu, et al., 2023) | News Articles | 1 | 1212 | - | News | ✗ |
| CNN (Zhang, Lyu, et al., 2023) | News Articles | 1 | 1190 | - | News | ✗ |

## 6.4 Challenges

Despite the numerous methods developed in neural-symbolic relation extraction, several challenges persist in this domain. One significant issue is the sheer volume of symbolic information, which, if fully utilized in neural models, might introduce bias or noise. There has been limited research on identifying which sources of symbolic information are most valuable for information extraction and on strategies to mitigate bias. Another challenge is aligning symbolic information with neural models. Neural models typically process text input to generate high-dimensional embeddings, whereas symbolic information can take various forms, including text and rules, with dimensions ranging from low to high. Some approaches use linear transformation or concatenation to align the dimensionality of the representations, while others integrate rules into the neural model training process, such as through the loss function or constrained generation. Successfully integrating these diverse types of information into a cohesive model remains a complex and ongoing task.

As for the datasets, it is evident that many datasets do not contain links to an existing KG or ontology. While this often is the case for datasets in the biomedical domain and to some extent for those which were initially annotated by distant-supervision, the majority are lacking such links. This is problematic due to several reasons. First, the linking needs to be done by an entity linker. However, the performance of different entity linkers vary vastly. Therefore, the actual linked entities might be different from method to method, introducing a variability in the starting conditions. Therefore, not only the developed method but also the quality of the entity linker has an impact on the results. While summarising the presented methods, it becomes evident that many papers did not state how the entity links were gathered. Second, the linking might be done against different versions of a KG. This again introduces more uncertainty in regard to differences between reported results. In the future, to better study the interaction between symbolic and neural relation extractions method, constructing datasets containing such links is vital. Furthermore, specifying against which KG the entities were annotated is very important as well.

## 6.5 Conclusion

In summary, RE is a pivotal task in NLP that focuses on identifying and classifying relationships between entities within text. The versatility of RE spans various types of text, including in-sentence, in-document, and cross-document contexts, making it a critical component in constructing KGs and answering complex queries. The advancement of machine learning models, particularly Pre-trained Language Models, has significantly enhanced the performance of RE by enabling efficient extraction of relations from unstructured text. However, these models often grapple with issues of robustness and generalization, especially when dealing with unseen entities and domain-specific texts.

To address these challenges, the integration of symbolic information, such as KGs and logical rules, has emerged as a promising approach. Neuro-symbolic RE methods aim to combine the interpretability and reasoning capabilities of symbolic systems with the flexibility and scalability of neural networks. Despite their potential, these methods face several challenges, including the effective alignment of symbolic and neural representations and the mitigation of bias introduced by symbolic information.

The current landscape of RE research highlights the necessity for improved datasets that include explicit links to existing KGs or ontologies. Such datasets would enable more consistent and comparable evaluations of RE methods. Additionally, further research is

needed to identify the most valuable sources of symbolic information and develop robust strategies for integrating them into neural models.

Looking forward, the continued exploration of neuro-symbolic approaches holds promise for developing more robust, interpretable, and generalizable RE models. By addressing the existing challenges and leveraging both neural and symbolic techniques, future work can significantly advance the state of relation extraction, ultimately contributing to more sophisticated and accurate NLP systems.

## Abbreviations used in this chapter

**AMR** Abstract Meaning Representation

**ASP** Answer Set Programming

**ASPER** ASP-enhanced Entity-Relation extraction

**BERT** Bidirectional Encoder Representations from Transformers

**CNN** Convolutional Neural Network

**DocRE** Document-level Relation Extraction

**DRK** Discriminative Rule-based Knowledge

**GAL** Global Adaptive Loss

**GNN** Graph Neural Network

**GM_GEN** Graph-based Model Generation

**ILP** Integer Linear Programming

**KG** Knowledge Graph

**KICE** Knowledge-evolving Framework by Iterative Consolidation and Expansion

**LSTM** Long Short-term Memory

**MILR** Mining and Injecting Logical Rules

**NER** Named Entity Recognition

**NLP** Natural Language Processing

**PLM** Pre-trained Language Model

**RE** Relation Extraction

**RECON** Relation Extraction using Knowledge Graph Context in a Graph Neural Network

**SKAMRR** Sememe Knowledge-enhanced Abstract Meaning Representation and Reasoning

**GAN** Graph Attention Network

**UMLS** Unified Medical Language System

# 7

# Bridging the Gap: Generating a Comprehensive Biomedical Knowledge Graph Question Answering Dataset

## Bibliographic Information

Xi Yan, Patrick Westphal, Jan Seliger, and Ricardo Usbeck. 2024. Bridging the Gap: Generating a Comprehensive Biomedical Knowledge Graph Question Answering Dataset. In *ECAI 2024,* 1198–1205. IOS Press. (Cited on pages 4, 6, 10 sq., 114).

## Abstract

Despite the plethora of resources such as large-scale corpora and manually curated Knowledge Graphs (KGs), the ability to perform reasoning with natural language inputs over biomedical graphs remains challenging due to insufficient training data. We propose a novel method for automatically constructing a Biomedical Knowledge Graph Question Answering (BioKGQA) dataset sourced from PrimeKG, the largest precision medicine-oriented KG. In total, we create 85,368 question-answer pairs along with their respective SPARQL queries. Our approach generates a diverse array of contextually relevant questions covering a wide spectrum of biomedical concepts and levels of complexity. We evaluate our method based on automatic metrics alongside manual annotations. We establish novel standards tailored for KGQA systems to highlight the linguistic correctness and semantical faithfulness of the generated questions based on extracted KG facts. The compiled dataset – PrimeKGQA – serves as a valuable benchmarking resource for advancing knowledge-driven biomedical research and evaluating KGQA systems.

## 7.1 Introduction

Biomedical KGs offer a powerful framework for organizing and semantically linking heterogeneous biomedical data, enabling comprehensive exploration and analysis of the underlying biological phenomena. However, the effective use of these KGs for

real-world applications, such as question answering (QA) systems (Lin et al., 2024), necessitates the availability of high-quality training datasets tailored to the intricacies of the biomedical domain.

There are three main challenges in developing large-scale bioKGQA datasets:

(i) The costs of hiring annotators with professional backgrounds are usually high. (ii) There are numerous biomedical KGs not aligned with common (but evolving) ontologies. For instance, DisGeNET (Piñero et al., 2016) is not fully aligned with common ontologies such as the Human Phenotype Ontology (HPO) (Brookes and Robinson, 2015) or the Disease Ontology (DO) (Schriml et al., 2012). (iii) Most existing automatic question-generation algorithms require (extensive) training data. As a result, there are only three bioKGQA datasets based on different KGs, with a total amount of 90 question-answer pairs.

Yet, recent advancements in pre-trained language models (PLMs) can tackle the above challenges by guiding the generation process with a small amount of annotated data without a costly training process, resolving the challenges (i) and (iii), by introducing prompt-based few-shot learning (Wang, Yao, et al., 2020). With a small number of samples ranging from one to twenty per class (Cao et al., 2020), PLMs generally demonstrate a strong capacity to generalize over unseen data.

Supporting the creation of a large-scale KG-based QA dataset in the biomedical domain, recently, a large database was built that integrates 20 high-quality and most cited databases in precision medicine,[1] PrimeKG (Chandak et al., 2023b). It focuses on ten major biological scales, including disease-associated protein perturbations, biological processes and pathways, anatomical and phenotypic scales, and the entire range of approved drugs with their therapeutic action, considerably expanding previous efforts in disease-rooted databases. The underlying KGs of the existing bioKGQA datasets can be mapped to PrimeKG, providing a chance to integrate the other bioKGQA datasets, too.

Thus, we built a large-scale KGQA dataset on top of PrimeKG utilizing few-shot learning on PLMs. First, we transform the PrimeKG database into an RDF-based KG and set up a SPARQL endpoint. Next, we follow a general triple-to-question pipeline: (i) We sample subgraphs of specific structures from PrimeKG as reasoning paths of the question-answer pair. (ii) The answers are selected using a specific anchor selection strategy. (iii) The reasoning paths and the updated answers are linearized and sent to the PLM as parts of the input prompt to generate the underlying questions. (iv) We test several PLMs and validate them on our own as well as three well-known QA datasets. Hence, this dataset can serve as a vital resource for advancing biomedical research, enabling the development and evaluation of QA systems that efficiently retrieve relevant biomedical knowledge.

Therefore, the contribution of this work is three-fold: (i) We present the first large-scale biomedical KGQA dataset. PrimeKGQA is factor 1000 larger than the second-largest KGQA dataset. (ii) We develop a novel framework to generate questions based on the KG triples. (iii) We initiate a novel anchored answer selection strategy. The developed model and dataset are publicly available on GitHub.[2]

## 7.2 Related work

In this section, we review the existing bioKGQA datasets, highlighting the need for new resources. We also discuss previous work on generating questions based on triples and the metrics used to evaluate the quality of these generated questions.

1. A data and KG-centered approach to disease diagnosis and treatment that accounts for the variability in genetics, environment, and lifestyle across individuals

2. https://github.com/xixi019/primeKGQG

**Table 7.1**: Statistics of the existing BioKGQA datasets.

| | QALD-4 | Bgee-QA | OMA-QA | CORDIS-QA |
|---|---|---|---|---|
| # q-a pairs | 50 | 20 | 10 | 30 |
| Underlying KG | DrugBank | Bgee | OMA | CORDIS |

### 7.2.1 Existing dataset

Two major problems of existing bioKGQA datasets are, that they are small in size and that they are built upon different KGs. Details are listed in Table 7.1. Four existing public datasets BgeeQA (Sima, Mendes de Farias, et al., 2021), OMA QA (Sima, Mendes de Farias, et al., 2021), CORDIS-QA (Sima, Mendes de Farias, et al., 2021) and Task 2 of QALD-4 (Unger et al., 2014a) are dependent on KGs of various sub-domains of biomedicine. For instance, Bgee (Bastian et al., 2020) contains information about genes and in which parts of the body (anatomical entity) a gene is expressed or absent, while DrugBank (Knox et al., 2024b) is a pharmaceutical database. Theoretically, those KGs could be mapped and grouped into a bigger KG (by ontology or ID mapping) so that it serves as the underlying KG for all KGQA datasets, in order to fully utilize the training data and to enhance model generalizability. Yet, no work has been done in this direction.

### 7.2.2 BioKG

There is a growing focus on constructing large-scale biomedical KGs by integrating resources, like BioKG (Zhang, Sui, et al., 2023), Hetionet (Himmelstein et al., 2017), OREGANO (Boudin et al., 2023), and PrimeKG (Chandak et al., 2023b), among others. Of these, PrimeKG stands out as one of the largest open-source biomedical knowledge graphs, incorporating the most widely used datasets. Unlike BioKG, Hetionet, and OREGANO, PrimeKG includes more up-to-date resources such as Bgee(Bastian et al., 2020), Drug Central(Ursu et al., 2017), and Uberon,[3] making it the most diverse dataset available to date.

The original artifacts of the PrimeKG project (Chandak et al., 2023b) are publicly available.[4] They comprise the collection of build and pre-processing scripts to compile the main PrimeKG dataset from the respective sources, and the final data files for download. The collected and integrated information stems from a variety of prominent sources for biomedical data, namely Bgee,[5] Comparative Toxicogenomics Database(CTD),[6] DisGeNET,[7]DrugBank,[8] Drug Central,[9] Entrez Gene,[10] Gene Ontology (GO),[11] Human

3. https://github.com/obophenotype/uberon
4. https://zitniklab.hms.harvard.edu/projects/PrimeKG/
5. https://www.bgee.org/
6. https://ctdbase.org
7. https://www.disgenet.org
8. https://www.drugbank.com/
9. https://drugcentral.org/
10. https://www.ncbi.nlm.nih.gov/gene
11. https://geneontology.org/

Phenotype Ontology (HPO),[12] Mondo Disease Ontology,[13] Reactome,[14] SIDER,[15] Uberon, and UMLS.[16]. The data is often in tabular form, except for the ontologies mentioned. The PrimeKG build scripts then generate an integrated view on the input sources with the core abstraction of having *nodes*, i.e. resources with certain properties and provenance information, and *edges*, which are typed relations between the nodes. This data constitutes the main PrimeKG, yet, it is provided in CSV format, not following the LinkedData principles.[17]

### 7.2.3 Triple-to-Question Generation

Most work in triple-to-question generation follows the supervised scheme of fitting and inferring, which means they fine-tune or pre-train a model based on a large-scale dataset and evaluate the trained model on the test set. GAIN (Shu and Yu, 2024b) fine-tunes a T5 model (Raffel et al., 2020a), to convert two node triples from freebase to natural questions. Fock (2022) (Fock, 2022) fits triples and quadruples, extracted from a temporal KG, YAGO11k (a subset of YAGO3 (Mahdisoltani et al., 2015)) into pre-defined question templates. JointGT (Ke et al., 2021) adds structural information of the input triple to the transformer layer and separates text generation into three sub-tasks to further pre-train the transformer models. Han and Gardent (2023) (Han and Gardent, 2023) designed a multitask model capable of generating questions from both textual and graphical inputs. They validated the generated questions by testing whether the corresponding answers from open-domain QA models based on the questions match the gold standard answer. Han, Ferreira, and Gardent (2022) (Han et al., 2022) pre-trained BART (Lewis et al., 2019) for the triple-to-question task, incorporating two additional pieces of information: question type and property information from the underlying knowledge graph. Kumar et al. (2019) (Kumar et al., 2019) feed the textualized graph to an encoder and decoder-based transformer with embedded answers and difficulty estimation to generate complex questions. Cheng et al. (2021) (Cheng, Li, et al., 2021) fine-tune GPT-2 on a self-constructive dataset for guiding the model to rewrite the simple questions into difficult questions.

Note that Rangel et al. (2024) (Rangel et al., 2024) utilize an automatic process to generate a large-scale biomedical KGQA dataset over Bgee. Yet, it is neither clear how the corresponding questions are generated, nor could we find the dataset under the given URL in their brief report.

All these supervised learning approaches use large training data and, thus, are not applicable in our case, where training data is hard to obtain due to high cost. Additionally, there are no existing triple-to-questions models in the biomedical domain and the available bioKGQA datasets are severely undersized for supervised training. Therefore, we decide to explore the power of few-shot learning using PLMs to synthesize a sufficiently large-scale dataset that is anchored in explicit domain facts.

### 7.2.4 Evaluation Metrics

There are two assessment strategies for examining quality and suitability: automatic evaluation and human evaluation (Osuji et al., 2024). Automatic evaluation metrics can be categorized into *n*-gram metrics, task-specific metrics, and information extraction

12. https://hpo.jax.org/app/
13. https://mondo.monarchinitiative.org/
14. https://reactome.org
15. http://sideeffects.embl.de/
16. https://www.nlm.nih.gov/research/umls/index.html
17. https://www.w3.org/DesignIssues/LinkedData.html

metrics (Osuji et al., 2024). 15 metrics are adopted in triple-to-question research, with the top three being BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004), which are established *n*-gram-based metrics for evaluating text generation quality. Thus, we will use these three to evaluate our approach.

