Knowledge Graph Efficient Construction: Embedding Chain-of-Thought into LLMs

Jixuan Nie
Beijing Information Science & Technology University
nie00331@163.com

Xia Hou*
Beijing Information Science & Technology University
HouXia@bistu.edu.cn

Wenfeng Song
Beijing Information Science & Technology University
songwenfenga@163.com

Xuan Wang
Beijing Information Science & Technology University
2022020634@bistu.edu.cn

Xinyu Zhang
Beijing Information Science & Technology University
zhangxinyu1@bistu.edu.cn

Xingliang Jin
Beijing Information Science & Technology University
xingliangjin276@gmail.com

Shuozhe Zhang
Beijing Information Science & Technology University
Z3607629726@163.com

Jiaqi Shi
Beijing Information Science & Technology University
3305191047@qq.com

ABSTRACT

Large Language Models (LLMs) are extensively utilized for extracting key information from unstructured data to construct Knowledge Graph (KG) due to their advanced language comprehension and generation capabilities. However, the diversity in natural language leads to varied relational expressions in the extracted triples for the data with similar meanings, often necessitating substantial manual annotation to ensure quality. We present a novel method that employs an ontology to define domain-specific knowledge, thereby guiding LLMs to extract more standardized triples. By constructing Chain-of-Thought (CoT) prompts that emulate the human cognitive process of understanding triple knowledge in unstructured data, we guide the model to extract higher-quality triples. Our approach significantly reduces the diversity of relational expressions, lowering the difficulty and workload associated with building domain-specific KG. Experiments conducted on the TekGen dataset demonstrate that our method can markedly decrease the diversity of relational expressions while preserving accuracy. We also discuss potential future research directions.

VLDB Workshop Reference Format:
Jixuan Nie, Xia Hou, Wenfeng Song, Xuan Wang, Xinyu Zhang, Xingliang Jin, Shuozhe Zhang, and Jiaqi Shi. Knowledge Graph Efficient Construction: Embedding Chain-of-Thought into LLMs. VLDB 2024 Workshop: LLM+KG.

1 INTRODUCTION

Knowledge Graph (KG) have been widely applied in traditional fields such as search engines [7], e-commerce [29], and social media [2, 23], expanded into multiple industries including finance [4], healthcare [33], and education [3]. KG is a knowledge base that consists of three elements: head entities, tail entities, and the relationships between them. Extracting these knowledge from unstructured data is an important foundation for constructing KG. Currently, commonly used knowledge extraction methods rely on large amounts of annotated data and utilize deep learning algorithms [11, 28] to achieve extraction. That is, given the types of entities and relationships to be extracted, models are trained through supervised learning to obtain specific entities and relationships. For instance, a classic algorithm for entity extraction is BiLSTM+CRF [6], while representative algorithms for relationship extraction include Convolutional Neural Network [17]. Such methods rely heavily on extensive manual annotation, which is labor-intensive. Additionally, when targeting specific domains, they require domain expertise, further complicating the annotation process. Although there are methods such as distant supervision [21], and transfer learning [27] to alleviate the issue of limited annotated data, their effectiveness remains imperfect for specific vertical domains. Therefore, the advancement of domain-specific KG heavily depends on the participation of industry experts, with data annotation emerging as a significant obstacle to their progress.

Recently, Large Language Models (LLMs) have demonstrated remarkable performance in tasks such as natural language understanding, text generation [1, 18]. These models leverage their deep neural network architectures to capture complex patterns and implicit relationships in language, thereby exhibiting immense potential in handling unstructured data. Through zero-shot learning [8], LLMs can directly extract triple knowledge from unstructured data based on instructions, without the need for further model training, thus aiding in the construction of KG.

However, the diversity of natural language expressions leads to instability in the output results of generative models. For example, the text "Stuff Stephanie in the Incinerator (originally titled In Deadly Heat) is a 1989 horror-comedy written and directed by Don Nardo." describes the relationship between a movie and its director. Different vocabulary such as "director" or "directed by" can be used

*Corresponding author

Proceedings of the VLDB Endowment. ISSN 2150-8097.

to express this relationship. When using LLMs for triple extraction, obtaining triples for similar sentences may result in variations, such as [Stuff Stephanie in the Incinerator|directed_by|Don Nardo] and [Stuff Stephanie in the Incinerator|director|Don Nardo]. When constructing KG, it is necessary to align semantically equivalent but differently expressed content, leading to certain errors that affect the quality of the KG.

Additionally, LLMs may also extract many unnecessary triples from unstructured data when building domain-specific KG, which is often due to the model's lack of understanding of domain knowledge. Discrimination is required to filter out these triples, resulting in additional workload during KG construction.

To address the aforementioned issues, we propose an innovative approach that utilizes ontology [22] to guide large models with specialized knowledge and employs Chain-of-Thought (CoT) [25] to mimic the human thought process of understanding triple knowledge from unstructured data. We will use a simplified ontology conceptual model as the basis, which focuses on providing domain knowledge without involving complex ontology axioms. In our method, we only include the mapping between entity types and relationship types. This aims to guide LLMs to extract higher-quality triples, thereby reducing the difficulty and workload of domain-specific KG construction.

The proposed method integrates ontology into the triple extraction process, making it an integral part of the CoT prompts. This ensures that the model's reasoning and extraction steps are aligned with the predefined ontology, guiding the model to extract triples that conform to the specified domain knowledge. CoT prompts direct the model through the extraction process in a step-by-step manner, including entity discovery, relationship extraction, and ontology alignment. The ontology serves as a reference throughout the process, ensuring that the extracted triples adhere to the defined entity types and relationships. This combined approach aims to reduce the diversity of relational expressions in the extracted triples and ensure their consistency with the predefined ontology, making the knowledge graph construction process more efficient and less reliant on manual annotation.

2 RELATED WORK

In this section, we delve into the application of LLMs in KG construction, exploring methodologies that incorporate ontology into the construction process. We examine approaches leveraging the zero-shot reasoning capabilities of LLMs and summarize the insights gathered from these methods.

