A Self-Improving, RAG-Enhanced Framework for Automatic Knowledge Graph Construction from Climate Event News

Xiaorui Chen
The Chinese University of Hong Kong (Shenzhen)
224030224@link.cuhk.edu.cn

Haiyan Hao
The Chinese University of Hong Kong (Shenzhen)
haohaiyan@cuhk.edu.cn

ABSTRACT

Knowledge Graphs (KGs) have emerged as important tools in information management and data analysis, with applications spanning recommendation systems, medical diagnostics, and complex event understanding. However, constructing KGs automatically from climate event news remains challenging due to the presence of redundant entities and intricate relationships that arise from semantic complexity of climate events. Traditional methods (e.g., those relying on fine-tuned BERT models) often require extensive annotated datasets and struggle to merge semantically similar entities. Recent advancements in Large Language Model (LLM)-based algorithms show promise but are prone to hallucination, leading to inaccurate KGs. To address these limitations, we propose a novel LLM-based framework that combines Retrieval-Augmented Generation (RAG) with an iterative self-improving mechanism to improve the quality and accuracy of KGs extracted from climate event news. We compared the proposed method with three other commonly used prompting techniques. The experimental results showed that our method can effectively trace the driving factors of complex climate events and construct precise and reliable KGs. This framework offers a scalable, cost-effective solution for constructing domain-specific KGs, contributing to informed decision-making in managing complex crises.

Keywords

Knowledge Graphs, Domain-Specific Entity Recognition, Retrieval-Augmented Generation, Large Language Model, Compound Weather Event.

1. INTRODUCTION

A knowledge graph (KG) is a structured representation of knowledge that organizes information into a set of entities (nodes) and their relationships (edges) (Pujara, J. et al., 2013). KGs have become essential tools in information management and data analysis in recent years with extensive applications in domains such as recommendation systems, medical diagnostics, and crisis management (Nadeau et al., 2007; Deng, S. et al., 2020). In the context of crisis management, climate event news serves as a valuable and dynamic information source for constructing KGs, facilitating disaster preparedness and supporting stakeholders' decision-making, such news articles often contain rich details, including place names, organizations, event triggers, and impacts. Particularly, a knowledge graph based on triplet structures (i.e., subject–predicate–object) enables the distillation of unstructured textual contents into clearly defined event entities and their relationships, which cannot be directly inferred from raw text alone.

Constructing KGs based on news articles enables a deeper understanding of the dynamic evolution and mechanisms of weather-related disasters. For example, Zou et al. (2024) used a multilevel knowledge representation model to capture spatiotemporal characteristics and relationships within urban rainstorm events. Normally, this task is achieved with human annotators, which is labor-intensive and time-consuming. To address this, automated approaches have emerged, utilizing advanced natural language processing (NLP) techniques. For instance, fine-tuned BERT models have been applied for domain-specific entity recognition in the field of climate science or finance (Jawahar et al., 2019). More recently, large language models (LLMs), such as the GPT series, have demonstrated exceptional language understanding and generation capabilities through prompt engineering (Floridi et al., 2020). Thus, some scholars have also explored the potential of LLMs for automated KG construction (Chen et al., 2023).

These automated approaches significantly reduce the cost of KG construction. They eliminate the need for manual feature design, enable automatic model generation through training on large datasets, and demonstrate strong performance in common named entity recognition (NER) tasks. However, they still present certain challenges. First, in the process of entity recognition for climate-related news, the widespread phenomenon of semantically similar expressions is evident—identical or similar climate events or elements are often described using different terminologies. Without semantic identification and normalization processing, this would lead to a large number of redundant entities in the resulting KG. For instance, words such as "rainstorm," "extreme rainfall," and "heavy rain" may refer to the same type of meteorological disaster. If these semantically similar expressions are treated as independent nodes, it would result in bulky graph structures, ultimately compromising the query efficiency and analytical accuracy of the KG. Second, BERT-based methods demand extensive high-quality annotated datasets for fine-tuning. These sequence labeling models typically require token-level BIO annotation for climate event extraction. For instance, in the sentence "*Three consecutive days of heavy rainfall caused flooding in southern mountainous areas*", each character in 'heavy rainfall' must be labeled as B-Hazard I-Hazard, while 'flooding' requires another B-Hazard I-Hazard tag, with all other tokens marked as O. This character-level annotation paradigm not only necessitates substantial time and financial investment for data collection and labeling, but also requires annotators to possess both linguistic skills and domain expertise. Such requirements severely limit the generalization capability of traditional BERT approaches to new domains or emerging events, making them particularly unsuitable for rapid-response climate crisis scenarios where timely analysis is critical.