As for task-specific metrics, embeddings or PLM-based metrics are used for enhancing the semantic alignment between text and the reference (Nedelchev et al., 2020). This is compliant with our use case since we want semantically aligned and linguistically varied question-answer pairs for the generalization capacity of the QA systems. Therefore, we also adopt BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) in the evaluation, which are the popular metrics under this category. Note that BLEURT is a learned metric that measures both fluency and the correspondence of the generated question to the reference in terms of the semantic meaning. It is a BERT model pre-trained on a synthetic sentence pair dataset and then fine-tuned based on public human ratings. The range of BLEURT is between -2 and around 1. The closer the score is to 1, the better the quality of the prediction is.

The information extraction metrics focus on content selection of the systems, often when multiple records are used as prediction sequences. Most of the question generation datasets do not contain multiple references. Therefore, this strategy is discarded.

Human evaluation is usually included since it is more precise in terms of semantic coherence, mismatch of the numerical values, and complexity. However, there are no established or unified standards for (costly) manual evaluation and different studies use various wording for a diverse range of aspects. According to a review paper in natural language generation (Osuji et al., 2024), Fluency, Grammaticality, Correctness, Adequacy, Coherence, Faithfulness, Naturalness, Conciseness, and Similarity are the top 10 indexes used in NLG publications, with fluency being the most used standard. Based on those metrics, we conclude three with consideration to our dataset evaluation: Consistency, Grammaticality and Coverage. This is explained in Section 7.4.4.

## 7.3 Method

Based on PrimeKG, we aim to facilitate a generalizable approach for generating comprehensive KGQA datasets. Additionally, we aim to address energy efficiency concerns in the age of PLMs, which have significant requirements for training resources, leading to considerable carbon dioxide emissions. To achieve these goals, we propose a training-free and knowledge graph-independent method. On top, this method can be easily adapted to any knowledge graph of the user's choice, enhancing its flexibility and usability.

An illustration of our pipeline is shown in Figure 7.1. We first convert PrimeKG to an RDF KG which can be accessed via SPARQL. Then this SPARQL endpoint is used to extract the 2- to 4-node-subgraphs based on network motifs (Milo et al., 2002). The subgraphs/triples are then linearized, i.e. transformed from formal KG triples into sentences, as part of the input for the PLM for generating the questions. On the other hand, based on the generated triples, we design SPARQL templates which take the entity and relations from each question to form a corresponding SPARQL query. And this query is then run against the endpoint to extract correct answers. For each subgraph, we collect the generated questions, SPARQL queries, and the answers as KGQA pairs.

### 7.3.1 Building an RDF KG for PrimeKG

The main motivation for developing an RDF-based KG is to build the downstream AI tasks (e.g., question answering, search engine, etc.) on established and standardized protocols

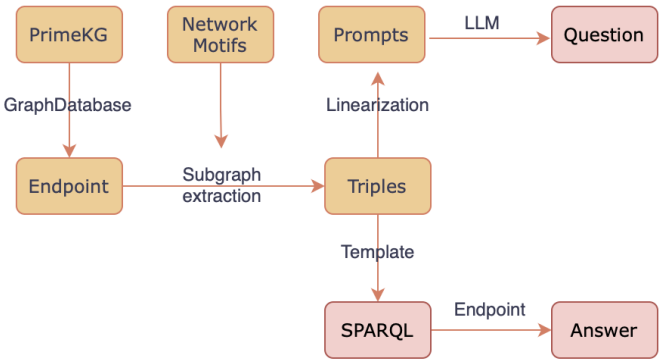



**Figure 7.1:** Our pipeline for automatic generation of PrimeKGQA. The pink blocks are the composing elements of the dataset, i.e., question in natural language, SPARQL query, and correct answer from the KG.

and formats and to be able to ease further integration steps with other RDF-based data sources. To represent the original PrimeKG resources, as well as the relations between them in a unique way, a generic IRI scheme was applied, with the node IDs, resp. relation type strings, becoming the local parts of the IRIs. The original PrimeKG CSV files were then translated to RDF straightforwardly with relations between resources becoming object properties, any selected additional resource features becoming literals assigned to resources via datatype properties, and the resource types are assigned by means of `rdf:type` triples. In the current version, the relation types are not translated to RDF, as this would require RDF reifications which were considered too costly in terms of the storage overhead. We filtered out any MONDO group resources from the original PrimeKG dataset as they are essentially a *collection* of actual MONDO classes, which represent a new concept without an unique identifier. These resources were removed as they do not fit our downstream processing workflow. The generated RDF triples were then loaded into a triple store to make them publicly accessible via SPARQL.[18] The number of triples in the triple store amounts to 8,580,967.

### 7.3.2 Subgraph Generation

To generate prompts which in turn should generate questions from triples, we need to extract subgraphs from PrimeKG. We sample triples with the numbers of nodes ranging from 2 to 4 using triple templates, namely, network motifs (Milo et al., 2002). We focus on 2- to 4-node subgraphs for the following two reasons. (i) A comprehensive KGQA dataset typically consists of both simple and complex questions to enhance the model's ability to generalize across various complexities. The categorization of questions into *simple* and *complex* is based on the number of hops. Questions involving 1-hop patterns are considered simple, whereas those involving patterns with two or more hops are deemed complex, corresponding to 2-nodes and 3-nodes or more in the triples, respectively. (ii) We analyze a real-world biomedical QA dataset, BioASQ (Tsatsaronis et al., 2012), which comprises questions with the most common number of entities ranging between two to four per question. Since the BioASQ dataset lacks named entity recognition (NER) annotations, and existing biomedical NER tools vary in granularity, we aim to compare the number of entities detected by tools designed for both fine and coarse granularities. Therefore, our approach begins by running two open-source biomedical

18. http://sems-coypu-4.informatik.uni-hamburg.de:8890/sparql/

named entity tagging tools with different granularities for the scientific and clinical subdomain in biomedicine.[19,20] The results indicate that most manually curated questions contain around two to four entities. Consequently, we utilize subgraphs with two, three, and four nodes as the underlying triples.

Sampling based on network motifs allows us to include diverse and complicated reasoning paths. Network motifs are well-defined network structures used across many fields of science, such as the World Wide Web, networks from biochemistry, neurobiology, ecology, and engineering (Milo et al., 2002). Motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks.

As for 2-node-subgraphs, there is only one pattern, as can be seen in Figure 7.2, since we do not consider cyclic graphs. In terms of 3-node-subgraphs, according to Milo et al. (2022) (Milo et al., 2002), there are 13 types, as shown in Figure 7.2. Certain types of them contain fully connected graphs. The extracted triples in this structure are, despite being meaningful subgraphs, hard to convert into a valuable question in the later step. For instance, type 5 (N3_5 in Figure 7.2) is a graph $G$, formally written as $G = \langle V, E \rangle$, with $V$ denoting a set of nodes $x_1$, $x_2$, $x_3$, $E$ denoting a set of edges $\{\langle x_1, x_2 \rangle, \langle x_3, x_2 \rangle, \langle x_3, x_1 \rangle \mid x_1 \neq x_2, x_1 \neq x_3, x_2 \neq x_3\}$. Under such a triangular structure, we observe that inevitably one edge would not be connected directly to a certain node in the graph. For instance, $\langle x_3, x_1 \rangle$ is not related to node $x_1$. In this case, the edge is not needed to generate a reasoning path to the question node, i.e., the information needed to generate a path is the same as contained in the fourth structure of 3-node-subgraph (denoted as N3_4 in Figure 7.2). And this shall hold for all the triangular-shaped subgraph patterns. Therefore, we decide to discard seven motifs, leaving only six as plausible motifs, due to the occurrence of the pattern. Also, we remove the subgraphs with duplicate edges.

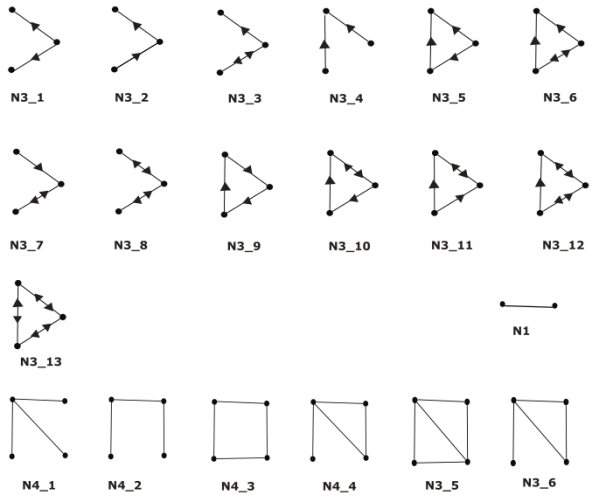



**Figure 7.2:** All types of network motifs for graphs with node numbers from two to four. N3_1 stands for "node number 3 subgraph type 1". Note that for 3-node-subgraphs, we discard N3_5, N3_6, N3_9, N3_10, N3_11, N3_12 and N3_13.

Regarding 4-node motifs, the number of possible motifs explodes, factoring the count of the possible 3-node-subgraph motifs. The basic structures of the motifs are listed in Figure 7.2, according to (Al-Thaedan and Carvalho, 2019). However, according to our statistics real-world QA datasets, 4-node questions are not the majority of the whole

19. https://huggingface.co/d4data/biomedical-ner-all
20. https://huggingface.co/Clinical-AI-Apollo/Medical-NER

corpus, taking up 0.04% and 0.19% according to different NER tagging tools. Consequently, we only sample motifs 1 (N4_1) and 2 (N4_2) amongst the set of 4-node motifs listed in Figure 7.2 and abandon other types for the same reason as explained for 3-node-motifs.

Based on those motifs, we generate SPARQL queries to extract subgraphs from the PrimeKG. An example SPARQL query based on the type 1 motif for 3-node-subgraph is illustrated in Figure 7.3:

```
SELECT
DISTINCT ?subj ?prop1 ?obj1 ?prop2 ?obj2
WHERE {{
        ?subj ?prop1 ?obj1.
        ?subj ?prop2 ?obj2.
        }}
```

**Figure 7.3:** An example SPARQL query.

Note, we also make use of the `BIND()`, `RANDOM()` and `FILTER()` functions to extract more diverse subgraphs and exclude terminological information. An example SPARQL query is contained in our project repository on GitHub.[21] In total, we have nine motifs from which we extract 90,000 subgraphs.

#### Answer selection

Most of the work on triple-to-question generation or automatic KGQA construction chooses the tail entity as the answer. In a KG, a triple is formed by a relation $r$ connecting two entities $h$ and $t$. It is noted as $\langle h, r, t \rangle$, where $h$ is the *head* entity, and $t$ is the *tail* entity with $h, t \in V$ and $r \in E$. Hence, the generated dataset is usually homogeneous and the reasoning path is easier to learn for the model. Besides, most of the answers are only nodes, sofar the edges are ignored in the subgraphs, which are also important in composing the subgraphs and reasoning path. Therefore, we decide on an anchor answer selection strategy. First, we include both edges and nodes. Second, we choose the edges or nodes with the highest connectivity in the graph. Specifically, we locate the node with at least two outgoing/incoming edges. Next, the chosen nodes and the edges are combined into a candidate answer list. We randomly sample one entry from the candidates list to be the designated answer. Based on the subgraph and the answer, a SPARQL query is formatted with the answer masked. This SPARQL query is used for SPARQL validation.

#### SPARQL validation

We utilize the SPARQL queries generated in the last step for extracting the answer from the PrimeKG. When the answer from the endpoint differs from the one extracted from the subgraph, we update the answer by the answers extracted from the SPARQL endpoint.

### 7.3.3 Question Generation

The input of this triple-to-question task is a subgraph $G_i = \langle V_i, E_i \rangle$ and an answer $a_i \in G_i$. We evaluate various prompts and incrementally add settings such as one-shot, few-shot, and chain-of-thought (CoT) (Wei et al., 2022). We collect the output and evaluate a few samples manually to choose the best settings. Upon initial exploration and experimentation, we narrow down the set of experiments. Our experiments show that enclosing edges and nodes in brackets (e.g., [nasal cavity epithelium] [presents the

21. https://github.com/xixi019/primeKGQG/blob/main/primekg/appendices.pdf

expression] [Anterior segment of eye aplasia]) in CoT prompts yields the most faithful and grammatically fluent results. An example of our final prompt setting can be found in the project repository on GitHub. We set up three metrics to check the question quality: *Grammaticality*, *Coverage*, and *Consistency*. They are explained in Section 7.4.4. We test ChatGPT,[22] Mistral,[23] (Jiang, Sablayrolles, Mensch, Bamford, Chaplot, Casas, et al., 2023) and a LLaMA-based medicine-PLM (med-PLM) (Cheng et al., 2023).[24] We keep the initial prompt as the baseline and the best group of prompt settings for generating the actual dataset.

## 7.4 Evaluation

We evaluate our model and the generated dataset using both automatic and manual metrics. We use Mistral (Jiang, Sablayrolles, Mensch, Bamford, Chaplot, Casas, et al., 2023) without few-shot samples, bracket, and CoT in the prompts as the baseline. Similar to the past work on question generation, and due to a lack of testing bioKGQA datasets, we examine the model on KGQA datasets in the encyclopedic domain. For manual evaluation, we sample from the generated dataset and ask three experts (two biologists and one IT expert) to annotate the samples. The sample size is 90 (10 samples from each type of network motif) due to the high annotation costs.

### 7.4.1 Dataset Description

We use SQB (Wu et al., 2019), LC-QuAD (Dubey et al., 2019c), and WebquestionSP (WebQSP) (Yih et al., 2016b) as the testing dataset since they cover simple and complex triples. Each dataset includes a natural language question, an answer, and a query based on Freebase or DBpedia. The questions from SQB and LC-QuAD are created by filling the entity and relations in question templates, while WebQSP is generated by annotating the natural language questions against a KG.

To obtain the input subgraph, we first process the SPARQL queries or the inference path (i.e., reasoning path) in the dataset, including the removal of namespace prefixes, converting IRIs to corresponding entity names, linearizing the relation representation, etc. Detailed statistics of the datasets can be accessed in Table 7.2.