2.1 Applications of LLMs in Knowledge Graph Construction

LLMs showed their great potential in various NLP tasks [16, 30]. In the field of KG construction, LLMs demonstrated significant capabilities and broad potential in executing various construction tasks [24]. For instance, the RECENT [11], by leveraging entity types to limit candidate relations, showcased the effectiveness of LLMs in KG construction. The KICGPT [26] combined LLMs with a triadic-based KG completion retriever, providing a new solution for the task of KG completion. Moreover, methods based on OpenIE [13] used LLMs to construct KG, integrating syntactic structures to more effectively represent knowledge within text, thus building richer and more coherent KGs. The GenKGC [28] transformed KG completion into a sequence-to-sequence generation task and utilized pre-trained language models for generation, achieving excellent performance in KG completion through relation-guided and entity-aware hierarchical decoding. Grapher [14] constructed KG through multi-stage design, using pretrained language models and classification/generation models. The DREEAM [12] approach aimed to improve evidence retrieval (ER) in document-level relation extraction (DocRE). DREEAM directly supervised the attention module of the DocRE system, focusing the model's attention on evidence sentences related to entity pairs, thereby enhancing both relation extraction and evidence retrieval. Curriculum-RE [19] proposed a curriculum learning-based sentence-level relation extraction method, which divided data by difficulty and then learned them in order of difficulty to improve the performance of relation extraction. Some of above works limited the extraction scope of the model by utilizing lists of entity types or relationships. This approach, to some extent, enhanced the accuracy of information extraction by focusing the model's attention on specific entities or relationships. However, a high-quality triple not only needed to contain correct entities and relationships, but more importantly, it should have ensured the correspondence between entities and relationships, meaning entities and relationships matched each other correctly. Additionally, the diversity of relational expressions was also a direction that needed to be considered in the process of information extraction, and these works that utilized LLMs for automated KG construction did not address this issue.

2.2 Applications of ontology in Knowledge Graph Construction

The ontology, as a comprehensive structural system, provided a solution for the correspondence between the aforementioned entities and relationships. The launch of the Text2KGBench benchmark [15] provided an important platform for evaluating language models' ability to generate KG that conformed to a given ontology from text. The Extract-Define-Canonicalize method [32] used LLMs to construct KG, employing open information extraction, ontology-definition-based graph schema, and relation-definition-based graph schema, as well as standardizing triples to eliminate redundancy and ambiguity, thereby enhancing the efficiency and quality of KG construction, essentially automating the process of expert annotation. Although these works demonstrated the feasibility of ontology application, they did not fully utilize the internal structure of the ontology, especially in terms of the correspondence between entities and relationships. Ontologies typically included hierarchical structures and relationships between entities, which could provide deeper understanding and richer semantic representations. There is currently no research focusing on gaining a thorough understanding of the ontology while using LLMs to assist in the construction of knowledge graphs.

2.3 Prompt engineering for Large Language Models

Prompt engineering is a novel field that focused on creating and refining prompts to maximize the effectiveness of LLMs across various applications and research areas [5, 20]. Generated knowledge prompting consisted of generating knowledge from a language model, then providing the knowledge as additional input when answering a question [9]. Automatic prompt engineer (APE) [34] was an automatic prompt generation method to improve the performance of LLMs. The "CoT" prompting method [25] included a series of intermediate reasoning steps in the prompts, assisting language models in deriving final answers step by step. In complex reasoning tasks such as mathematical problems, common-sense reasoning, and symbolic reasoning, CoT prompts significantly improved the performance of LLMs compared to standard prompts. The zero-shot prompting technique [8] prompted chained thinking by adding "Let's think step by step" before each answer. The paper showed that large-scale language models possessed underutilized general zero-shot reasoning abilities, beyond just few-shot learning capabilities. This provided important insights into the advanced multi-task zero-shot cognitive abilities hidden within language models [10].

These research findings clearly indicated that using LLMs for KG construction was not only technically feasible but also held great potential for improving construction efficiency and significantly reducing labor costs [31]. In this process, ontology, as the core basis for knowledge extraction and structuring, presented a promising direction for development. The advanced natural language processing capabilities of LLMs enabled them to automatically identify and extract entities, relationships, and attributes from text, which were key components in building KG. There was no related work on using CoT prompts for LLMs applied to information extraction.

3 METHODOLOGY

In this section, we introduce our novel method which innovatively applies ontology and CoT to information extraction. This method leverages the synergistic powers of ontology and LLMs to enhance the extraction of triples that adhere to predefined specifications. Existing methods have demonstrated that ontologies can be used to guide large language models in information extraction, but their understanding of the content of the ontologies is still insufficient. We construct the process of triple extraction as a CoT prompt, and integrate the content of the ontology into this process. By employing this method, we aim to decrease dependency on expert annotation during the KG construction phase, thereby streamlining the KG construction process and ensuring the integrity and precision of the captured knowledge.

3.1 Framework

Our approach utilizes CoT and ontology, which steers the model through the triple extraction process. Figure 1 contrasts our proposed methodological framework with the conventional approach that relies exclusively on LLMs for triple extraction. The A part of the figure illustrates the extraction of triples from natural language text using LLMs alone. This method depends on LLMs to comprehend natural language text and produce a set of triples. The outcomes of this approach display a variety of expressions for each relation, necessitating time-consuming manual annotation by experts and knowledge alignment for KG construction. The B part of the figure denotes our proposed method. We use CoT prompts to guide the model through the process of triple extraction and integrate a set of ontology concepts into this process. The ontology concepts here focus more on ensuring that the entities at both ends of a relationship are within the list of entity types. This manifests in the overall structure such that all entity relationship structures belonging to the domain should be subgraphs of the ontology graph structure. This strategy yields more consistent and precise relational expressions, thus reducing the reliance on expert annotation to a certain degree. In the following, we detail the key components of our approach.

3.2 Ontology

In our approach, ontology serves a crucial role as a structured form of knowledge representation. It not only defines entity types within the domain with clarity but also describes the relationships among them. This integrated framework constitutes an valuable prior knowledge base, which both guides and confines the model's behavior in recognizing and extracting information for KG.

Through the integration of ontological knowledge, we are able to embed rule-based prior knowledge into the model, which is essential for managing unstructured data. This integration allows the model to conduct identification and extraction within predetermined frameworks, substantially narrowing the search space and enabling the model to concentrate on the concepts and relationships as defined by the ontology.

Ontologies should encompass entities and relationships with a high degree of consistency, which means that in the process of constructing a knowledge graph, the entities and their associations should strictly adhere to the categories and relationship frameworks defined in the ontology. Specifically, for any given relationship, the entity types at both ends must strictly conform to the predefined entity type list in the ontology. This consistency ensures the accuracy and reliability of the knowledge graph because it follows a clear, predefined conceptual model.

For example, if an ontology defines "book" and "author" as entity types, and "writes" as a relationship between them, then any "writes" relationship extracted in the knowledge graph should connect two entities, one of which is a "book" and the other an "author". Any relationship that does not match this type of pairing would be considered inconsistent with the ontology and could lead to errors or confusion in the knowledge graph.