**Table 7.2:** Statistics of the evaluation. *Simple* and *Complex* stand for simple and complex questions in the dataset. *Paraphrase* indicates whether the edges and nodes in the triple are replaced by the synonyms in the generated question, which makes it harder for the model to generate a similar question based on *n*-gram metrics. We use the test/validation sets for evaluation.

| | Simple | Complex | Eval_size | Rephrase | Template |
|---|---|---|---|---|---|
| SQB | ✓ | ✗ | 21,483 | ✓ | ✓ |
| LC-QuAD | ✓ | ✓ | 2,000 | ✓ | ✓ |
| WebQSP | ✓ | ✓ | 1,639 | ✓ | ✗ |

22. https://platform.openai.com/docs/models/gpt-3-5-turbo
23. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
24. https://huggingface.co/AdaptLLM/medicine-LLM

**Table 7.3:** Evaluation result of different methods on SQB.

| Result | BLEU | ROUGE | METEOR | BS | BLEURT |
|---|---|---|---|---|---|
| GAIN | **0.3060** | **0.5927** | **0.5361** | **0.8709** | **-0.2934** |
| med-PLM. | 0.0750 | 0.4507 | 0.4744 | 0.8208 | -0.4904 |
| Mistral | 0.0520 | 0.4020 | 0.4502 | 0.8140 | -0.4423 |
| baseline | 0.0198 | 0.2501 | 0.3340 | 0.7674 | -0.9215 |

**Table 7.4:** Evaluation result of different methods on WebquestionSP.

| Result | BLEU | ROUGE | METEOR | BS | BLEURT |
|---|---|---|---|---|---|
| GAIN | **0.0585** | **0.4790** | 0.4110 | 0.8320 | -0.3213 |
| med-PLM | 0.0487 | 0.4471 | **0.4929** | **0.8785** | -0.1675 |
| Mistral | 0.0151 | 0.3710 | 0.4528 | 0.8460 | **1.0370** |
| baseline | 0.0108 | 0.3046 | 0.4201 | 0.8075 | -0.5606 |

### 7.4.2 Automatic Evaluation

Our methods are compared to GAIN (Shu and Yu, 2024a), which is tested on the same datasets (SQB, LC-QuAD, and WebQSP). We intended to compare our approach to DiffQG, however, we were unable to obtain the necessary resources and models from the authors. The evaluation results show that our method significantly improves upon the baseline across all three datasets. The settings used in our best-performing model outperform GAIN by a large margin on most evaluation metrics across almost all datasets, except for SQB. This difference can primarily be attributed to SQB being a dataset comprised solely of 2-node triple-to-question pairs, with shorter question text lengths. In contrast, our models tend to utilize all information from the triple and generate longer questions that are semantically more faithful to the original questions. For datasets with more complex questions (LC-QuAD and WebQuestionsSP), our models demonstrate better performance. This would be further explained in Section 7.4.5.

**Automatic Metrics** As mentioned in Section 7.2.4 We utilize both *n*-gram and PLM-based metrics. *n*-gram based metrics include BLEU, METEOR, and ROUGE. As for PLM-based metrics, we utilize BERTSCORE(BS) and BLEURT.

### 7.4.3 Automatic Evaluation Result and Analysis

The performance of GAIN and the performance of our methods over three different PLMs are listed in Table 7.3, Table 7.4 and Table 7.5. The best performances are marked in bold. We also explain why on SQB our methods are worse than GAIN.

Note that for the dataset SQB, we turn the relational facts from Freebase into a triple format to include structural and domain knowledge, which shows improvement in several metrics than non-relational included counterparts.

Overall, med-PLM has the best performance overall metrics across different datasets. On SQB, med-PLM has a similar score compared to the best GAIN model. In terms

**Table 7.5:** Evaluation result of different methods on LC-QuAD.

| Result | BLEU | ROUGE | METEOR | BS | BLEURT |
|---|---|---|---|---|---|
| GAIN | 0.0692 | 0.3575 | 0.2396 | 0.8008 | -0.7671 |
| med-PLM | **0.1649** | **0.4151** | **0.4314** | **0.8399** | **-0.4550** |
| Mistral | 0.0822 | 0.3440 | 0.3651 | 0.8249 | -0.6351 |
| baseline | 0.0590 | 0.2756 | 0.3309 | 0.7601 | -0.8353 |

of WebquestionSP, med-PLM has the highest score on METEOR and BS, with a small difference (less than 0.15) on BLEU, ROUGE, and BLEURT. As for LC-QuAD, med-PLM performs the best across all metrics with a big margin.

Upon *n*-gram-based metrics, for all the evaluated models there is still room for improvement. This might be due to the paraphrasing applied in the original dataset generation process. For instance, in SQB, the question *"What is a hong kong netflix film?"* is related to triple: *"hong kong"* (head entity), *"media_common.netflix_genre.titles"* (relation), *"Saviour of the Soul"* (tail entity and answer). Note that this pair is problematic since the tail entity is not the only node corresponding to the reasoning path.[25] To predict the mention of *"film"* from the relation "media_common.netflix_genre.titles" seems to be a daunting task for PLMs in general, since different subchunks of the relation can all be suitable for generating such a question. The predictions of the PLM for this triple vary from:

- *"what is the title of hong kong"* (GAIN)
- *"Which Hong Kong action film from the 1990s was particularly popular and went by the name Saviour of the Soul?"* (baseline)
- *"What is an example of a popular Hong Kong TVB drama series from the 1990s?"* (Mistral)
- *"What is the title of the movie that is available on Netflix and is set in Hong Kong?"* (med-PLM)
- *"what is a hong kong netflix film?"* (reference)

As humans, we can discern that med-PLM preserves the closest semantic meaning to the original sentence. However, because *n*-gram-based metrics assess similarity on the string-based gram level and disregard contextual meaning, the scores are relatively low.

This example also explains why our models are worse than GAIN on SQB. From our observation of the generated questions, the PLMs we use tend to produce lengthy questions, which can contribute to low scores on BLEU, METEOR, and ROUGE, since on SQB the question spans are relatively short. Meanwhile, GAIN is optimized for SQB, resulting in shorter generated text. Besides, GAIN effectively learns the template after the fine-tuning, resulting in better performance on SQB. Mistral and med-PLM, on the other hand, have not been specifically optimized for simple or complex questions and strive to retain all the information provided to preserve semantics. Consequently, they achieve relatively low scores across different metrics for SQB, as the reference questions in SQB have shorter spans. However, for LC-QuAD and WebQuestionsSP, which contain

25. There are multiple entities connected to the entity *"Hong Kong"* by the relation *"media_common.netflix_genre.titles"* in Freebase. The answer should be validated by running SPARQL queries which correspond to the reasoning path and extract the connected tail entities. This problem is fixed in our dataset creation process by the SPARQL validation step.

**Table 7.6:** Aggregated annotation result on the sample question pair.

| | Grammar | Coverage | Consistency |
|---|---|---|---|
| Evaluation scores | 0.6111 | 0.7555 | 0.4555 |

both simple and complex triples and reference questions with longer text spans, our models have better scores.

### 7.4.4 Manual Metrics

As discussed in the related work, we have scrutinized the existing manual evaluation metrics and identified key indices that are pertinent to the demands of a comprehensive KGQA dataset. Specifically, we find that Consistency, Grammaticality, and Coverage are essential for nurturing high-performing KGQA systems.

**Consistency** pertains to the fidelity of the generated question with respect to the provided answer. Annotators are tasked with assessing the alignment between questions and the designated answer node/edge extracted from the SPARQL endpoint. This evaluation criterion is pivotal for determining whether the generated question can be effectively answered and accurately reflects real entities or relationships within the original subgraph. For example, given the subgraph with the head entity *"muscle organ"*, relation *"is associated with"*, and tail entity *"esophagus carcinoma in situ"* (the answer), a question such as *"What are the possible associations of muscle organ with esophagus carcinoma in situ?"* would be deemed incorrect, as it focuses on the relation rather than the tail entity. A preferred formulation would be *"What are the possible associations of the muscle organ?"* which directly targets the tail entity.

**Grammaticality** assesses whether the question adheres to linguistic correctness in terms of vocabulary, grammar, and structure. This criterion is crucial for ensuring that questions are understandable and interpretable by domain experts.

**Coverage** evaluates the fidelity of the generated question to the underlying subgraph or reasoning path. This aspect is vital for our KGQA dataset, as many KGQA systems, whether based on information retrieval or semantic parsing, heavily rely on the alignment between natural language questions and reasoning paths. For instance, a 2-hop subgraph should not be associated with a 1-hop question.

These standards are also highlighted as the weak points of current PLMs in the sense that the hallucinated generation of the model would be detected as false. **Model hallucination** is a focus of evaluation on the generated text by big models nowadays. In the context of triple-to-question, hallucination refers to the generation of content that lacks fidelity or is not supported by the source data provided. In this work, hallucination can be seen as a divergence of the generated questions to the input: i.e., the corresponding subgraph and the answer, which are the consistency and coverage in our manual metrics.

### 7.4.5 Manual Evaluation Result and Analysis

We ask three English-proficient annotators to independently provide feedback on each generated question based on the following three criteria, including two biomedical experts and one IT expert. The unannotated sample can be found in our project repository on GitHub. We aggregate the result and list it in Table 7.6. A detailed score from the annotators is also illustrated in Table 7.7.

As can be seen in Table 7.6, we have relatively high scores for Grammar and Coverage, while Consistency appears to be lower compared to the other two metrics. This can be

**Table 7.7**: The detailed scores from different annotators.

| | Grammaticality | Coverage | Consistency |
|---|---|---|---|
| Biologist-1 | 55 | 8 | 24 |
| Biologist-2 | 22 | 33 | 41 |
| IT Expert | 55 | 61 | 16 |

**Table 7.8**: $\kappa$ scores for the Grammaticality, Coverage and Consistency metrics.

| | Grammaticality | Coverage | Consistency |
|---|---|---|---|
| $\kappa$ | 0.12 | 0.17 | 0.15 |

due to the complexity of reasoning: Questions that follow a reasoning path often require understanding complex relationships, concepts, and logical structures within the graph. In the context of KGQA, the reasoning path refers to the multiple-fact triples in the KG corresponding to capturing this complexity. Replicating it in a question generation system can be difficult even for the PLM, especially for open-ended or higher-order thinking questions. Reasoning has always been proven challenging for PLMs and they're not optimized for this special task.

While Grammaticality and Coverage seem relatively satisfactory, efforts could be directed towards enhancing Consistency, possibly through refining the generation process or providing more context for the generated content or some post-editing/filtering techniques.

### 7.4.6 Inter-Annotator Agreement

We utilize Fleiss' Kappa as the metric for checking the reliability, i.e., coefficients of agreement among our annotators. The $\kappa$ score for each metric is listed in Table 7.8.

The measured agreement is rather low, showing a disparity between annotators. Biologist-1 and the IT Expert have given higher scores for Grammaticality compared to Biologist-2. There's some disparity between annotators, especially evident in the Grammaticality and Consistency scores.

The disparity can be attributed to external reasons. This can be due to the inexperience of the biologist experts hired who have limited knowledge in annotating an NLP dataset and with KGs. On the other hand, the IT Expert gives a Grammaticality score similar to Biologist-1 and a different score in Consistency. This is mostly due to the lack of domain knowledge in the biomedicine domain to match the concept in the question and answer. Nonetheless, we decided to base our quality analyses on the majority vote. On this basis, we did not deem it necessary to remove any examples for a lack of quality.

## 7.5 Generated Dataset

In total we have 85,368 question-answer pairs, since we filter out MONDO group resources from PrimeKG, as explained in Section 7.3.1. The generated dataset is partitioned into *train*, *test*, and *validation* set with a ratio of 6:2:2.

**Table 7.9**: The distribution of questions based on the number of nodes in their corresponding subgraphs. Also, the total number of relations (# rel.) and entities (# ent.) are listed.

| | 2-node | 3-node | 4-node | # q-a pairs | # rel. | # ent. |
|---|---|---|---|---|---|---|
| Train | 5,769 | 34,118 | 11,333 | 51,220 | 131,775 | 263,792 |
| Test | 1,955 | 11,272 | 3,847 | 17,074 | 44,035 | 87,786 |
| Val. | 2,008 | 11,276 | 3,790 | 17,074 | 43,932 | 87,840 |

### 7.5.1 Statistics

The numbers of 2-node, 3-node, and 4-node based questions, question-answer pairs, relations, and entities in each separation are exhibited in Table 7.9.

The majority of questions in each subset has 3 nodes, followed by 2-node and 4-node questions. As the number of nodes in the subgraph increases, the number of questions decreases, which is expected as subgraphs with more nodes are likely to be less common or more complex. This also aligns with the analysis from the existing biomedical QA datasets. The distribution of questions across different node counts is consistent across subsets, indicating that the dataset is well-balanced in terms of subgraph complexity across training, testing, and validation sets. This will ensure representative sampling during different stages of developing a model, such as training and testing.

## 7.6 Ethical Statement and Acknowledgement

All subjects gave their informed consent for inclusion before they participated in the study. This project was supported by the Ministry of Research and Education within the SifoLIFE project RESCUE-MATE (project number 13N16836), and by the Federal Ministry for Economic Affairs and Climate Action of Germany in the project CoyPu (project number 01MK21007G). We utilized two NVIDIA RTX A5000 graphics cards with 24GB of RAM, kindly provided by the NVIDIA Academic Hardware Grant Program.

## 7.7 Conclusion and Future Work

In this paper, we introduced a novel approach for addressing the challenge of generating high-quality question-answer pairs for BioKGQA systems. Leveraging PLMs and the PrimeKG, we devised a methodology to automatically construct a large-scale BioKGQA dataset. Our approach resulted in the creation of PrimeKGQA, a benchmarking resource comprising 83999 question-answer pairs alongside their corresponding SPARQL queries. This is so far the largest dataset in BioKGQA and is 1000 factors more than the second biggest dataset in this domain. Through a rigorous evaluation process involving both automatic metrics and manual annotations by domain experts, we established novel standards tailored specifically for assessing the linguistic correctness and semantic faithfulness of the generated questions. This ensures that PrimeKGQA serves as a reliable

benchmark for evaluating the performance of KGQA systems in the biomedical domain. On top, the dataset generation framework is training-free, adaptable to other domains, and supports evolving KGs, making it suitable for automatic dataset generation across various fields for automatic dataset generation.

While our work represents a significant step forward in addressing the dearth of large-scale BioKGQA datasets, several avenues for future research and improvement remain: Refined Question Generation, i.e., investigating methodologies to refine the generated questions, enabling examination of the output in desired dimensions and facilitating post-editing strategies for error correction. Application-oriented Evaluation, i.e., conducting current KGQA systems using PrimeKGQA to assess their effectiveness in supporting real-world biomedical tasks, such as clinical decision support and drug discovery.

# 8 Conclusion

## Contents



## 8.1 Summary

This dissertation has explored the domain of biomedical knowledge graph question answering from two complementary perspectives: the development of datasets and the design of information extraction systems. The contributions directly address Research Questions 1 and 2 formulated in the Chapter 1.

### 8.1.1 BioKGQA Dataset

In response to the motivation outlined in Chapter 1, we explored and reviewed methods for effectively leveraging structured resources in combination with LLMs in chapter 3 and 7. To tackle the challenge of generating high-quality biomedical KGQA data. We first examined existing research on active KGQA datasets, then manually created a KGQA dataset to study the process and characteristics of real-world datasets. Building on these findings, we designed a training-free, generalizable framework for KGQA dataset generation that is compatible with evolving KGs and scalable to other domains.

Research Question 1

How to generate KGQA datasets that exhibit a natural linguistic style and reflect a realistic distribution of both simple and complex questions?

In Chapter 3, we present a large-scale analysis of KGQA datasets, systems, and demonstrations. By surveying 100 publications and 98 systems over the past decade, we developed an open, centralized leaderboard for KGQA benchmarks[1], providing the first systematic meta-analysis of evaluation practices in this domain. From a dataset perspective, we examined factors contributing to inconsistencies in reported results, including differences in underlying knowledge graphs (e.g., DBpedia, Wikidata), dataset size, question complexity, and multilingual coverage. We focus on four widely used benchmarks representing both manual and automatic creation: **QALD series**: Manually created, multilingual datasets. QALD-8 contains 219 training and 42 test pairs, while QALD-9 offers 558 questions over DBpedia, each paired with a SPARQL query, answer URI, and answer type. **LC-QuAD series**: Automatically generated datasets with 5,000 ($v1.0$) and 30,000 ($v2.0$) question–SPARQL pairs over DBpedia, widely adopted for QA system development. Our comparison shows that most systems achieve consistent scores across publications, with standard deviations typically below 1%, thanks to an available evaluation, ongoing support for the KG endpoint, and a clear separation of the test and train set.