The ontology functions as a "filter" or "compass" for the model, guiding it to recognize and extract pertinent entities and relationships while ignoring extraneous information. Given that our research focus is on the precise extraction of target triples from natural language texts in specific domains, we have narrowed our scope to the correct correspondence between entities and relationships. In this process, we have chosen not to extend our attention to the more complex axiomatic levels within ontology, but instead to concentrate on ensuring the semantic accuracy of the extracted entities and their corresponding relationships. Furthermore, the ontology serves as an evaluation criterion for performing consistency checks

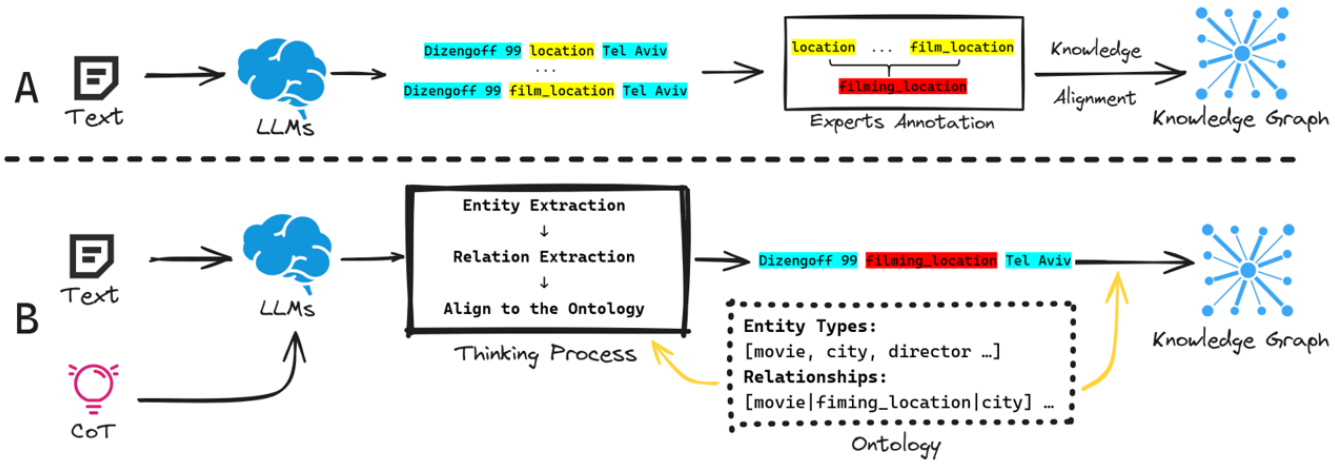

Figure 1: Comparison between our method Extraction with Thinking and the method using only LLMs to extract triples.

and quality assessments on the results produced by the model. By contrasting the model's outputs with the ontological specifications, we can swiftly pinpoint and rectify any errors or discrepancies in the extraction process.

3.3 CoT

We innovatively integrate the CoT method, which performs well in the question-answering field, into the process of extracting triples using LLMs. Our aim is to guide the model to extract triples step by step, mainly including three parts: entity discovery, extracting relationships between entities, and corresponding with a predefined ontology. The design philosophy of CoT is similar to the human thought process. When humans extract triples that conform to the ontology definition from natural language texts, they first identify entities in the text, then analyze the relationships between these entities, and finally correspond the identified entities and relationships to the definitions in the ontology. Using models to complete these three tasks separately would result in computational loss during the intermediate process. CoT incorporates these three steps into prompts, guiding LLMs (Large Language Models) to perform information extraction in the same steps, thus simulating the human thought process. LLMs can better understand domain-specific concepts and relationships, thereby extracting triples that more closely align with the ontology definition.

We have developed CoT prompts that efficiently direct the model's reasoning and extraction processes in line with predefined logical structures. Our main objective is to extract triples that fulfill specific criteria, rather than explicitly revealing the extraction procedure. To this end, we employ a zero-shot prompting approach for the CoT component, enabling the model to engage in reasoning and produce outputs based on the given prompts, without requiring any further training.

We utilize the ICL prompting format, which is an acronym for Instruction, Context, and Learning. This format mimics a dialogue-like architecture, crafting instructions for the system role, user example inputs, and expected output examples. These elements are conveyed in a conversational style as inputs to LLMs. This scenario-based learning approach assists the model in comprehending task requirements more thoroughly and produces high-quality outputs in real-world applications.

The Figure 2 depicts the process we aim to use for guiding the model's reasoning through CoT and its integration with the ontology. Initially, the system role provides the Instruction, which includes information about the ontology and the CoT process. In the CoT section is where we design prompts to direct the model to first recognize entities, then to discover relationships between these entities within the sentence, and finally to align with the ontology's definitions. The Ontology section presents a collection of entities and relationships within the domain. We express these relationships using a triple format, where each relationship's endpoints correspond to entities from the entity list, thereby showcasing the ontology's overall structure. In the Example section, the User and Expected Output roles are used to create a sample dialogue for triple extraction within the domain, demonstrating the desired format for the model's output. This example only specifies the format requirements for the output, excluding any content from the CoT. In the Sentence section, the User role inputs a sentence. In the Thinking Process section, the LLMs engage in a step-by-step reasoning process guided by the instructions and examples from the previous sections. Ultimately, the LLMs outputs triples in the given format.

We aim for LLMs to present key information as concisely as possible, specifically by displaying triples in the format "[Entity1 | Relation | Entity2]". This format not only contains the fundamental components of constructing a KG clearly but also reduce the reasoning process. Our objective is to have the model concentrate on extracting essential information in this simplified output structure, without the distraction of elaborate reasoning details. To moderate the influence of example content on the model's output, we employ an invariant example strategy, using just a single example for each domain. This implies that the example content remains consistent, regardless of the type of relationship present in the input sentence. The reason behind this approach is that in practical applications,

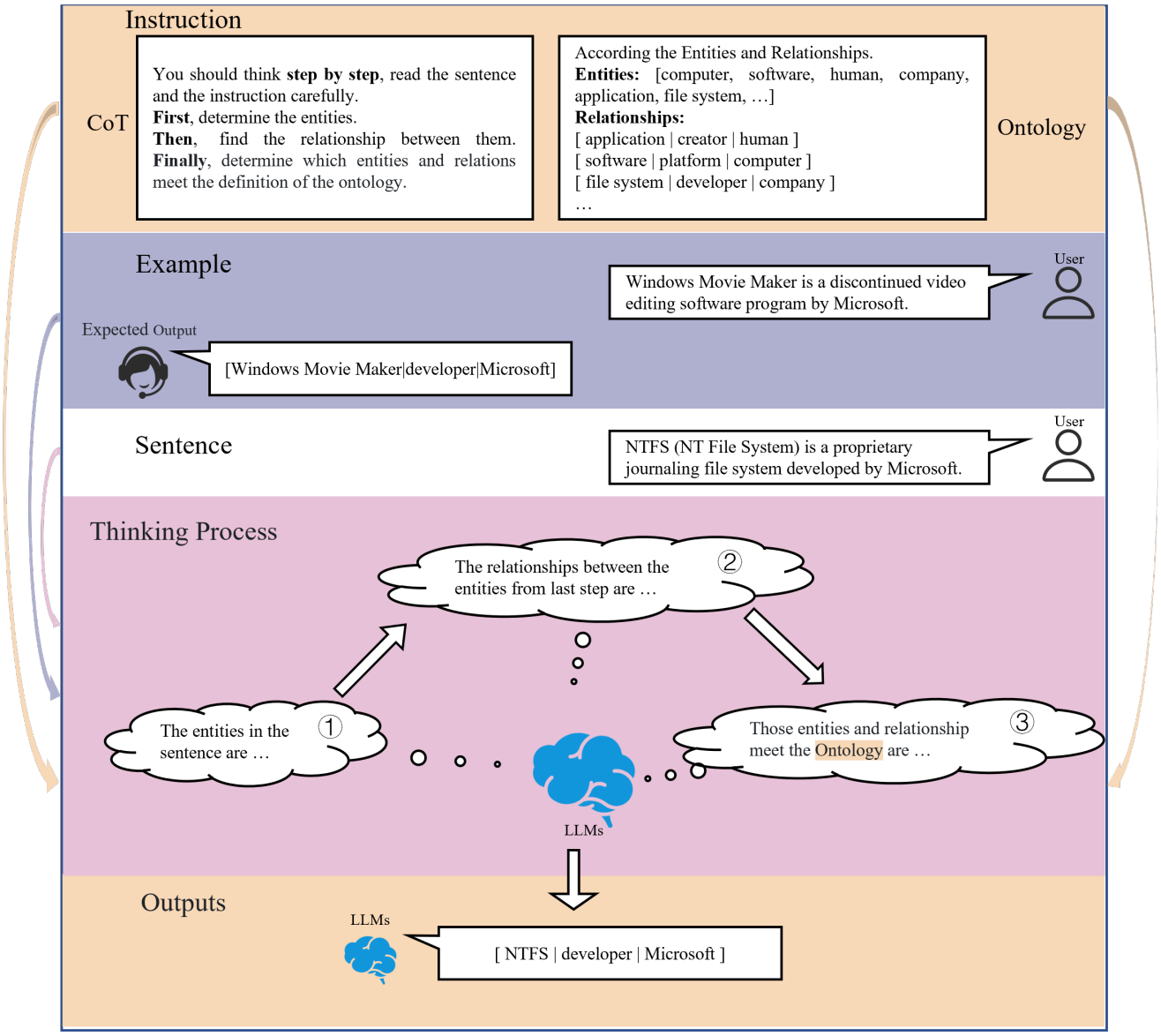

Figure 2: An example of CoT_Ontology

it's unlikely to have an ideal example for every text extraction scenario, and creating comprehensive examples can be labor-intensive. Hence, we focus the extraction process more on the ontology and the LLMs' capabilities, rather than relying on examples tailored to specific sentences. With the invariant example strategy, we seek to decrease the model's dependence on example content, prompting it to draw more on its contextual understanding and application of the ontology. This method not only enhances the model's ability to generalize across diverse domain data but also ensures more precise extraction of the required triple information when dealing with real-world data.

In summary, our method guides LLMs to efficiently and accurately extract more specific relationships from unstructured data by offering a simplified output format and a single example selection. The deployment of this technique is anticipated to significantly advance the automation of KG creation and to deliver robust knowledge services across various sectors. Our vision is to realize the automation and sophistication of KG construction through this innovative approach, equipping industries with more detailed and streamlined knowledge services.

4 EXPERIMENTS

In Section subsection 4.1, we detail the dataset employed for our experiments. Section subsection 4.2 outlines the key experimental parameters and setup. Section subsection 4.3 presents the assessment of conventional metrics, encompassing Precision, Recall, and

F1 Score for relational tasks. Section subsection 4.4 delves into the evaluation of the specificity in relational expression. Concludingly, Section subsection 4.5 offers a thorough analysis and discussion of the experimental outcomes.

4.1 Dataset

In the experimental segment of our research, we utilize the TekGen dataset [15]. This dataset covers 10 distinct sub-domains, offering a wealth of diversity. Crucially, it defines a comprehensive set of ontologies and their associated relationships for each domain. In such a scenario, our results might suffer from a lack of credibility. The dataset is constructed through a specialized process: initially, a predefined set of entity types and relationships are established, and a collection of triples is created based on this framework. These triples are then aligned with natural language paragraphs from the corpus, thereby compiling the sentences that constitute the dataset. Once sentences are identified, any other relationships in the corpus connected to these sentences are scrutinized, and if they belong to the intended ontology, they are integrated into the dataset as well. To enhance the portability and generalizability of our method across various domains, we don't make any domain-specific modifications to the dataset for our model. Furthermore, since our approach does not require model training, we combine the training and testing subsets of the dataset to create a unified experimental dataset. We aim for this method to be applicable to the automatic construction of domain-specific knowledge graphs without the need for domain-specific training. The merged dataset's statistics are presented in Table 1.

The primary reason we choose the TekGen dataset is its pre-defined simple ontology for each domain. The TekGen dataset provides pre-defined entity types and relationship types for each domain, which serves as a valuable benchmark and validation standard for our method. If we were to use another dataset, we would need to define the types of entities and relationships within the dataset ourselves. This would not only increase the amount of preparatory work but could also lead to inconsistencies and biases in the definitions, thus undermining the advantages of our method. More importantly, the advantages of answers obtained under the guidance of an ontology defined by ourselves are not fully convincing. Due to the lack of authoritative and widely recognized ontologies as a reference, we cannot be sure that the ontologies we define can accurately reflect the knowledge and structure of specific domains, thereby affecting the accuracy and reliability of the extraction results.

4.2 Experiments Settings

All of our experiments were conducted on a pair of NVIDIA Geforce RTX3090 24G graphics cards. Additionally, during the experimental process, we observed a significant decrease in inference speed as the dataset progressed, and in some cases, the system would even freeze when dealing with particularly complex queries. Consequently, in order to enhance the speed and efficiency of inference, we adopt the vllm accelerated reasoning framework and integrate the powerful ChatGLM3-6B 128K model for all experiments. The model has 6 billion parameters and can process complex text data. We choose the 128K context length model because we want the model to better understand the prompts of long context. When reasoning with models with shorter context lengths, the ICL group always takes less time than the Context group, indicating that models with shorter context lengths understand Context format prompts less well than the ICL group. We do not want such gaps to arise due to model capabilities. Therefore, we choose to use a 128k context length model, which is sufficiently large for both ICL format and Context format, reducing the differences brought about by prompt length. In the process of model generating triples, we use the beam_search algorithm, which is a common sequence generation strategy that can explore multiple possible output paths when generating text. We set the number of output results to five times for subsequent evaluation, in order to get more convincing scores. In the subsequent quantitative indicators, the score of each sentence corresponds to the average score of the 5 output results. In light of our aim to demonstrate the potential advantages of reducing the diversity of relational expression through a series of meticulously designed experiments, current experimental research on the TekGen dataset has solely focused on accuracy as the measurement criterion. Moreover, these studies exhibit significant discrepancies in terms of experimental environment, model selection, parameter configuration, whether fine-tuning is applied, and the specific content of the prompts. To address this, we have decided to establish six groups of comparative experiments under completely consistent environmental conditions. This approach is intended to ensure the reliability and comparability of the experimental results, thereby more deeply uncovering the effectiveness of the method we propose.