These findings directly address the challenge of creating a benchmark dataset in bioKGQA. Beyond linguistic naturalness, datasets should foster reproducibility and consistent reporting. The analysis suggests that research outcomes depend less on whether datasets are manual or automatic, and more on the standardization of evaluation tools and the rigor of publication practices.

Built on the findings from Chapter 4, we generate KGQA datasets using both manual and automatic approaches, partially addressing **RQ 1**. We introduce QALD 10—a multilingual, complex KGQA dataset designed for realistic and challenging evaluations. Questions were collected from humans to ensure realistic linguistic usage. They were annotated against Wikidata before being verified.

To capture question characteristics and complexity, QALD-10 was analyzed across features such as the number of triple patterns, join operators, join vertex degree, and SPARQL modifiers (e.g., `LIMIT`, `ORDER BY`, `GROUP BY`). Modifier occurrences and structural complexity were examined using the mean and standard deviation of the three query features. QALD-10 exhibits high variation across these features, indicating more diverse and complex queries, while the mean number of triple patterns remains between 1 and 2—consistent across datasets. These patterns provide guidance for generating KGQA datasets that reflect realistic question structures.

We address biomedical KGQA in chapter 7 by introducing PrimeKGQA, a large-scale dataset automatically constructed from PrimeKG—the largest precision medicine KG. PrimeKGQA comprises 85,368 question–answer pairs with corresponding SPARQL queries, covering diverse biomedical concepts and varying complexity levels, making it the largest BioKGQA dataset to date and a robust benchmark for knowledge-driven biomedical systems.

We designed a training-free, generalizable framework for KGQA dataset generation, compatible with evolving KGs and scalable to other domains. Subgraphs of varying complexity were sampled based on network motifs (graph structures) (Yan et al., 2024), converted into SPARQL queries to retrieve answers, and then paired with questions

1. https://kgqa.github.io/leaderboard/

generated by a PLM. The quality of both the framework and the dataset was validated through automatic and manual evaluations, demonstrating its effectiveness.

### 8.1.2 Information Extraction for bioKGQA

Chapter 6 and 5 address the motivation outlined in Chapter 1. To tackle the associated challenges, we explored existing and emerging methods for integrating LLMs with structured biomedical knowledge to support robust and interpretable systems for information extraction (IE) and KGQA. We conducted a comprehensive review of neural-symbolic IE methods, examining the sources of extra information and how they are incorporated into training and inference frameworks. Additionally, we evaluated the effectiveness of leveraging KG information as a pre-training objective in entity linking (EL) tasks.

**Research Question 2**

How to enhance information extraction of bioKGQA systems?

Existing neural-symbolic RE methods can be grouped into two categories based on how symbolic information is incorporated. First, symbolic knowledge can enhance distant supervision by using ontologies to mine or refine textual labels. Second, symbolic information can be directly integrated during training and inference to improve RE performance. External information is not only concatenated with text but also used to generate constraints applied during both training and inference. Symbolic knowledge in these methods includes KGs, prior meta-information, inferred logical rules, linguistic cues, temporal data, and commonsense knowledge.

Among those methods, KG information is the most commonly used extra knowledge, including attributes, synonyms, descriptions, etc. Some of them include enhancing the representation of the input via concatenating word embeddings and embeddings of subjects, objects, and relations. While some work adds the KG as part of the pre-training objectives, or uses the KG info in the prompt.

In chapter 5, we introduce a novel framework for pre-training generative LLMs using a corpus derived from a KG. Our evaluations, however, did not demonstrate a substantial benefit from incorporating synonym, description, or relational information. This preliminary work highlights the challenges of leveraging semantic knowledge in LLMs and suggests avenues for further research in scientific document processing.

We focused on enhancing bioEL by integrating linearized triples and synonym information. Unexpectedly, these additions resulted in only marginal gains in EL performance, emphasizing the inherent complexity of biomedical EL tasks. Future research could explore more effective strategies to inject KG knowledge into LLMs, such as utilizing the graph structure via Graph Neural Networks (GNNs) instead of relying solely on linearized representations.

## 8.2 Limitations and Future Work

In chapter 3, we presented a novel community resource to track advances in KGQA research. This platform serves as a central leaderboard to ensure that the community remains aligned. However, due to its current design, all datasets and systems must still be manually added, which requires considerable effort and time for maintenance. As a potential direction for future work, we propose the use of LLM-based workflows to streamline the process of adding and updating entries, thereby reducing manual overhead and ensuring the leaderboard remains current (Singh et al., 2024).

In addition, multiple leaderboards currently exist for individual datasets, and it would be valuable to explore automated synchronization between the KGQA leaderboard and other reporting platforms. Moreover, the current evaluations are primarily based on reported results in the literature, without direct analysis of the actual models or datasets. Future work should incorporate improved evaluation metrics applied directly to models, source code, or platform-based evaluations, thereby enabling more comprehensive and in-depth analyses of QA system capabilities.

For chapter 4, we created a multilingual KGQA dataset through manual construction based on Wikidata. The dataset currently covers English, German, Chinese, and Russian, which are relatively well-studied languages. A promising future direction would be to extend this work to low-resource languages or specialized domains, such as the biomedical domain, to foster greater inclusiveness. Moreover, as the KGQA community increasingly adopts logical forms in place of SPARQL queries, it would be worthwhile to incorporate logical representations into the dataset in future iterations.

For chapter 5, our work focused on integrating linearized triples and synonym information into biomedical entity linking. However, contrary to expectations, the inclusion of these elements resulted in only minimal performance improvements. Future research should investigate which types of knowledge are actually learned during the process, as well as analyze error cases before and after training to identify potential areas for improvement. Based on the findings, a possible extension is to develop more sophisticated methods for instructing LLMs to learn from external knowledge, ensuring efficient injection that benefits downstream tasks. For example, instead of relying solely on linearized representations, KG information could be incorporated by leveraging graph structures through Graph Neural Networks, an adapter (Chen et al., 2022) architecture that injects knowledge without changing the model parameter, RAG pipeline (Gao et al., 2023), which injects knowledge with linking, enabling the development of multiple enhanced approaches, etc. Moreover, our experiments were limited to BART, an encoder–decoder-based language model, and have not been extended to larger and more widely used LLMs such as Mixtral or LLama. Exploring these models could reveal different behaviors and benefits at larger scales. A comparative study that examines not only encoder–decoder models but also encoder-only and decoder-only architectures could provide valuable insights for future research on injecting knowledge into LLMs.

For chapter 6, which reviewed existing neuro-symbolic relation extraction methods, the continued exploration of such approaches holds promise for developing more robust, interpretable, and generalizable RE models. In our study, we focused on a structural analysis rather than a direct performance comparison. Consequently, we are unable to provide quantitative comparisons in terms of extensibility, performance gains, or other empirical measures. Future work should therefore combine both structural and performance-based evaluations to enable a more comprehensive assessment. Moreover, the establishment of standardized evaluation criteria for these methods is essential—encompassing not only performance, but also extensibility, reproducibility, and other key aspects (Schimmler et al., 2023).

In the chapter 7, we introduced a novel approach to address the challenge of generating high-quality question-answer pairs for BioKGQA systems. Leveraging large language models and PrimeKG, we devised a methodology to automatically construct a large-scale BioKGQA dataset. Our approach resulted in the creation of **PrimeKGQA**, a benchmarking resource comprising 83,999 question-answer pairs along with their corresponding SPARQL queries. Through rigorous evaluation using both automatic metrics and manual annotations by domain experts, we established novel standards specifically designed to assess the linguistic correctness and semantic faithfulness of generated questions. This ensures that PrimeKGQA serves as a reliable benchmark for evaluating the performance of

KGQA systems in the biomedical domain. Additionally, our dataset generation framework is training-free, adaptable to other domains, and compatible with evolving knowledge graphs, making it suitable for automatic dataset generation across diverse fields.

While our work from chapter 7 represents a significant step forward in addressing the lack of large-scale BioKGQA datasets, several avenues for future research remain. As indicated by the manual evaluation, the quality of the dataset is still far from perfect. Future work could focus on refining the generated questions, enabling analysis along specific dimensions, and supporting post-editing strategies for error correction. Additionally, evaluating current KGQA systems using PrimeKGQA to measure their effectiveness in real-world biomedical tasks, such as clinical decision support and drug discovery, could provide practical insights and guide further improvements. Another promising direction is to explore richer integration of knowledge graph information, potentially leveraging graph-structured representations and advanced LLMs, to enhance downstream task performance.

From a broader perspective, the contributions of this dissertation to BioKGQA research could be further strengthened by establishing experimental baselines for BioKGQA systems. This could involve exploring information retrieval-based or query translation-based approaches. Developing a demonstration system would also provide tangible benefits to the community by allowing users to interact with and evaluate the methods. Additionally, with the growing popularity of agent-based methods for complex tasks, investigating how LLM-based agents can decompose and solve complex BioKGQA questions represents a promising direction for future research.

# References


Abien Fred Agarap. 2018. Deep Learning using Rectified Linear Units (ReLU). *CoRR* abs/1803.08375. arXiv: 1803.08375. (Cited on pages 15 sq.).

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2021. Entity Linking and Discovery via Arborescence-based Supervised Clustering. *CoRR* abs/2109.01242. arXiv: 2109.01242. (Cited on page 64).

Ahmad Aghaebrahimian, Maria Anisimova, and Manuel Gil. n.d. Ontology-Aware Biomedical Relation Extraction, (cited on pages 8, 79, 83).

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events,* 1–8. (Cited on page 73).

Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based Inference for Biomedical Entity Linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021,* edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 2598–2608. Association for Computational Linguistics. (Cited on page 64).

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical Report. Anthropic. (Cited on page 25).

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007a. Dbpedia: A nucleus for a web of open data. In *international semantic web conference,* 722–735. Springer. (Cited on pages 28, 34).

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007b. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007,* edited by Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, 4825:722–735. Lecture Notes in Computer Science. Springer. (Cited on page 90).

Sören Auer, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and Mohamad Yaser Jaradeh. 2020. Improving Access to Scientific Literature with Knowledge Graphs. *Bibliothek Forschung und Praxis* 44 (3): 516–529. (Cited on pages 7, 36, 42).

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2022, Hybrid: Seattle, Washington, USA + Online, July 10-15, 2022,* edited by Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min, 209–220. Association for Computational Linguistics. (Cited on page 64).

Michael Azmy, Peng Shi, Jimmy Lin, and Ihab Ilyas. 2018. Farewell freebase: Migrating the simplequestions dataset to dbpedia. In *Proceedings of the 27th international conference on computational linguistics,* 2093–2103. (Cited on page 37).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings,* edited by Yoshua Bengio and Yann LeCun. (Cited on page 20).

Debayan Banerjee, Pranav Ajit Nair, Jivat Neet Kaur, Ricardo Usbeck, and Chris Biemann. 2022a. Modern baselines for SPARQL semantic parsing. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2260–2265. (Cited on page 6).

Debayan Banerjee, Pranav Ajit Nair, Jivat Neet Kaur, Ricardo Usbeck, and Chris Biemann. 2022b. Modern Baselines for SPARQL Semantic Parsing. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* (cited on page 52).

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization,* 65–72. (Cited on page 101).

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,* 2503–2514. (Cited on page 37).

Nikita Baramiia, Alina Rogulina, Sergey Petrakov, Valerii Kornilov, and Anton Razzhigaev. 2022. Ranking Approach to Monolingual Question Answering over Knowledge Graphs. *Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference (ESWC 2022),* (cited on page 55).

Frederic B Bastian, Julien Roux, Anne Niknejad, Aurélie Comte, Sara S Fonseca Costa, Tarcisio Mendes de Farias, Sébastien Moretti, Gilles Parmentier, Valentine Rech de Laval, Marta Rosikiewicz, Julien Wollbrett, Amina Echchiki, Angélique Escoriza, Walid H Gharib, Mar Gonzales-Porta, Yohan Jarosz, Balazs Laurenczy, Philippe Moret, Emilie Person, Patrick Roelli, Komal Sanjeev, Mathieu Seppey, and Marc Robinson-Rechavi. 2020. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research* 49, no. D1 (October): D831–D847. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D831/35364714/gkaa793.pdf. (Cited on page 99).

Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021,* edited by Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, 1673–1685. ACM / IW3C2. (Cited on pages 77, 83).

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA,* edited by Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, 932–938. MIT Press. (Cited on pages 18 sq.).

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing,* 1533–1544. (Cited on page 37).

Rajarshi Bhowmik, Karl Stratos, and Gerard de Melo. 2021. Fast and Effective Biomedical Entity Linking Using a Dual Encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, LOUHI@EACL, Online, April 19, 2021,* edited by Eben Holderness, Antonio Jimeno-Yepes, Alberto Lavelli, Anne-Lyse Minard, James Pustejovsky, and Fabio Rinaldi, 28–37. Association for Computational Linguistics. (Cited on page 64).

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia-A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web* 7 (3): 154–165. (Cited on page 27).

Olivier Bodenreider. 2004a. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32 (suppl_1): D267–D270. (Cited on page 3).

Olivier Bodenreider. 2004b. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32 (Database-Issue): 267–270. (Cited on page 79).

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008a. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data,* 1247–1250. (Cited on page 28).

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008b. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* 1247–1250. SIGMOD '08. Vancouver, Canada: Association for Computing Machinery. (Cited on page 48).

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008c. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* 1247–1250. SIGMOD '08. Vancouver, Canada: Association for Computing Machinery. (Cited on page 87).

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR* abs/2108.07258. arXiv: 2108.07258. (Cited on page 24).

Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L Hamilton. 2022. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics* 23, no. 6 (September): bbac404. eprint: https://academic.oup.com/bib/article-pdf/23/6/bbac404/47144248/bbac404.pdf. (Cited on page 2).

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale Simple Question Answering with Memory Networks. *CoRR* abs/1506.02075. arXiv: 1506.02075. (Cited on page 37).

Andreas Both, Dennis Diefenbach, Kuldeep Singh, Saedeeh Shekarpour, Didier Cherix, and Christoph Lange. 2016. Qanary–A Methodology for Vocabulary-Driven Open Question Answering Systems. In *European Semantic Web Conference,* 625–641. Springer. (Cited on page 48).

Andreas Both, Aleksandr Perevalov, Johannes Richard Bartsch, Paul Heinze, Rostislav Iudin, Johannes Rudolf Herkner, Tim Schrader, Jonas Wunsch, Ann Kristin Falkenhain, and René Gürth. 2021. A Question Answering System for retrieving German COVID-19 data driven and quality-controlled by Semantic Technology. In *Joint Proceedings of the Semantics co-located events: Poster&Demo track and Workshop on Ontology-Driven Conceptual Modelling of Digital Twins co-located with Semantics 2021, Amsterdam and Online, September 6-9, 2021,* edited by Ilaria Tiddi, Maria Maleshkova, Tassilo Pellegrini, and Victor de Boer, vol. 2941. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 34).

Marina Boudin, Gayo Diallo, Martin Drancé, and Fleur Mougin. 2023. The OREGANO knowledge graph for computational drug repurposing. *Scientific data* 10 (1): 871. (Cited on page 99).

Anthony J Brookes and Peter N Robinson. 2015. Human genotype–phenotype databases: aims, challenges and opportunities. *Nature Reviews Genetics* 16 (12): 702–715. (Cited on page 98).

Mikhail S. Burtsev, Alexander V. Seliverstov, Rafael Airapetyan, Mikhail Y. Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-Source Library for Dialogue Systems. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations,* edited by Fei Liu and Thamar Solorio, 122–127. Association for Computational Linguistics. (Cited on page 55).

Pei-Xuan Cai, Yao-Chung Fan, and Fang-Yie Leu. 2022. Compare Encoder-Decoder, Encoder-Only, and Decoder-Only Architectures for Text Generation on Low-Resource Datasets. In *Advances on Broad-Band Wireless Computing, Communication and Applications,* edited by Leonard Barolli, 216–225. Cham: Springer International Publishing. (Cited on page 25).

Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 423–433. (Cited on page 37).

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net. (Cited on page 65).

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual Autoregressive Entity Linking. *Trans. Assoc. Comput. Linguistics* 10:274–290. (Cited on page 65).

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022,* edited by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, 6101–6119. Association for Computational Linguistics. (Cited on page 7).

Tianshi Cao, Marc T. Law, and Sanja Fidler. 2020. A Theoretical Analysis of the Number of Shots in Few-Shot Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net. (Cited on page 98).

Tommaso Caselli and Piek Vossen. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. In *Proceedings of the Events and Stories in the News Workshop@ACL 2017, Vancouver, Canada, August 4, 2017,* edited by Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard H. Hovy, Teruko Mitamura, and David Caswell, 77–86. Association for Computational Linguistics. (Cited on pages 89 sq.).

Nilesh Chakraborty, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann, and Asja Fischer. 2021. Introduction to neural network-based question answering over knowledge graphs. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 11 (3). (Cited on page 36).

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023a. Building a knowledge graph to enable precision medicine. *Scientific Data* 10 (1): 67. (Cited on pages 2 sq., 27).

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023b. Building a knowledge graph to enable precision medicine. *Scientific Data* 10 (1): 67. (Cited on pages 64, 98 sq.).

Andreas Chandra, Affandy Fahrizain, Simon Willyanto Laufried, et al. 2021. A Survey on non-English Question Answering Dataset. *arXiv preprint arXiv:2112.13634,* (cited on page 34).

Lu Chang, Ruihuan Zhang, Jia Lv, Weiguang Zhou, and Yunli Bai. 2022. A review of biomedical named entity recognition. *J. Comput. Methods Sci. Eng.* 22 (3): 893–900. (Cited on page 29).

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Ishaan Singh Rawal, Cheston Tan, Dongkyu Choi, Bo Xiong, and Bo Ai. 2024. LLM-Based Multi-Hop Question Answering with Knowledge Graph Integration in Evolving Environments. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 14438–14451. Miami, Florida, USA: Association for Computational Linguistics, November. (Cited on page 6).

Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys* 56 (12): 1–32. (Cited on page 24).

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. 2022. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534,* (cited on page 116).

Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. 2020. Knowledge Graph Completion: A Review. *IEEE Access* 8:192435–192456. (Cited on page 72).

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530,* (cited on page 105).

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. HacRED: A Large-Scale Relation Extraction Dataset Toward Hard Cases in Practical Applications. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021,* edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, vol. ACL/IJCNLP 2021, 2819–2831. Findings of ACL. Association for Computational Linguistics. (Cited on pages 89 sq.).

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the Growth: Difficulty-Controllable Question Generation through Step-by-Step Rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 5968–5978. Online: Association for Computational Linguistics, August. (Cited on page 100).

Philipp Cimiano, Vanessa Lopez, Christina Unger, Elena Cabrio, Axel-Cyrille Ngonga Ngomo, and Sebastian Walter. 2013. Multilingual Question Answering over Linked Data (QALD-3): Lab Overview. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization,* 321–332. Springer. (Cited on page 50).

Fuze Cong, Wenping Hu, Qiang Huo, and Li Guo. 2019. A Comparative Study of Attention-Based Encoder-Decoder Approaches to Natural Scene Text Recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR),* 916–921. (Cited on page 25).

UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic acids research* 43 (D1): D204–D212. (Cited on page 88).

Council of Europe. 2001. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe. (Cited on page 51).

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2021. Multilingual Compositional Wikidata Questions. *arXiv preprint arXiv:2108.03509,* (cited on page 37).

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022. Compositional Generalization in Multilingual Semantic Parsing over Wikidata. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 10:937–955. (Cited on page 49).

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. n.d. Universal Graph based Relation Extraction, (cited on pages 8, 76).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),* 4171–4186. (Cited on pages 20 sq., 23).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers),* edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–4186. Association for Computational Linguistics. (Cited on pages 73 sq.).

Dennis Diefenbach, Andreas Both, Kamal Singh, and Pierre Maret. 2020. Towards a question answering system over the semantic web. *Semantic Web* 11 (3): 421–439. (Cited on page 41).

Dennis Diefenbach, Vanessa López, Kamal Deep Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowl. Inf. Syst.* 55 (3): 529–569. (Cited on pages 36 sq.).

Dennis Diefenbach, Kamal Singh, and Pierre Maret. 2017a. WDAqua-core0: A Question Answering Component for the Research Community. In *Semantic Web Challenges,* edited by Mauro Dragoni, Monika Solanki, and Eva Blomqvist, 84–89. Cham: Springer International Publishing. (Cited on page 41).

Dennis Diefenbach, Kamal Singh, and Pierre Maret. 2018. WDAqua-Core1: A Question Answering Service for RDF Knowledge Bases, 1087–1091. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee. (Cited on page 48).

Dennis Diefenbach, Kamal Deep Singh, and Pierre Maret. 2017b. WDAqua-core0: A Question Answering Component for the Research Community. In *Semantic Web Challenges - 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28 - June 1, 2017, Revised Selected Papers,* edited by Mauro Dragoni, Monika Solanki, and Eva Blomqvist, 769:84–89. Communications in Computer and Information Science. Springer. (Cited on pages 54 sq.).

Dennis Diefenbach, Thomas Pellissier Tanon, Kamal Deep Singh, and Pierre Maret. 2017a. Question Answering Benchmarks for Wikidata. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017.* (Cited on page 37).

Dennis Diefenbach, Thomas Pellissier Tanon, Kamal Deep Singh, and Pierre Maret. 2017b. Question Answering Benchmarks for Wikidata. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017,* edited by Nadeschda Nikitina, Dezhao Song, Achille Fokoue, and Peter Haase, vol. 1963. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 48).

Dennis Diefenbach, Max De Wilde, and Samantha Alipio. 2021. Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph. In *International Semantic Web Conference,* 631–647. Springer. (Cited on page 34).

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal.* European Language Resources Association. (Cited on page 84).

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Informatics* 47:1–10. (Cited on page 67).

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019a. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II,* edited by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 11779:69–78. Lecture Notes in Computer Science. Springer. (Cited on pages 34, 37).

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019b. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II,* edited by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 11779:69–78. Lecture Notes in Computer Science. Springer. (Cited on pages 49, 52).

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019c. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II,* edited by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 11779:69–78. Lecture Notes in Computer Science. Springer. (Cited on page 105).

Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri, and Jens Lehmann. 2018. EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I,* edited by Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, 11136:108–126. Lecture Notes in Computer Science. Springer. (Cited on page 37).

Joe Ellis, Jeremy Getman, Justin Mott, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, and Jonathan Wright. 2013. Linguistic Resources for 2013 Knowledge Base Population Evaluations. In *Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013.* NIST. (Cited on page 84).

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the linked data web. In *International semantic web conference,* 50–65. Springer. (Cited on pages 27, 34, 48, 72).

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020,* edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 4846–4853. Association for Computational Linguistics. (Cited on page 42).

Shengda Fan, Shasha Mo, and Jianwei Niu. 2022. Boosting Document-Level Relation Extraction by Mining and Injecting Logical Rules. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 10311–10323. (Cited on pages 8, 79, 82 sq.).

Christaine Fellbaum. 1998. WordNet: An Electronic Lexical Database. MIT Press. (Cited on page 80).

Javier D. Fernández, Miguel A. Martínez-Prieto, Claudio Gutierrez, Axel Polleres, and Mario Arias. 2013. Binary RDF representation for publication and exchange (HDT). *J. Web Semant.* 19:22–41. (Cited on page 61).

Hagen Aad Fock. 2022. Knowledge Graph Expansion Using Question Answering By Leveraging Pre-Trained Language Models. Master's thesis, Utrecht University. (Cited on page 100).

Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 international conference on management of data,* 1433–1445. (Cited on page 27).

Bin Fu, Yunqi Qiu, Chengguang Tang, Yang Li, Haiyang Yu, and Jian Sun. 2020. A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges. *CoRR* abs/2007.13069. arXiv: 2007.13069. (Cited on pages 34, 36).

Tingzhao Fu, Jianfa Zhang, Run Sun, Yuyao Huang, Wei Xu, Sigang Yang, Zhihong Zhu, and Hongwei Chen. 2024. Optical neural networks: progress and challenges. *Light: Science & Applications* 13 (1): 263. (Cited on page 16).

Mikhail Galkin, Priyansh Trivedi, Gaurav Maheshwari, Ricardo Usbeck, and Jens Lehmann. 2020. Message Passing for Hyper-Relational Knowledge Graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020,* edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 7346–7359. Association for Computational Linguistics. (Cited on page 72).

Fernando Gallego, Pedro Ruas, Francisco M Couto, and Francisco J Veredas. 2025. Enhancing cross-encoders using knowledge graph hierarchy for medical entity linking in zero-and few-shot scenarios. *Knowledge-Based Systems* 314:113211. (Cited on page 65).

Artur Galstyan. 2022. A systematic survey of relation extraction for knowledge graph question answering. Universität Hamburg. (Cited on page 73).

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling Document-level Causal Structures for Event Causal Relation Identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 1808–1817. (Cited on page 81).

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019,* edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 6249–6254. Association for Computational Linguistics. (Cited on pages 84, 86 sq.).

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2 (1). (Cited on page 116).

Tobias Glasmachers. 2017. Limits of end-to-end learning. In *Asian conference on machine learning,* 17–32. PMLR. (Cited on page 29).

Goran Glavas, Jan Snajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A Corpus for Extracting Event Hierarchies from News Stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014,* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, 3678–3683. European Language Resources Association (ELRA). (Cited on pages 89 sq.).

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. Deep learning. Vol. 1. 2. MIT press Cambridge. (Cited on page 17).

Paul Gorrell et al. 1995. Syntax and parsing. Vol. 76. Cambridge University Press Cambridge. (Cited on page 73).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783,* (cited on page 25).

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26. (Cited on page 24).

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond IID: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021,* 3477–3488. (Cited on page 37).

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. n.d. Beyond IID: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021,* 3477–3488. ACM. (Cited on page 61).

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Informatics* 45 (5): 885–892. (Cited on pages 84 sqq.).

Negacy Hailu, Lawrence Hunter, and K Bretonnel Cohen. 2013. UColorado_SOM: extraction of drug-drug interactions from biomedical text using knowledge-rich and knowledge-poor features. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* 684–688. (Cited on pages 84 sqq.).

Kelvin Han, Thiago Castro Ferreira, and Claire Gardent. 2022. Generating questions from Wikidata triples. In *13th Edition of its Language Resources and Evaluation Conference.* (Cited on page 100).

Kelvin Han and Claire Gardent. 2023. Generating and Answering Simple and Complex Questions from Text and from Knowledge Graphs. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers),* 285–304. (Cited on page 100).

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction. *arXiv preprint arXiv:2004.03186,* (cited on pages 31, 72).

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* edited by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 4803–4809. Brussels, Belgium: Association for Computational Linguistics, October. (Cited on page 87).

Chen Haotian, Chen Yijiang, and Zhou Xiangdong. 2024. Understanding More Knowledge Makes the Transformer Perform Better in Document-level Relation Extraction. In *Asian Conference on Machine Learning,* 231–246. PMLR. (Cited on pages 82 sq.).

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *J. Biomed. Informatics* 46 (5): 914–920. (Cited on pages 85 sq.).

Lars Patrick Hillebrand, Tobias Deußer, Tim Dilmaghani Khameneh, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. KPI-BERT: A Joint Named Entity Recognition and Relation Extraction Model for Financial Reports. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022,* 606–612. IEEE. (Cited on page 32).

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 6:e26726. (Cited on page 99).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9 (8): 1735–1780. (Cited on page 19).

Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2017. Survey on challenges of Question Answering in the Semantic Web. *Semantic Web* 8 (6): 895–920. (Cited on pages 36 sq.).

Aidan Hogan, Andreas Harth, and Stefan Decker. 2006. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems.* (Cited on page 8).

Alexander Miserlis Hoyle, Ana Marasović, and Noah A. Smith. 2021. Promoting Graph Awareness in Linearized Graph-to-Text Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 944–956. Online: Association for Computational Linguistics, August. (Cited on page 64).

Sen Hu, Lei Zou, Jeffrey Xu Yu, Haixun Wang, and Dongyan Zhao. 2018. Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs (Extended Abstract). In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018,* 1815–1816. IEEE Computer Society. (Cited on page 34).

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual,* edited by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. (Cited on page 34).

Xiaofeng Huang, Jixin Zhang, Zisang Xu, Lu Ou, and Jianbin Tong. 2021. A knowledge graph based question answering method for medical domain. *PeerJ Comput. Sci.* 7:e667. (Cited on page 7).

Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. 2024. Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada,* edited by Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, 18327–18335. AAAI Press. (Cited on pages 79, 83).

Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. 2024. Knowledge-Driven Cross-Document Relation Extraction. *CoRR* abs/2405.13546. arXiv: 2405.13546. (Cited on page 80).

Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* 230–247. Springer. (Cited on pages 8, 79, 83).

Yixin Ji, Kaixin Wu, Juntao Li, Wei Chen, Mingjie Zhong, Jia Xu, and Min Zhang. 2024. Retrieval and Reasoning on KGs: Integrate Knowledge Graphs into Large Language Models for Complex Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 7598–7610. Association for Computational Linguistics. (Cited on page 7).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55 (12): 1–38. (Cited on pages 4, 25).

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018,* 1057–1062. (Cited on page 37).

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex Temporal Question Answering on Knowledge Graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management,* 792–802. (Cited on page 37).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825,* (cited on page 105).

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088,* (cited on page 25).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR* abs/2310.06825. arXiv: 2310.06825. (Cited on pages 14, 24).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv: 2310.06825 [cs.CL]. (Cited on page 25).

Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, and Ji-Rong Wen. 2024. Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. *arXiv preprint arXiv:2402.11163,* (cited on page 7).

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: a new factoid QA data set matching Trivia-style question-answer pairs with freebase. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 318–323. (Cited on page 37).

Longquan Jiang and Ricardo Usbeck. 2022. Knowledge Graph Question Answering Datasets and Their Generalizability: Are They Enough for Future Research? In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022,* edited by Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, 3209–3218. ACM. (Cited on page 61).