- **Context_base** The prompt contains only the instruction to extract triples.
- **Context_Ontology** The ontology concepts and relationship list are added to the prompt.
- **Context_CoT_Ontology** A method combining CoT and ontology.
- **ICL_base** The prompt contains only the instruction to extract triples.
- **ICL_Ontology** The ontology concepts and relationship list are added to the prompt.
- **ICL_CoT_Ontology** A method combining CoT and ontology.

As Figure 3 shows, "context" refers to the input of instructions, examples, and prior knowledge in a hole part to the model, with "ICL" having been previously mentioned. We meticulously designed the comparative experiments between the Context group and the ICL group with the intention that: due to the additional computational overhead introduced by CoT_Ontology compared to traditional Ontology. To validate the practical effectiveness of the CoT (Chain of Thought) mechanism, we adopted two different methods to deliver prompts to the model while keeping the content of the prompts consistent. Through this design of control experiments, our aim is to highlight the key role played by the CoT and the benefits it brings within the framework of our proposed method. Additionally, in Section subsection 4.3 and subsection 4.4, we only conduct internal comparisons in Context or ICL, while the inter-group comparison between Context and ICL will be conducted in Section subsection 4.5.

Table 1: Statistics on the number of ontologies, relationships, and sentences contained in various domains within the integrated TekGen dataset

Type	Movie	Music	Sport	Book	Military	Computer	Space	Politics	Nature	Culture
Ontology	12	13	20	20	13	15	15	13	14	15
Relation	15	13	10	12	8	4	7	9	13	8
Sentence	1960	1571	1129	1271	528	524	468	489	1094	363

- Your task is to extract the knowledge graph triplet from a given sentence. The relationships in the triplet you are extracting should be in the Ontology Relationships list, and if they are not in the list, they should not be extracted. You should think step by step, you should first determine the entities first and then find the relationship between them, finally, map the extracted relationships to the Ontology Relationships list after read the sentence and the instruction carefully. You can't have a relationship name that doesn't exist in the Ontology Relationships in your answer.
- CONTEXT:
  - Ontology Concepts: {ontology list}
  - Ontology Relationships: {relation list}
  - Example Input: Extract the triplet from the sentence where the relationship should exist in the Ontology Relationships.
  - Ontology Relationships: {relation list}.Sentence: {sample_sentence}
  - Example Output: {sample_triples}
  - Test Input: Extract the triplet from the sentence where the relationship should exist in the Ontology Relationships.
  - Ontology Relationships: {relation list}.Sentence: {sentence}
  - Test Output:

Figure 3: Example of Context Prompt

4.3 Traditional Evaluation

Firstly, we employ the conventional evaluation approach to quantitatively assess the precision, recall, and F1 score of the extracted triples in comparison to the dataset's standard answers. Table 2 shows the indicators for each field in the Context group, while Table 3 displays the indicators for each field in the ICL group. Across all three metrics, the ICL_Ontology and ICL_CoT_Ontology groups show superior performance to the ICL_Base group, suggesting that the incorporation of ontology and CoT can significantly enhance the precision of the information extraction process. Furthermore, in the precision metric evaluation, the ICL_CoT_Ontology group achieves a higher score than the ICL_Ontology group in eight domains. In the recall metric evaluation, the ICL_CoT_Ontology group outperforms the ICL_Ontology group in five domains. Similarly, in the F1 score evaluation, the ICL_CoT_Ontology group achieves a higher score than the ICL_Ontology group in seven domains. However, the trend does not hold for the Context group, which we will examine further in section subsection 4.5. Our method, grounded in ICL, demonstrates that our CoT approach can effectively enhance the accuracy of triple extraction.

As previously mentioned, the construction method of this dataset doesn't disambiguate pronouns in the sentences, which led to a frequent occurrence of pronouns such as "The movie" or "The film" within many sentences. When our model extracts information from these sentences, it often fails to accurately identify the specific entities that these pronouns represent. Consequently, the model employing the CoT method tends to refrain from extracting relations that involve entities without clear references, leading to a lower recall value for the CoT_Ontology group compared to the Ontology group in certain domains.

Another trend is that, for both the Context and ICL groups, the recall indicator scores across all 10 domains of the experiments are higher than their corresponding precision scores. We meticulously analyzed the model's extraction results and the dataset construction process to uncover some key insights.

In the dataset construction process, there is a scenario where relationship expressions in the corpus fall outside the ontology's definition but can be mapped to relationships that are included in the ontology. In these instances, the relationships are not incorporated into the dataset, resulting in a degree of incompleteness in the dataset's reference answers. This manifests in the experiment

Table 2: Statistics on Precision, Recall and F1 of Context Group. A means Base, B means Ontology, C means CoT_Ontology. The bold numbers represent the best scores.

Type	Precision			Recall			F1		
	A	B	C	A	B	C	A	B	C
movie	0.1301	**0.3068**	0.3062	0.1582	**0.3279**	0.3221	0.1428	**0.3170**	0.3140
music	0.1402	0.3040	**0.3049**	0.1842	**0.3451**	0.3422	0.1592	**0.3232**	0.3225
sport	0.1308	0.2379	**0.2404**	0.1522	**0.2623**	0.2613	0.1407	0.2495	**0.2504**
book	0.2457	**0.2808**	0.2776	**0.3442**	0.3235	0.3133	0.2867	**0.3006**	0.2944
military	0.2886	**0.5275**	0.5232	0.5278	**0.6515**	0.6459	0.3732	**0.5829**	0.5781
computer	0.3160	0.3697	**0.3716**	0.4641	0.5064	**0.5073**	0.3760	0.4274	**0.4290**
space	0.1746	**0.5459**	0.5378	0.3284	**0.6895**	0.6716	0.2280	**0.6094**	0.5973
politics	0.0232	0.1304	**0.1319**	0.0277	0.1383	**0.1431**	0.0253	0.1343	**0.1373**
nature	0.1205	**0.2729**	0.2657	0.1160	**0.2763**	0.2678	0.1182	**0.2746**	0.2668
culture	0.1086	**0.3102**	0.2883	0.1102	**0.3102**	0.2887	0.1094	**0.3102**	0.2885

Table 3: Statistics on Precision, Recall and F1 of ICL Group. A means Base, B means Ontology, C means CoT_Ontology. The bold numbers represent the best scores.

Type	Precision			Recall			F1		
	A	B	C	A	B	C	A	B	C
movie	0.1042	**0.3484**	0.3457	0.1709	**0.5870**	0.5734	0.1295	**0.4372**	0.4313
music	0.0920	**0.2959**	0.2949	0.1646	**0.4963**	0.4741	0.1181	**0.3707**	0.3637
sport	0.0465	0.2118	**0.2155**	0.0869	**0.3918**	0.3757	0.0606	**0.2749**	0.2739
book	0.1935	0.3150	**0.3284**	0.3727	0.5408	**0.5501**	0.2547	0.3982	**0.4113**
military	0.1946	0.4741	**0.5088**	0.4307	**0.7965**	0.7950	0.2680	0.5944	**0.6205**
computer	0.2609	0.3204	**0.3346**	0.4484	0.5719	**0.5753**	0.3299	0.4107	**0.4231**
space	0.1926	0.5117	**0.5338**	0.3323	0.7434	**0.7449**	0.2439	0.6061	**0.6219**
politics	0.0298	0.2687	**0.2715**	0.0430	0.3932	**0.3975**	0.0352	0.3192	**0.3226**
nature	0.0812	0.2174	**0.2256**	0.0793	**0.3094**	0.2981	0.0803	0.2554	**0.2568**
culture	0.0876	0.2928	**0.3248**	0.0876	0.2931	**0.3251**	0.0876	0.2930	**0.3249**

as a mismatch in granularity and direction between our model's output and the standard answers. During the experiment, since we don't employ approximate matching for each sentence, we provide identical input for example content and prior knowledge across sentences. This approach means that the model had to autonomously identify and extract all relationships it perceived to exist, and then map these to the relationship list we provided. Consequently, our results exhibit a higher recall value compared to the precision value.

This further highlights the challenges in dataset construction and answer annotation. Future work can focus on addressing these issues to improve the overall performance of the model in the task of triple relationship extraction.

4.4 Specificity of relationship expression

In this experiment, we focus on assessing the specificity of relationship expressions extracted from the text. To thoroughly evaluate this attribute, we employ two complementary metrics: the first metric is the count of unique relationship types that the model extracts for each domain dataset. This metric reflects the model's capability to consolidate various expressions of a single relationship into a unified representation. The second metric is OC (Ontology Conformance), which is calculated based on the relationships defined within the ontology and the proportion of these relationships that the model successfully extracts. This metric, inspired by the TEXT2KGBENCH paper [15], indicates the model's ability in aligning the extracted relationships with those predefined in the ontology. A decrease in the number of relationship types coupled with an increase in the OC value suggests that the extracted relationship expressions are more precise and align more closely with the ontology's definitions.

4.4.1 Kinds of Relationship Expressions. The evaluation of the number of relationship expressions involves counting the total count of unique relationship types across all five extraction results for each sentence within every domain. The experimental findings are detailed in Table 4. In both the Context and ICL groups, we noted that the CoT_Ontology group markedly reduced the variety of relationship types in the extracted triples across the ten domain datasets when compared to the base group, which lacks any scope limitations or supplementary prompts. Specifically, the Context_CoT_Ontology group exhibited a lower count of relationship types in 8 out of the 10 domains when contrasted with the corresponding domains in the Context_Ontology group. Similarly, the ICL_CoT_Ontology group showed a reduction in the number of

relationship types in 6 domains when compared to the corresponding domains in the ICL_Ontology group. These results indicate that our method effectively decreases the diversity of relationship types that the model extracts from triples.

4.4.2 OC. Merely reducing the number of relationship types is insufficient. When constructing a KG with minimal manual annotation, it is essential to have a higher number of relationships that adhere to the ontology's definitions. Consequently, we also computed the OC (Ontology Conformance) metric, which quantifies the ratio of relationship expressions within the extracted results that align with the predefined ontology list, relative to the aggregate number of extracted triples.

The experimental findings are depicted in Figure 4 and Figure 5. Across each domain, whether in the Context group or the ICL group, we observed that the CoT_Ontology group markedly enhanced the OC value of the triples extracted from the ten domain datasets when contrasted with the base group. The ICL_CoT_Ontology group outperformed the corresponding domains in the ICL_Ontology group in 8 out of the 10 domains. This suggests that, within the ICL prompt framework, CoT is effective in understanding the ontology and extracting relationships that are in line with the ontology's definitions. Additionally, we noted that the performance of the Context_CoT_Ontology group surpassed that of the corresponding domains in the Context_Ontology group in only 2 domains. This indicates that CoT exhibits superior performance in the ICL format prompt compared to when it is applied within the Context format prompt. Further analysis of these results is provided in section subsection 4.5.

4.5 Analysis

Our results reveal that the ICL group's performance in evaluating the specificity of relationship expressions is a little less robust compared to the Context group, despite its significant advantages in traditional metrics. To gain further insights, we examine a sample of the model's extraction results and conduct a thorough analysis across all indicators. While the ICL group's outcomes include new relationship types that contribute to a lower OC value, the expression of relationships within other triples was superior to that of the corresponding sentences in the Context group under equivalent conditions. This suggests that the model demonstrates a degree of autonomy but maintaining basic accuracy. Upon considering the total number of relationship types, we note that although the ICL_CoT_Ontology group had a slightly higher count compared to the Context_CoT_Ontology group, the discrepancy in OC values are minimal. This indicates that, under the ICL prompt format, the model does extract some relationship types that deviate from the ontology's definitions, but this does not detract from the accurate extraction of relationships that align with the ontology.

In the majority of domains, the traditional metrics for the Context_CoT_Ontology group don't not prove as effective as those for the Context_Ontology group, a stark contrast to the performance observed with ICL. This suggests that the CoT method under the Context format did not yield the anticipated benefits. This outcome validates our decision to construct CoT prompts within the ICL framework. The rationale behind this is as follows: with ICL, instructions and examples are integrated into the model's input as dialogue history, signifying that the model processes and learns information through successive interactions. In contrast, when using the Context format, instructions, examples, and other content are fed into the model all at once, resulting in an excessively long context that dilutes the potency of the reasoning chain. Within the confines of CoT prompts and ontology, the model demonstrates enhanced performance across a range of metrics. This discovery highlights the significance of mimicking human interaction in the learning process and offers a novel perspective for future research endeavors in natural language processing.