Longquan Jiang, Xi Yan, and Ricardo Usbeck. 2023a. A Structure and Content Prompt-based Method for Knowledge Graph Question Answering over Scholarly Data. In *Joint Proceedings of Scholarly QALD 2023 and SemREC 2023 co-located with 22nd International Semantic Web Conference ISWC 2023, Athens, Greece, November 6-10, 2023,* edited by Debayan Banerjee, Ricardo Usbeck, Nandana Mihindukulasooriya, Gunjan Singh, Raghava Mutharaju, and Pavan Kapanipathi, vol. 3592. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 6).

Longquan Jiang, Xi Yan, and Ricardo Usbeck. 2023b. A Structure and Content Prompt-based Method for Knowledge Graph Question Answering over Scholarly Data. In *Joint Proceedings of Scholarly QALD 2023 and SemREC 2023 co-located with 22nd International Semantic Web Conference ISWC 2023, Athens, Greece, November 6-10, 2023,* edited by Debayan Banerjee, Ricardo Usbeck, Nandana Mihindukulasooriya, Gunjan Singh, Raghava Mutharaju, and Pavan Kapanipathi, vol. 3592. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 11).

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)* 55 (2): 1–36. (Cited on page 2).

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. GenIE: Generative Information Extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022,* edited by Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, 4626–4643. Association for Computational Linguistics. (Cited on page 74).

Daniel Jurafsky and James H Martin. 2018. Speech and language processing (draft). *preparation [cited 2022 January 4] Available from: https://web. stanford. edu/~ jurafsky/slp3,* (cited on page 34).

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd. Online manuscript released January 12, 2025. (Cited on page 14).

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *CoRR* abs/2001.08361. arXiv: 2001.08361. (Cited on page 24).

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 2526–2538. Online: Association for Computational Linguistics, August. (Cited on page 100).

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713,* (cited on page 37).

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net. (Cited on page 49).

Sun Kim, Rezarta Islamaj Dogan, Andrew Chatr-Aryamontri, Mike Tyers, W John Wilbur, and Donald C Comeau. 2015. Overview of biocreative v bioc track. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, Sevilla, Spain,* 1–9. (Cited on page 88).

Gerhard Klager and Axel Polleres. 2023a. Is GPT fit for KGQA? - Preliminary Results. In *Joint Proceedings of the Second International Workshop on Knowledge Graph Generation From Text and the First International BiKE Challenge co-located with 20th Extended Semantic Conference (ESWC 2023), Hersonissos, Greece, May 29th, 2023,* edited by Sanju Tiwari, Nandana Mihindukulasooriya, Francesco Osborne, Dimitris Kontokostas, Jennifer D'Souza, Mayank Kejriwal, and Edgard Marx, 3447:171–191. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 7).

Gerhard Georg Klager and Axel Polleres. 2023b. Is GPT fit for KGQA. In *Proceedings of the International Workshop on Knowledge Graph Generation from Text, co-located with Extended Semantic Web Conference,* 93. (Cited on pages 3, 25).

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. 2024a. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic acids research* 52 (D1): D1265–D1275. (Cited on page 85).

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. 2024b. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Research* 52 (D1): D1265–D1275. (Cited on page 99).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022,* edited by Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh. (Cited on page 25).

Vladislav Korablinov and Pavel Braslavski. 2020a. RuBQ: a Russian dataset for question answering over Wikidata. In *International Semantic Web Conference,* 97–110. Springer. (Cited on page 37).

Vladislav Korablinov and Pavel Braslavski. 2020b. RuBQ: A Russian Dataset for Question Answering over Wikidata. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II,* edited by Jeff Z. Pan, Valentina A. M. Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal, 12507:97–110. Lecture Notes in Computer Science. Springer. (Cited on page 49).

Franz Krause, Xi Yan, Baptiste Darnala, and Michel Dumontier. 2023. On the Combination of Event Calculus and Empirical Semantic Drifts. In *Joint Proceedings of the ESWC 2023 Workshops and Tutorials co-located with 20th European Semantic Web Conference (ESWC 2023), Hersonissos, Greece, May 28-29, 2023,* edited by Mehwish Alam, Cássia Trojahn, Sven Hertling, Catia Pesquita, Christian Aebeloe, Hidir Aras, Amr Azzam, Juan Cano, John Domingue, Simon Gottschalk, Olaf Hartig, Katja Hose, Sabrina Kirrane, Pasquale Lisena, Francesco Osborne, Philipp D. Rohde, Luc Steels, Ruben Taelman, Aisling Third, Ilaria Tiddi, and Rima Türker, vol. 3443. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 11).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25. (Cited on page 19).

Ruben Kruiper, Julian F. V. Vincent, Jessica Chen-Burger, Marc P. Y. Desmulliez, and Ioannis Konstas. 2020. In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020,* edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, 1489–1500. Association for Computational Linguistics. (Cited on page 84).

Sanjay Kukreja, Tarun Kumar, Amit Purohit, Abhijit Dasgupta, and Debashis Guha. 2024. A literature survey on open source large language models. In *Proceedings of the 2024 7th International Conference on Computers in Management and Business,* 133–143. (Cited on page 26).

Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. 2018. An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers),* edited by Marilyn A. Walker, Heng Ji, and Amanda Stent, 97–106. Association for Computational Linguistics. (Cited on page 84).

Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-controllable multi-hop question generation from knowledge graphs. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18,* 382–398. Springer. (Cited on page 100).

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be Overkill: A Tiny but Effective Biomedical Entity Linker based on Residual Convolutional Neural Networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021,* 1631–1639. Punta Cana, Dominican Republic: Association for Computational Linguistics, November. (Cited on page 68).

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021,* edited by Zhi-Hua Zhou, 4483–4491. ijcai.org. (Cited on page 56).

Ora Lassila and Ralph R Swick. 1999. Resource description framework (RDF) model and syntax specification. (Cited on page 26).

Trung Hoang Le, Huiping Cao, and Tran Cao Son. 2023. ASPER: Answer Set Programming Enhanced Neural Network Models for Joint Entity-Relation Extraction. *Theory and Practice of Logic Programming* 23, no. 4 (July): 765–781. (Cited on pages 8, 76).

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1 (4): 541–551. (Cited on page 19).

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal,* (cited on page 48).

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691,* (cited on page 25).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461,* (cited on page 100).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020,* edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, 7871–7880. Association for Computational Linguistics. (Cited on pages 14, 21, 23, 68).

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016a. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016. (Cited on pages 88 sq.).

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016b. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation* 2016. (Cited on page 67).

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016c. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation* 2016. (Cited on page 88).

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* 34 (1): 50–70. (Cited on page 29).

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021. Few-shot Knowledge Graph-to-Text Generation with Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 1558–1568. Online: Association for Computational Linguistics, August. (Cited on page 64).

Mingchen Li, Huixue Zhou, Han Yang, and Rui Zhang. 2024. RT: a Retrieving and Chain-of-Thought framework for few-shot medical named entity recognition. *Journal of the American Medical Informatics Association* 31 (9): 1929–1938. (Cited on page 3).

Wanli Li and Tieyun Qian. 2022. Graph-based Model Generation for Few-Shot Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 62–71. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, December. (Cited on pages 8, 81, 83).

Yanzeng Li, Bowen Yu, Xue Mengge, and Tingwen Liu. 2020. Enhancing Pre-trained Chinese Character Representation with Word-aligned Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 3442–3448. Online: Association for Computational Linguistics, July. (Cited on page 79).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out,* 74–81. (Cited on page 101).

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu. 2024. BioKGBench: A Knowledge Graph Checking Benchmark of AI Agent for Biomedical Science. *arXiv preprint arXiv:2407.00466,* (cited on page 98).

Carolyn E Lipscomb. 2000a. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88 (3): 265. (Cited on page 3).

Carolyn E Lipscomb. 2000b. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88 (3): 265. (Cited on page 88).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437,* (cited on page 3).

Bang Liu, Fred X Han, Di Niu, Linglong Kong, Kunfeng Lai, and Yu Xu. 2020. Story Forest: Extracting Events and Telling Stories from Breaking News. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14 (3): 1–28. (Cited on page 80).

Yijun Liu, Feifei Dai, Xiaoyan Gu, Haihui Fan, Dong Liu, Bo Li, and Weiping Wang. 2023. Powering Fine-Tuning: Learning Compatible and Class-Sensitive Representations for Domain Adaption Few-shot Relation Extraction. In *Database Systems for Advanced Applications - 28th International Conference, DASFAA 2023, Tianjin, China, April 17-20, 2023, Proceedings, Part IV,* edited by Xin Wang, Maria Luisa Sapino, Wook-Shin Han, Amr El Abbadi, Gill Dobbie, Zhiyong Feng, Yingxiao Shao, and Hongzhi Yin, 13946:121–131. Lecture Notes in Computer Science. Springer. (Cited on pages 8, 78, 83).

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-Shot Entity Linking by Reading Entity Descriptions. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers,* edited by Anna Korhonen, David R. Traum, and Lluís Màrquez, 3449–3460. Association for Computational Linguistics. (Cited on page 64).

Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. 2013. Evaluating Question Answering over Linked Data. *Journal of Web Semantics* 21 (June). (Cited on page 50).

Yilin Lu, Xiaoqiang Wang, Haofeng Yang, and Siliang Tang. 2023. KICE: A Knowledge Consolidation and Expansion Framework for Relation Extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 37:13336–13343. 11. (Cited on pages 8, 81, 83).

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602,* (cited on pages 88 sq.).

Linhao Luo, Zicheng Zhao, Chen Gong, Gholamreza Haffari, and Shirui Pan. 2024. Graph-constrained Reasoning: Faithful Reasoning on Knowledge Graphs with Large Language Models. *CoRR* abs/2410.13080. arXiv: 2410.13080. (Cited on page 3).

John Lyons. 1995. Linguistic semantics: An introduction. Cambridge University Press. (Cited on page 14).

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *Seventh Biennial Conference on Innovative Data Systems Research, CIDR 2015, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings.* www.cidrdb.org. (Cited on page 100).

Gaurav Maheshwari, Priyansh Trivedi, Denis Lukovnikov, Nilesh Chakraborty, Asja Fischer, and Jens Lehmann. 2019. Learning to Rank Query Graphs for Complex Question Answering over Knowledge Graphs. In *The Semantic Web – ISWC 2019,* edited by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 487–504. Cham: Springer International Publishing. (Cited on page 41).

Christopher D Manning. 2008. Introduction to Information Retrieval. Syngress Publishing, (cited on page 53).

Peter Hugoe Matthews. 1981. Syntax. Cambridge University Press. (Cited on page 14).

Peter Hugoe Matthews. 1991. Morphology. Cambridge university press. (Cited on page 14).

Carolyn J Mattingly, Glenn T Colby, John N Forrest, and James L Boyer. 2003. The comparative toxicogenomics database (CTD). *Environmental health perspectives* 111 (6): 793–795. (Cited on page 88).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26. (Cited on page 79).

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings,* edited by Yoshua Bengio and Yann LeCun. (Cited on page 18).

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298 (5594): 824–827. eprint: https://www.science.org/doi/pdf/10.1126/science.298.5594.824. (Cited on pages 101 sqq.).

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore,* edited by Keh-Yih Su, Jian Su, and Janyce Wiebe, 1003–1011. The Association for Computer Linguistics. (Cited on page 75).

MN Mohammed, Ammar Al Dallal, Mariam Emad, Abdul Qader Emran, and Malak Al Qaidoom. 2024. A comparative analysis of artificial hallucinations in GPT-3.5 and GPT-4: Insights into AI progress and challenges. *Business Sustainability with Artificial Intelligence (AI): Challenges and Opportunities: Volume 2,* 197–203. (Cited on page 25).

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019.* (Cited on page 67).

Cedric Möller and Ricardo Usbeck. 2025. Analyzing the Influence of Knowledge Graph Information on Relation Extraction. In *The Semantic Web,* edited by Edward Curry, Maribel Acosta, Maria Poveda-Villalón, Marieke van Erp, Adegboyega Ojo, Katja Hose, Cogan Shimizu, and Pasquale Lisena, 460–480. Cham: Springer Nature Switzerland. (Cited on page 3).

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches. *Applied Sciences* 11 (12). (Cited on page 34).

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30 (1): 3–26. (Cited on page 28).

Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck. 2020. Language Model Transformers as Evaluators for Open-domain Dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020,* edited by Donia Scott, Núria Bel, and Chengqing Zong, 6797–6808. International Committee on Computational Linguistics. (Cited on page 101).

Qiang Ning, Hao Wu, and Dan Roth. 2018. A Multi-Axis Annotation Scheme for Event Temporal Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers,* edited by Iryna Gurevych and Yusuke Miyao, 1318–1328. Association for Computational Linguistics. (Cited on pages 89 sq.).

Kristian Noullet, Ayoub Ourgani, and Michael Färber. 2023. A Full-Fledged Framework for Combining Entity Linking Systems and Components. In *Proceedings of the 12th Knowledge Capture Conference 2023,* 148–156. K-CAP '23. Pensacola, FL, USA: Association for Computing Machinery. (Cited on page 64).

OpenAI. 2023. GPT-4 Technical Report. arXiv: 2303.08774 [cs.CL]. (Cited on pages 3, 20, 24 sq.).

Abdelghny Orogat and Ahmed El-Roby. 2021. CBench: Demonstrating Comprehensive Evaluation of Question Answering Systems over Knowledge Graphs Through Deep Analysis of Benchmarks. *Proc. VLDB Endow.* 14 (12): 2711–2714. (Cited on page 42).

Abdelghny Orogat, I-Chien Liu, and Ahmed El-Roby. 2021. CBench: Towards Better Evaluation of Question Answering Over Knowledge Graphs. *Proc. VLDB Endow.* 14:1325–1337. (Cited on pages 38, 40).

Chinonso Cynthia Osuji, Thiago Castro Ferreira, and Brian Davis. 2024. A Systematic Review of Data-to-Text NLG. *arXiv preprint arXiv:2402.08496,* (cited on pages 100 sq.).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35:27730–27744. (Cited on page 24).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics,* 311–318. (Cited on page 101).

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Trans. Assoc. Comput. Linguistics* 5:101–115. (Cited on page 88).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL,* edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans, 1532–1543. ACL. (Cited on page 18).

Arnaldo Pereira, Alina Trifan, Rui Pedro Lopes, and José Luís Oliveira. 2022. Systematic review of question answering over knowledge bases. *IET Software* 16 (1): 1–13. (Cited on pages 7, 73).

Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. 2022b. QALD-9-plus: A Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers, 229–234. January. (Cited on pages 50 sq.).

Aleksandr Perevalov, Axel-Cyrille Ngonga Ngomo, and Andreas Both. 2022. Enhancing the Accessibility of Knowledge Graph Question Answering Systems through Multilingualization. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC),* 251–256. (Cited on page 48).

Aleksandr Perevalov, Xi Yan, Liubov Kovriguina, Longquan Jiang, Andreas Both, and Ricardo Usbeck. 2022. Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2998–3007. Marseille, France: European Language Resources Association, June. (Cited on pages 48, 53 sq.).

Aleksandr Perevalov*, Xi Yan*, Liubov Kovriguina, Longquan Jiang, Andreas Both, and Ricardo Usbeck. 2022. Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, 2998–3007. * Shared first authorship. Marseille, France: European Language Resources Association, June. (Cited on pages 7, 10 sq., 28).

Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. 2009. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)* 34 (3): 1–45. (Cited on page 27).

Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. 2016. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research,* gkw943. (Cited on page 98).