When analyzing the performance of a group by comprehensively considering all indicators, we found that in domains where accuracy is better, the OC value is also higher, indicating that increasing the specificity of relationship expression can effectively improve the accuracy of the information extraction process.

In summary, in the process of extracting triples using LLMs, our method can reduce the diversity of relationship expression while ensuring a certain level of accuracy.

5 CONCLUSION AND FUTURE WORK

In this paper, our goal is to address the core challenge of extracting KG triples from unstructured data, namely the diversity of relational expressions resulting from the variability of natural language expressions. Our proposed approach leverages the advantages of ontologies and CoT to guide LLMs in performing more consistent and accurate triple extraction. By incorporating a clear system of concepts and relational guidelines through the ontology framework, our method not only ensures the consistency of the KG but also enhances its semantic accuracy and clarity. The ontology serves as a filter and guidebook for the model, focusing its attention on relevant concepts and relationships while reducing the variety of relational expressions in the extracted triples. Furthermore, the innovative use of CoT has enabled the model to better understand complex tasks and maintain logical consistency during the triple extraction process. This form of guidance ensures that the model adheres to the rules and standards set forth, resulting in a more controlled and stable generation of KG triples.

In conclusion, our research contributes to the advancement of KG construction by introducing a novel approach that combines ontology and CoT. This work paves the way for more intelligent and precise knowledge services across various industries, relying less on manual annotation and moving towards more automated and efficient construction of KG.

Based on the current approach, there are several directions for future work that can further enhance the effectiveness and accuracy of KG triple extraction using LLMs. These include:

- **Standardized Dataset Construction** Developing a more standardized and comprehensive dataset is crucial for training and evaluating LLMs in KG triple extraction tasks. Future work should focus on creating datasets that cover a wide range of domains and include diverse natural language expressions and complete and correct triples. This will enable LLMs to learn from a rich set of examples and improve their ability to handle the diversity of relational expressions in unstructured data.

Table 4: Statistics on the number of relationship types in 6 groups of experiments. The bold numbers represent the best scores.

Type	Context			ICL		
	base	Ontology	CoT_Ontology	base	Ontology	CoT_Ontology
movie	664	**209**	212	1472	**246**	**246**
music	529	**106**	109	1161	136	**126**
sport	187	45	**43**	440	56	**55**
book	784	118	**116**	1308	118	**105**
military	221	38	**37**	539	**27**	28
computer	412	53	**52**	560	68	**62**
space	203	33	**31**	284	**34**	36
politics	294	40	**38**	490	86	**71**
nature	224	65	**61**	441	**86**	89
culture	151	72	**63**	202	**34**	37

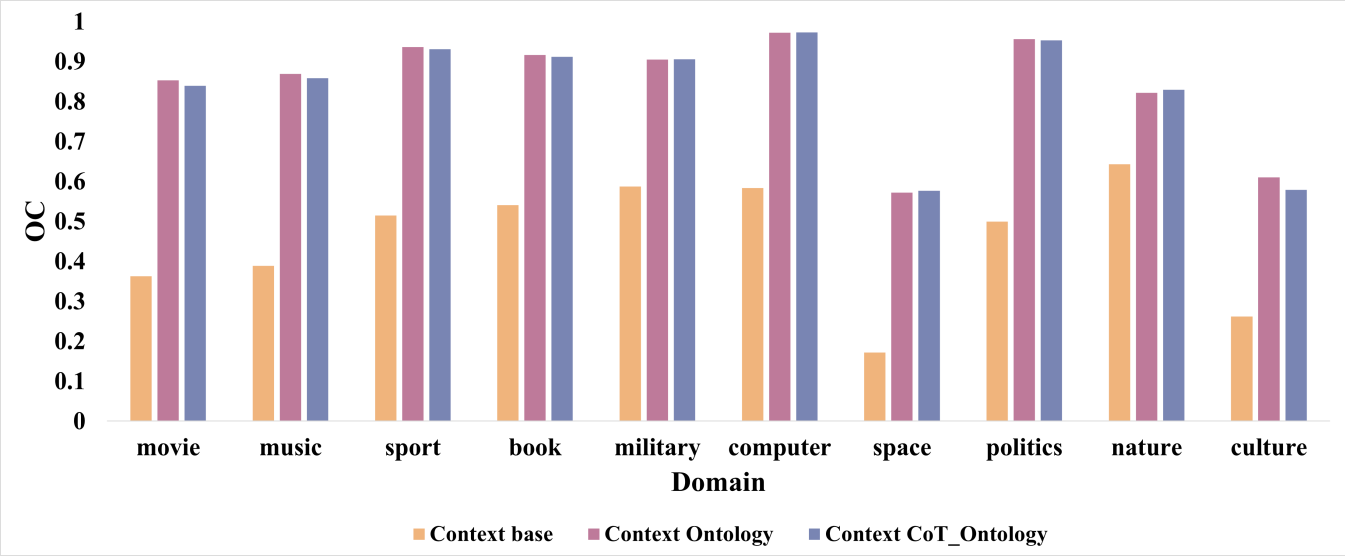

Figure 4: OC Value Statistics of Context Group

- **Advanced Ontology Representation** Improving the way ontologies are represented and integrated into LLMs can significantly enhance their understanding of domain concepts and relationships. Future work could explore more sophisticated ontology representation techniques, such as using graph neural networks or embedding methods, to capture the complex interrelationships between concepts. This will provide LLMs with a richer and more nuanced understanding of the ontology, enabling them to make more informed decisions during the triple extraction process.
- **Ontology-Aware Pre-training** Pre-training LLMs on data that is annotated with ontology information can help them develop a better understanding of domain-specific concepts and relationships. Future work could involve pre-training LLMs on large-scale datasets that are annotated with ontology labels, allowing the models to learn the ontology structure directly from the data. This will enable LLMs to better recognize and extract KG triples that adhere to the ontology guidelines.
- **Evaluation and Benchmarking** Establishing rigorous evaluation criteria and benchmarks is crucial for assessing the performance of KG extraction methods.

Future work should involve the development of comprehensive datasets and metrics to accurately measure the quality and utility of the extracted KG triples, rather than only evaluating the effectiveness of entity extraction and relation extraction separately. By addressing these areas of future work, we can continue to advance the field of KG construction, making it more accessible, efficient, and reliable for a wide range of applications.

ACKNOWLEDGMENTS

This paper is supported by Beijing Natural Science Foundation (L232102, 4222024), National Natural Science Foundation of China (62102036, 62441201).