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinform.* 8. (Cited on page 84).

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019. OpenHowNet: An Open Sememe-based Lexical Knowledge Base. *arXiv preprint arXiv:1901.09957,* (cited on page 79).

Muhammad Reza Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are Decoder-Only Language Models Better than Encoder-Only Language Models in Understanding Word Meaning? In *Findings of the Association for Computational Linguistics: ACL 2024,* edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 16339–16347. Bangkok, Thailand: Association for Computational Linguistics, August. (Cited on page 25).

Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. 2020. Few-shot Relation Extraction via Bayesian Meta-learning on Relation Graphs. In *Proceedings of the 37th International Conference on Machine Learning,* edited by Hal Daumé III and Aarti Singh, 119:7867–7876. Proceedings of Machine Learning Research. PMLR, 13–18 Jul. (Cited on page 74).

Chris Quirk and Hoifung Poon. 2017. Distant Supervision for Relation Extraction beyond the Sentence Boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers,* edited by Mirella Lapata, Phil Blunsom, and Alexander Koller, 1171–1182. Association for Computational Linguistics. (Cited on page 88).

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training, (cited on pages 20 sq., 23).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21 (140): 1–67. (Cited on pages 21, 100).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21:140:1–140:67. (Cited on pages 55, 74).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016,* edited by Jian Su, Xavier Carreras, and Kevin Duh, 2383–2392. The Association for Computational Linguistics. (Cited on page 34).

Julio C. Rangel, Tarcisio Mendes de Farias, Ana Claudia Sima, and Norio Kobayashi. 2024. SPARQL Generation: an analysis on fine-tuning OpenLLaMA for Question Answering over a Life Science Knowledge Graph. arXiv: 2402.04627 [cs.AI]. (Cited on page 100).

Andrinandrasana David Rasamoelina, Fouzia Adjailia, and Peter Sinčák. 2020. A review of activation function for artificial neural network. In *2020 IEEE 18th world symposium on applied machine intelligence and informatics (SAMI),* 281–286. IEEE. (Cited on page 16).

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922,* (cited on page 74).

Matthew Restall. 2003. A history of the new philology and the new philology in history. *Latin American Research Review* 38 (1): 113–134. (Cited on page 14).

Julio Cesar Rangel Reyes, Tarcisio Mendes de Farias, Ana Claudia Sima, and Norio Kobayashi. 2024. SPARQL generation: an analysis on fine-tuning OpenLLaMA for Question Answering over a Life Science Knowledge Graph. In *15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, SWAT4HCLS 2024, Leiden, The Netherlands, February 20th to 26th, 2024,* edited by Marco Roos, Annika Jacobsen, Andrea Splendiani, M. Scott Marshall, Andra Waagmeester, Leyla Jael García Castro, Katherine Wolstencroft, Kristina M. Hettne, and Rutger A. Vos, 36–45. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 7).

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010a. Modeling Relations and Their Mentions without Labeled Text. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III,* edited by José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, 6323:148–163. Lecture Notes in Computer Science. Springer. (Cited on pages 84, 86).

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010b. Modeling Relations and Their Mentions without Labeled Text. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III,* edited by José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, 6323:148–163. Lecture Notes in Computer Science. Springer. (Cited on page 87).

Dan Roth and Wen-tau Yih. 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning, CoNLL 2004, Held in cooperation with HLT-NAACL 2004, Boston, Massachusetts, USA, May 6-7, 2004,* edited by Hwee Tou Ng and Ellen Riloff, 1–8. ACL. (Cited on pages 84, 86 sq.).

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323 (6088): 533–536. (Cited on pages 15, 18 sq.).

Ivan Rybin, Vladislav Korablinov, Pavel Efimov, and Pavel Braslavski. 2021a. RuBQ 2.0: An Innovated Russian Question Answering Dataset. In *European Semantic Web Conference,* 532–547. Springer. (Cited on page 37).

Ivan Rybin, Vladislav Korablinov, Pavel Efimov, and Pavel Braslavski. 2021b. RuBQ 2.0: An Innovated Russian Question Answering Dataset. In *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings,* edited by Ruben Verborgh, Katja Hose, Heiko Paulheim, Pierre-Antoine Champin, Maria Maleshkova, Óscar Corcho, Petar Ristoski, and Mehwish Alam, 12731:532–547. Lecture Notes in Computer Science. Springer. (Cited on page 49).

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md. Arafat Sultan, and Christopher Potts. 2023. UDAPDR: Unsupervised Domain Adaptation via LLM Prompting and Distillation of Rerankers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023,* edited by Houda Bouamor, Juan Pino, and Kalika Bali, 11265–11279. Association for Computational Linguistics. (Cited on page 5).

Ahmad Sakor, Kuldeep Singh, and Maria-Esther Vidal. 2024. BioLinkerAI: Capturing Knowledge Using LLMs to Enhance Biomedical Entity Linking. In *International Conference on Web Information Systems Engineering,* 262–272. Springer. (Cited on page 65).

Muhammad Saleem, Muhammad Intizar Ali, Aidan Hogan, Qaiser Mehmood, and Axel-Cyrille Ngonga Ngomo. 2015. LSQ: The Linked SPARQL Queries Dataset. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II,* edited by Marcelo Arenas, Óscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, and Steffen Staab, 9367:261–269. Lecture Notes in Computer Science. Springer. (Cited on pages 56 sq.).

Muhammad Saleem, Samaneh Nazari Dastjerdi, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2017. Question Answering Over Linked Data: What is Difficult to Answer? What Affects the F scores? In *BLINK/NLIWoD3@ ISWC.* (Cited on pages 34, 55, 57).

Muhammad Saleem, Gábor Szárnyas, Felix Conrads, Syed Ahmad Chan Bukhari, Qaiser Mehmood, and Axel-Cyrille Ngonga Ngomo. 2019. How Representative Is a SPARQL Benchmark? An Analysis of RDF Triplestore Benchmarks. In *The World Wide Web Conference,* 1623–1633. WWW '19. San Francisco, CA, USA: Association for Computing Machinery. (Cited on page 57).

Manuel Alejandro Borroto Santana, Francesco Ricca, Bernardo Cuteri, and Vito Barbara. 2022. SPARQL-QA enters the QALD challenge. *Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference (ESWC 2022),* (cited on pages 54 sq.).

Mourad Sarrouti, Carson Tao, and Yoann Mamy Randriamihaja. 2022. Comparing Encoder-Only and Encoder-Decoder Transformers for Relation Extraction from Biomedical Texts: An Empirical Study on Ten Benchmark Datasets. In *Proceedings of the 21st Workshop on Biomedical Language Processing,* edited by Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, 376–382. Dublin, Ireland: Association for Computational Linguistics, May. (Cited on page 25).

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question Answering Over Temporal Knowledge Graphs. *arXiv preprint arXiv:2106.01515,* (cited on page 37).

Sonja Schimmler, Bianca Wentzel, Arnim Bleier, Stefan Dietze, Saurav Karmakar, Peter Mutschke, Angelie Kraft, Tilahun A Taffa, Ricardo Usbeck, Zeyd Boukhers, et al. 2023. NFDI4DS infrastructure and services, (cited on page 116).

Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* 40 (D1): D940–D946. (Cited on page 98).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347. arXiv: 1707.06347. (Cited on page 24).

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 7881–7892. Online: Association for Computational Linguistics, July. (Cited on page 101).

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807,* (cited on page 37).

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web* 13 (3): 527–570. (Cited on page 75).

Jiaxin Shi, Shulin Cao, Liangming Pan, Yutong Xiang, Lei Hou, Juanzi Li, Hanwang Zhang, and Bin He. 2020. Kqa pro: A large diagnostic dataset for complex question answering over knowledge base. *arXiv e-prints,* arXiv–2007. (Cited on page 37).

Jiyun Shi, Zhimeng Yuan, Wenxuan Guo, Chen Ma, Jiehao Chen, and Meihui Zhang. 2023. Knowledge-graph-enabled biomedical entity linking: a survey. *World Wide Web (WWW)* 26 (5): 2593–2622. (Cited on pages 30 sq.).

Kanchan Shivashankar, Khaoula Benmaarouf, and Nadine Steinmetz. 2022. From Graph to Graph: AMR to SPARQL. *Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference (ESWC 2022),* (cited on pages 54 sq.).

Yiheng Shu and Zhiwei Yu. 2024a. Distribution Shifts Are Bottlenecks: Extensive Evaluation for Grounding Language Models to Knowledge Bases. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024: Student Research Workshop, St. Julian's, Malta, March 21-22, 2024,* edited by Neele Falk, Sara Papi, and Mike Zhang, 71–88. Association for Computational Linguistics. (Cited on pages 6, 106).

Yiheng Shu and Zhiwei Yu. 2024b. Distribution Shifts Are Bottlenecks: Extensive Evaluation for Grounding Language Models to Knowledge Bases. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop,* edited by Neele Falk, Sara Papi, and Mike Zhang, 71–88. St. Julian's, Malta: Association for Computational Linguistics, March. (Cited on page 100).

Lucia Siciliani, Pierpaolo Basile, Pasquale Lops, and Giovanni Semeraro. 2021. MQALD: Evaluating the impact of modifiers in question answering over knowledge graphs. *Semantic Web,* (cited on page 42).

Lucia Siciliani, Pierpaolo Basile, Pasquale Lops, and Giovanni Semeraro. 2022. MQALD: Evaluating the impact of modifiers in question answering over knowledge graphs. *Semantic Web* 13 (2): 215–231. (Cited on page 51).

Ana Claudia Sima, Tarcisio Mendes de Farias, Maria Anisimova, Christophe Dessimoz, Marc Robinson-Rechavi, Erich Zbinden, and Kurt Stockinger. 2021. Bio-SODA: Enabling Natural Language Question Answering over Knowledge Graphs without Training Data. In *SSDBM 2021: 33rd International Conference on Scientific and Statistical Database Management, Tampa, FL, USA, July 6-7, 2021,* edited by Qiang Zhu, Xingquan Zhu, Yicheng Tu, Zichen Xu, and Anand Kumar, 61–72. ACM. (Cited on pages 6 sq.).

Ana Claudia Sima, Tarcisio Mendes de Farias, Maria Anisimova, Christophe Dessimoz, Marc Robinson-Rechavi, Erich Zbinden, and Kurt Stockinger. 2021. Bio-SODA: Enabling Natural Language Question Answering over Knowledge Graphs without Training Data. In *Proceedings of the 33rd International Conference on Scientific and Statistical Database Management,* 61–72. SSDBM '21. Tampa, FL, USA: Association for Computing Machinery. (Cited on page 99).

Kuldeep Singh, Andreas Both, Dennis Diefenbach, Saedeeh Shekarpour, Didier Cherix, and Christoph Lange15. 2016. Qanary–the Fast Track to Creating a Question Answering System with Linked Data Technology. In *ESWC.* (Cited on page 7).

Kuldeep Singh, Arun Sethupat Radhakrishna, Andreas Both, Saeedeh Shekarpour, Ioanna Lytra, Ricardo Usbeck, Akhilesh Vyas, Akmal Khikmatullaev, Dharmen Punjani, Christoph Lange, Maria-Esther Vidal, Jens Lehmann, and Sören Auer. 2018. Why Reinvent the Wheel: Let's Build Question Answering Systems Together. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018,* edited by Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, 1247–1256. ACM. (Cited on page 37).

Kuldeep Singh, Muhammad Saleem, Abhishek Nadgeri, Felix Conrads, Jeff Z. Pan, Axel-Cyrille Ngonga Ngomo, and Jens Lehmann. 2019. QaldGen: Towards Microbenchmarking of Question Answering Systems over Knowledge Graphs. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II,* edited by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 11779:277–292. Lecture Notes in Computer Science. Springer. (Cited on page 35).

Shruti Singh, Shoaib Alam, Husain Malwat, and Mayank Singh. 2024. LEGOBench: Scientific Leaderboard Generation Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 14598–14613. (Cited on page 115).

Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. 2021. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics* 22 (6): bbab282. (Cited on page 29).

Yiqing Song, Wenfa Li, Guiren Dai, and Xinna Shang. 2023. Advancements in complex knowledge graph question answering: a survey. *Electronics* 12 (21): 4395. (Cited on pages 7, 28).

Daniil Sorokin and Iryna Gurevych. 2017. Context-Aware Representations for Knowledge Base Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017,* edited by Martha Palmer, Rebecca Hwa, and Sebastian Riedel, 1784–1789. Association for Computational Linguistics. (Cited on page 87).

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-QA: A dataset for event-centric question answering over knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management,* 3157–3164. (Cited on page 37).

Claus Stadler, Muhammad Saleem, Qaiser Mehmood, Carlos Buil-Aranda, Michel Dumontier, Aidan Hogan, and Axel-Cyrille Ngonga Ngomo. 2022. LSQ 2.0: A linked dataset of SPARQL query logs. *Semantic Web,* no. Preprint, 1–23. (Cited on page 56).

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-TACRED: Addressing Shortcomings of the TACRED Dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021,* 13843–13850. AAAI Press. (Cited on pages 86 sq.).

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2024. Knowledge graph based agent for complex, knowledge-intensive qa in medicine. *arXiv e-prints,* arXiv–2410. (Cited on page 7).

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 562–572. (Cited on page 37).

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643,* (cited on page 37).

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022,* edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 8472–8487. Association for Computational Linguistics. (Cited on pages 75, 88 sq.).

Thomas Pellissier Tanon, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From Freebase to Wikidata: The Great Migration. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016,* edited by Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, 1419–1428. ACM. (Cited on page 48).

Abbas Al-Thaedan and Marco Carvalho. 2019. Online estimation of motif distribution in dynamic networks. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC),* 0758–0764. IEEE. (Cited on page 103).

Nhuan D. To and Marek Reformat. 2020. Question-Answering System with Linguistic Terms over RDF Knowledge Graphs. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC),* 4236–4243. (Cited on page 41).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971,* (cited on page 25).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971. arXiv: 2302.13971. (Cited on page 74).

Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II,* edited by Claudia d'Amato, Miriam Fernández, Valentina A. M. Tamma, Freddy Lécué, Philippe Cudré-Mauroux, Juan F. Sequeda, Christoph Lange, and Jeff Heflin, 10588:210–218. Lecture Notes in Computer Science. Springer. (Cited on page 37).

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. BioASQ: A challenge on large-scale biomedical semantic indexing and Question Answering. In *2012 AAAI Fall Symposium Series.* (Cited on page 102).

Meimei Tuo and Wenzhong Yang. 2023. Review of entity relation extraction. *Journal of Intelligent & Fuzzy Systems* 44 (5): 7391–7405. (Cited on page 79).

Unger, Christina, Cimiano, Philipp, López, Vanessa, Motta, Enrico, Buitelaar, Paul, and Cyganiak, Richard, eds. 2012. Proceedings of the Workshop on Interacting with Linked Data, Heraklion, Greece, May 28, 2012. Vol. 913. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 50).

Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. 2014a. Question Answering over Linked Data (QALD-4). In *Working notes for CLEF 2014 conference.* (Cited on pages 6, 99).

Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. 2014b. Question Answering over Linked Data (QALD-4). In *CLEF,* 1172–1180. (Cited on page 50).

Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. 2015. Question Answering over Linked Data (QALD-5). In *CLEF.* (Cited on page 50).

Christina Unger, Axel-Cyrille Ngonga Ngomo, and Elena Cabrio. 2016. 6th Open Challenge on Question Answering over Linked Data (QALD-6). In *Semantic Web Challenges: Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers,* 171–177. Cham: Springer International Publishing. (Cited on page 50).

Oleg Ursu, Jayme Holmes, Jeffrey Knockel, Cristian Bologa, Jeremy J. Yang, Stephen L. Mathias, Stuart J. Nelson, and Tudor I. Oprea. 2017. DrugCentral: online drug compendium. *Nucleic Acids Res.* 45 (Database-Issue): D932–D939. (Cited on page 99).

Ricardo Usbeck, Ria Hari Gusmita, Axel-Cyrille Ngonga Ngomo, and Muhammad Saleem. 2018a. 9th Challenge on Question Answering over Linked Data (QALD-9) (invited paper). In *Semdeep/NLIWoD@ISWC.* (Cited on pages 50 sq.).

Ricardo Usbeck, Ria Hari Gusmita, Axel-Cyrille Ngonga Ngomo, and Muhammad Saleem. 2018b. 9th Challenge on Question Answering over Linked Data (QALD-9) (invited paper). In *Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, California, United States of America, October 8th - 9th, 2018,* edited by Key-Sun Choi, Luis Espinosa Anke, Thierry Declerck, Dagmar Gromann, Jin-Dong Kim, Axel-Cyrille Ngonga Ngomo, Muhammad Saleem, and Ricardo Usbeck, 2241:58–64. CEUR Workshop Proceedings. CEUR-WS.org. (Cited on pages 34, 37).

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Felix Conrads, Michael Röder, and Giulio Napolitano. 2018. 8th Challenge on Question Answering over Linked Data (QALD-8) (invited paper). In *Semdeep/NLIWoD@ISWC.* (Cited on pages 50 sq.).

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th Open Challenge on Question Answering over Linked Data (QALD-7). In *Semantic Web Challenges,* edited by Mauro Dragoni, Monika Solanki, and Eva Blomqvist, 59–69. Cham: Springer International Publishing. (Cited on pages 50 sq.).

Ricardo Usbeck, Michael Röder, Michael Hoffmann, Felix Conrad, Jonathan Huthmann, Axel-Cyrille Ngonga-Ngomo, Christian Demmler, and Christina Unger. 2018. Benchmarking Question Answering Systems. *Semantic Web Journal,* (cited on pages 7, 51, 53).

Ricardo Usbeck, Michael Röder, Michael Hoffmann, Felix Conrads, Jonathan Huthmann, Axel-Cyrille Ngonga Ngomo, Christian Demmler, and Christina Unger. 2019. Benchmarking question answering systems. *Semantic Web* 10 (2): 293–304. (Cited on pages 35, 37, 41).

Ricardo Usbeck, Xi Yan, Aleksandr Perevalov, Longquan Jiang, Julius Schulz, Angelie Kraft, Cedric Möller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, Muhammad Saleem, and Andreas Both. 2022. QALD-10 Wikidata Dump, December. (Cited on page 52).

Ricardo Usbeck*, Xi Yan*, Aleksandr Perevalov*, Longquan Jiang*, Julius Schulz*, Angelie Kraft, Cedric Möller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, et al. 2024. Qald-10–the 10th challenge on question answering over linked data: Shifting from dbpedia to wikidata as a kg for kgqa. * Shared first authorship, *Semantic Web* 15 (6): 2193–2207. (Cited on pages 10 sq.).

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Medical Informatics Assoc.* 18 (5): 552–556. (Cited on pages 85 sq.).

Svitlana Vakulenko, Javier David Fernandez Garcia, Axel Polleres, Maarten de Rijke, and Michael Cochez. 2019. Message passing for complex question answering over knowledge graphs. In *Proceedings of the 28th acm international conference on information and knowledge management,* 1431–1440. (Cited on page 40).

Maya Varma, Laurel J. Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-Domain Data Integration for Named Entity Disambiguation in Biomedical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021,* edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 4566–4575. Association for Computational Linguistics. (Cited on page 64).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30. (Cited on pages 3, 20, 22).

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57 (10): 78–85. (Cited on page 28).

Mengru Wang, Jianming Zheng, Fei Cai, Taihua Shao, and Honghui Chen. 2022. DRK: Discriminative Rule-based Knowledge for Relieving Prediction Confusions in Few-shot Relation Extraction. In *Proceedings of the 29th International Conference on Computational Linguistics,* edited by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, 2129–2140. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, October. (Cited on pages 78, 83).

Shouhui Wang and Biao Qin. 2024. No Need for Large-Scale Search: Exploring Large Language Models in Complex Knowledge Base Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy,* edited by Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 12288–12299. ELRA / ICCL. (Cited on page 3).

Xuren Wang, Mengbo Xiong, Yali Luo, Ning Li, Zhengwei Jiang, and Zihan Xiong. 2020. Joint Learning for Document-Level Threat Intelligence Relation Extraction and Coreference Resolution Based on GCN. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom),* 584–591. IEEE. (Cited on pages 89 sq.).

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Comput. Surv.* (New York, NY, USA) 53, no. 3 (June). (Cited on pages 25, 98).

Yihan Wang, Jatin Chauhan, Wei Wang, and Cho-Jui Hsieh. 2023. Universality and limitations of prompt tuning. *Advances in Neural Information Processing Systems* 36:75623–75643. (Cited on page 25).

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a Semantic Parser Overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers,* 1332–1342. The Association for Computer Linguistics. (Cited on page 6).

Yuxing Wang, Kaiyin Zhou, Mina Gachloo, and Jingbo Xia. 2019. An Overview of the Active Gene Annotation Corpus and the BioNLP OST 2019 AGAC Track Tasks. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019, Hong Kong, China, November 4, 2019,* edited by Jin-Dong Kim, Claire Nédellec, Robert Bossy, and Louise Deléger, 62–71. Association for Computational Linguistics. (Cited on pages 85 sq.).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35:24824–24837. (Cited on page 104).

Mengxi Wei, Yifan He, Qiong Zhang, and Luo Si. 2019. Multi-Instance Learning for End-to-End Knowledge Base Question Answering. *CoRR* abs/1903.02652. arXiv: 1903.02652. (Cited on page 48).

Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, and Barend Mons. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3 (March). (Cited on pages 7, 53).

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3 (1): 1–9. (Cited on pages 28, 35).

David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research* 36 (suppl_1): D901–D906. (Cited on page 3).

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020,* edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 6397–6407. Association for Computational Linguistics. (Cited on page 64).

Peiyun Wu, Yunjie Wu, Linjuan Wu, Xiaowang Zhang, and Zhiyong Feng. 2021. Modeling Global Semantics for Question Answering over Knowledge Bases. arXiv: 2101.01510 [cs.AI]. (Cited on page 41).

Peng Wu, Shujian Huang, Rongxiang Weng, Zaixiang Zheng, Jianbing Zhang, Xiaohui Yan, and Jiajun Chen. 2019. Learning Representation Mapping for Relation Detection in Knowledge Base Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* edited by Anna Korhonen, David Traum, and Lluís Màrquez, 6130–6139. Florence, Italy: Association for Computational Linguistics, July. (Cited on page 105).

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32 (1): 4–24. (Cited on page 69).

Xi Yan, Meriem Beloucif and Usbeck, Ricardo, eds. 2022. Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference (ESWC 2022). (Cited on pages 50, 55).

Rui Xing, Jie Luo, and Tengwei Song. 2020. BioRel: towards large-scale biomedical relation extraction. *BMC bioinformatics* 21:1–13. (Cited on pages 85 sq.).

Wang Xu, Kehai Chen, and Tiejun Zhao. 2023. Document-Level Relation Extraction with Path Reasoning. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22 (4): 1–14. (Cited on pages 81, 83).

Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. 2024. Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 18410–18430. Association for Computational Linguistics. (Cited on page 7).

Zhenran Xu, Yulin Chen, and Baotian Hu. 2023. Improving Biomedical Entity Linking with Cross-Entity Interaction. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, no. 11 (June): 13869–13877. (Cited on pages 64 sq.).

Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics,* edited by Emily M. Bender, Leon Derczynski, and Pierre Isabelle, 2145–2158. Santa Fe, New Mexico, USA: Association for Computational Linguistics, August. (Cited on page 73).

Jianhao Yan, Lin He, Ruqin Huang, Jian Li, and Ying Liu. 2019. Relation Extraction with Temporal Reasoning Based on Memory Augmented Distant Supervision. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers),* edited by Jill Burstein, Christy Doran, and Thamar Solorio, 1019–1030. Association for Computational Linguistics. (Cited on page 84).

Xi Yan, Cedric Möller, and Ricardo Usbeck. 2025. Biomedical Entity Linking with Triple-aware Pre-Training. In *Proceedings of the Third International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data (SemTech4STLD 2025), co-located with the Extended Semantic Web Conference (ESWC 2025),* edited by Rima Dessí, Joy Jeenu, Danilo Dessì, Francesco Osborne, and Hidir Aras. To appear. Portoroz, Slovenia: CEUR-WS.org, June. (Cited on pages 10, 12).

Xi Yan, Aida Usmanova, Cedric Möller, Patrick Westphal, and Ricardo Usbeck. 2025. Neuro-Symbolic Relation Extraction. In *Handbook on Neurosymbolic AI and Knowledge Graphs,* 400:550–576. Frontiers in Artificial Intelligence and Applications. IOS Press. (Cited on pages 8, 10, 12).

Xi Yan, Patrick Westphal, Jan Seliger, and Ricardo Usbeck. 2024. Bridging the Gap: Generating a Comprehensive Biomedical Knowledge Graph Question Answering Dataset. In *ECAI 2024,* 1198–1205. IOS Press. (Cited on pages 4, 6, 10 sq., 114).

Mohammad Yani and Adila Alfa Krisnadhi. 2021. Challenges, techniques, and trends of simple knowledge graph question answering: a survey. *Information* 12 (7): 271. (Cited on pages 7, 27).

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. CodRED: A Cross-Document Relation Extraction Dataset for Acquiring Knowledge in the Wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 4452–4472. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November. (Cited on page 91).

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers,* edited by Anna Korhonen, David R. Traum, and Lluís Màrquez, 764–777. Association for Computational Linguistics. (Cited on pages 75, 88 sq.).

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative Knowledge Graph Construction: A Review. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022,* edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 1–17. Association for Computational Linguistics. (Cited on page 73).

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers),* edited by Jill Burstein, Christy Doran, and Thamar Solorio, 2810–2819. Association for Computational Linguistics. (Cited on page 75).

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016a. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 201–206. (Cited on page 37).

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016b. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers.* The Association for Computer Linguistics. (Cited on page 105).

Xiaoyu Yin, Dagmar Gromann, and Sebastian Rudolph. 2019. Neural Machine Translating from Natural Language to SPARQL. *arXiv preprint arXiv:1906.09302,* (cited on page 37).

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-Based Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020,* edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, 4927–4940. Association for Computational Linguistics. (Cited on pages 84, 87).

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. In *Proceedings of the 21st Workshop on Biomedical Language Processing,* 97–109. Dublin, Ireland: Association for Computational Linguistics, May. (Cited on page 68).

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022,* edited by Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, 4038–4048. Association for Computational Linguistics. (Cited on pages 64 sqq.).

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: An entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.* 58 (4): 102563. (Cited on page 89).

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence,* 34:9507–9514. 05. (Cited on page 32).

Chong Zhang, Jiagao Lyu, and Ke Xu. 2023. A storytree-based model for inter-document causal relation extraction from news articles. *Knowledge and Information Systems* 65 (2): 827–853. (Cited on pages 80, 83, 91).

Jiawen Zhang, Jiaqi Zhu, Yi Yang, Wandong Shi, Congcong Zhang, and Hongan Wang. 2021. Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021,* edited by Feida Zhu, Beng Chin Ooi, and Chunyan Miao, 2183–2191. ACM. (Cited on pages 8, 78, 83).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* (Cited on page 101).

Tongxuan Zhang, Hongfei Lin, Michael M Tadesse, Yuqi Ren, Xiaodong Duan, and Bo Xu. 2021. Chinese medical relation extraction based on multi-hop self-attention mechanism. *International Journal of Machine Learning and Cybernetics* 12:355–363. (Cited on pages 88 sq.).

Wei Zhang, Jun Tanida, Kazuyoshi Itoh, and Yoshiki Ichioka. 1988. Shift-invariant pattern recognition neural network and its optical architecture. In *Proceedings of annual conference of the Japan Society of Applied Physics,* vol. 564. Montreal, CA. (Cited on page 19).

Yuan Zhang, Xin Sui, Feng Pan, Kaixian Yu, Keqiao Li, Shubo Tian, Arslan Erdengasileng, Qing Han, Wanjing Wang, Jianan Wang, et al. 2023. BioKG: a comprehensive, large-scale biomedical knowledge graph for AI-powered, data-driven biomedical research. *bioRxiv,* (cited on page 99).

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017,* edited by Martha Palmer, Rebecca Hwa, and Sebastian Riedel, 35–45. Association for Computational Linguistics. (Cited on page 87).

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alex Smola, and Le Song. 2018. Variational Reasoning for Question Answering with Knowledge Graph. In *AAAI.* (Cited on page 37).

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Xue Mengge, Tingwen Liu, and Li Guo. 2021. From What to Why: Improving Relation Extraction with Rationale Graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 86–95. Online: Association for Computational Linguistics. (Cited on pages 8, 80, 83).

Qi Zhao, Hongyu Yang, Qi Song, Xin-Wei Yao, and Xiangyang Li. 2025. KnowPath: Knowledge-enhanced Reasoning via LLM-generated Inference Paths over Knowledge Graphs. *CoRR* abs/2502.12029. arXiv: 2502.12029. (Cited on page 6).

Qihui Zhao, Tianhan Gao, and Nan Guo. 2023. Document-level relation extraction based on sememe knowledge-enhanced abstract meaning representation and reasoning. *Complex & Intelligent Systems* 9 (6): 6553–6566. (Cited on pages 8, 79, 83).

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers. *ACM Comput. Surv.* 56 (11): 293:1–293:39. (Cited on page 32).

Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, and Jian Wang. 2016. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinform.* 32 (22): 3444–3453. (Cited on page 84).

Weiguo Zheng and Mei Zhang. 2019. Question Answering over Knowledge Graphs via Structural Query Patterns. *ArXiv* abs/1910.09760. (Cited on page 38).

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An interpretable reasoning network for multi-relation question answering. *arXiv preprint arXiv:1801.04726,* (cited on page 37).

Jianyun Zou, Min Yang, Lichao Zhang, Yechen Xu, Qifan Pan, Fengqing Jiang, Ran Qin, Shushu Wang, Yifan He, Songfang Huang, et al. 2021. A Chinese Multi-type Complex Questions Answering Dataset over Wikidata. *arXiv preprint arXiv:2111.06086,* (cited on page 37).

Jianyun Zou, Min Yang, Lichao Zhang, Yechen Xu, Qifan Pan, Fengqing Jiang, Ran Qin, Shushu Wang, Yifan He, Songfang Huang, and Zhou Zhao. 2021. A Chinese Multi-type Complex Questions Answering Dataset over Wikidata. *CoRR* abs/2111.06086. arXiv: 2111.06086. (Cited on pages 48 sq.).