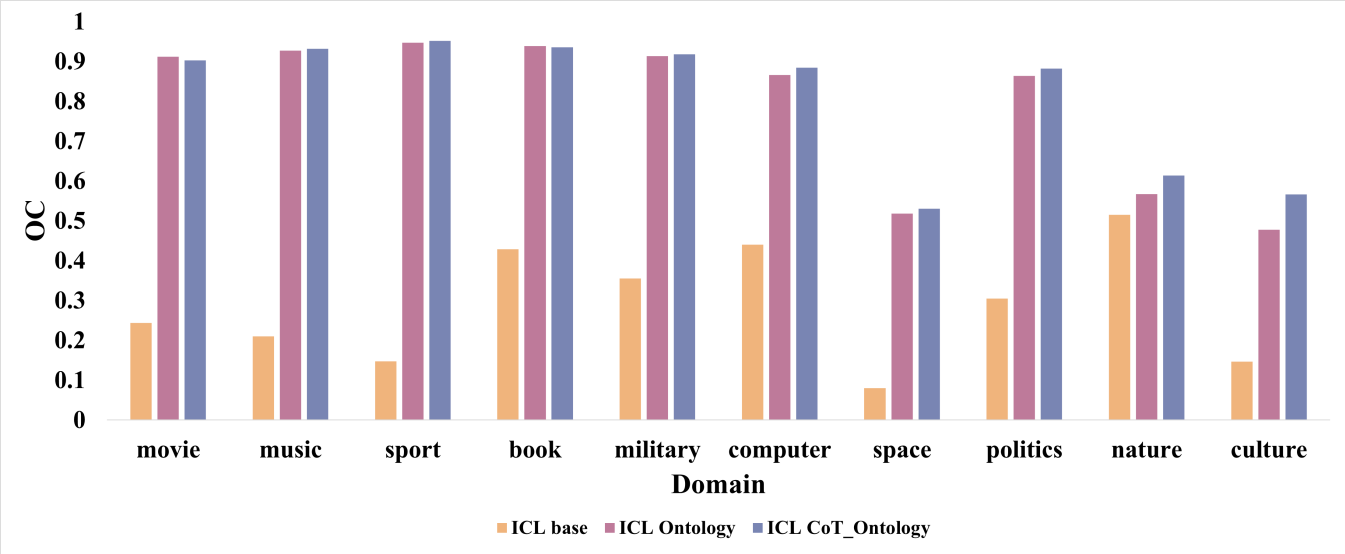

Figure 5: OC Value Statistics of ICL Group

REFERENCES

[1] Christopher Buss, Jasmin Mousavi, Mikhail Tokarev, Arash Termehchy, David Maier, and Stefan Lee. 2023. Generating Data Augmentation Queries Using Large Language Models. In *NeurIPS Table Representation Learning Workshop*.

[2] Lei Cao, Huijun Zhang, and Ling Feng. 2020. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia* 24 (2020), 87–102.

[3] Penghe Chen, Yu Lu, Vincent W Zheng, Xiyang Chen, and Boda Yang. 2018. Knowedu: A system to construct knowledge graph for education. *Ieee Access* 6 (2018), 31553–31563.

[4] Dawei Cheng, Fangzhou Yang, Xiaoyang Wang, Ying Zhang, and Liqing Zhang. 2020. Knowledge graph-based event embedding framework for financial quantitative investments. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2221–2230.

[5] Louie Giray. 2023. Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering* 51, 12 (2023), 2629–2633.

[6] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv* abs/1508.01991 (2015).

[7] Mayank Kejriwal and Pedro Szekely. 2017. Knowledge graphs for social good: An entity-centric search engine for the human trafficking domain. *IEEE Transactions on Big Data* 8, 3 (2017), 592–606.

[8] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[9] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387* (2021).

[10] Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. 139–151.

[11] Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. *arXiv preprint arXiv:2105.08393* (2021).

[12] Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. *arXiv preprint arXiv:2302.08675* (2023).

[13] Jose L Martinez-Rodriguez, Ivan López-Arévalo, and Ana B Rios-Alvarado. 2018. Openie-based approach for knowledge graph construction from text. *Expert Systems with Applications* 113 (2018), 339–355.

[14] Igor Melnyk, Pierre Dognin, and Payel Das. 2021. Grapher: Multi-stage knowledge graph construction using pretrained language models. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*.

[15] Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F Enguix, and Kusum Lata. 2023. Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In *International Semantic Web Conference*. 247–265.

[16] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[17] Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the workshop on vector space modeling for natural language processing*. 39–48.

[18] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[19] Seongsik Park and Harksoo Kim. 2021. Improving sentence-level relation extraction through curriculum learning. *arXiv preprint arXiv:2107.09332* (2021).

[20] Elvis Saravia et al. 2022. Prompt engineering guide. *GitHub. URL: https://github.com/dair-ai/Prompt-Engineering-Guide* (2022).

[21] Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *Comput. Surveys* 51, 5 (2018), 1–35.

[22] Barry Smith. 2012. Ontology. In *The furniture of the world*. 47–68.

[23] Wenfeng Song, Xinyu Zhang, Yuting Guo, Shuai Li, Aimin Hao, and Hong Qin. 2023. Automatic generation of 3d scene animation based on dynamic knowledge graphs and contextual encoding. *International Journal of Computer Vision* 131, 11 (2023), 2816–2844.

[24] Julio Vizcarra, Shuichiro Haruta, and Mori Kurokawa. 2024. Representing the Interaction between Users and Products via LLM-assisted Knowledge Graph Construction. In *IEEE International Conference on Semantic Computing*. 231–232.

[25] 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 (2022), 24824–24837.

[26] Yanbin Wei, Qiushi Huang, James T Kwok, and Yu Zhang. 2024. KICGPT: Large Language Model with Knowledge in Context for Knowledge Graph Completion. *arXiv preprint arXiv:2402.02389* (2024).

[27] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3 (2016), 1–40.

[28] Xin Xie, Ningyu Zhang, Zhoubo Li, Shumin Deng, Hui Chen, Feiyu Xiong, Mosha Chen, and Huajun Chen. 2022. From discrimination to generation: Knowledge graph completion with generative transformer. In *Companion Proceedings of the Web Conference*. 162–165.

[29] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Product knowledge graph embedding for e-commerce. In *Proceedings of international conference on web search and data mining*. 672–680.

[30] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* 18, 6 (2024), 1–32.

[31] Shuang Yu, Tao Huang, Mingyi Liu, and Zhongjie Wang. 2023. BEAR: Revolutionizing Service Domain Knowledge Graph Construction with LLM. In *International Conference on Service-Oriented Computing*. 339–346.

[32] Bowen Zhang and Harold Soh. 2024. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. *arXiv preprint arXiv:2404.03868* (2024).

[33] Yong Zhang, Ming Sheng, Rui Zhou, Ye Wang, Guangjie Han, Han Zhang, Chunxiao Xing, and Jing Dong. 2020. HKGB: an inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. *Information Processing & Management* 57, 6 (2020), 102324.

[34] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022).