Large language model (LLM)-based methods demonstrate significant advantages in addressing the two aforementioned challenges. With strong contextual understanding and zero- or few-shot capabilities, LLMs can perform entity recognition and normalization without requiring detailed manual annotation. By leveraging well-designed prompts, LLMs can identify and unify semantically equivalent expressions, such as "rainstorm" and "extreme rainfall", and directly generate structured triplets, eliminating the need for token-level BIO tagging. This makes LLM-based approaches particularly suitable for scenarios where annotated training data is scarce. However, LLMs are also associated with several notable limitations. They may suffer from hallucinations, occasionally producing entities or relations that are inconsistent with the original text. Additionally, when dealing with long or complex narratives, LLMs may also fail to capture key causal chains.

Zscheischler et al. (2020) categorized climate-related events/phenomena into four general categories including hazards, modulators, drivers, and impacts, and summarized common patterns. These patterns provide insights into the complicated spatiotemporal relationships among climate events, inspiring our work on automatically mining relations among various climate events and related factors by constructing KGs from historical textual records, such as news articles. We consider such KGs can standardize the analysis of compound climate events, facilitating a more systematic understanding of their underlying dynamics, and uncovering historical patterns of compound climate phenomena to improve future event predictions.

Building on Zscheischler et al. (2020)'s categorization, this study proposes a framework for domain-specific entity and relationship extraction from climate event news. Our framework leverages Retrieval-Augmented Generation (RAG) to provide LLMs with domain-specific knowledge and incorporates a self-improving mechanism to progressively refine extraction outcomes. Specifically, the framework enumerates potential triplets (i.e., entity-relation-entity) and constructs high-quality KGs based on the four core climate event elements defined by Zscheischler et al. (2020), i.e., hazards, modulators, drivers, and impacts. Triplets are vectorized, and relevant ones are retrieved from presented news articles using similarity matching, providing contextual support for the LLM. A self-improving mechanism then iteratively evaluates and refines the KG, with a counter mechanism determining whether further iterations are needed, ensuring more accurate and comprehensive results.

We tested the proposed pipeline on 100 real climate event news collected from Shenzhen, China. Experimental results demonstrate that our proposed framework effectively reduces semantic redundancy, identifies core driving entities of complex climate events, and significantly improves the accuracy of extracted entities and relationships from news articles. This work offers a low-cost, scalable method for KG construction in domains requiring high timeliness and expertise. It also presents a feasible strategy for external knowledge enhancement and iterative extraction using LLMs.

2. LITERATURE REVIEW

The early version of automated KG construction work focused on NER, which aims to identify entities mentioned in a sentence and classify them into predefined entity types (Nadeau, D. et al., 2007). In general domains, these entity types often include names of people, places, organizations, times, and more. Deep learning-based methods have been widely used for NER tasks, achieving advanced performances (Li et al., 2020). Compared to traditional methods that rely on manually engineered features, deep learning-based approaches automatically learn features by training on large datasets, eliminating the need for manual design. However, such methods require a substantial amount of high-quality annotated data to achieve optimal performance.

LLMs mitigate this limitation through their ability to leverage vast amounts of minimally or unannotated data, combined with pre-existing knowledge from pretraining. The emergence of LLMs has revolutionized various NLP tasks (He et al., 2023), enabling their application in downstream tasks, including KG construction. Prompt engineering is a critical technique that adapts LLMs for specific NLP tasks. Among various prompt engineering techniques, three major paradigms stand out.

- Few-Shot Prompting (Perez et al., 2021), which improves the model's capabilities using a few examples but its generalization capacity for complex semantics is limited.
- Chain-of-Thought (CoT) Prompting (Wei et al., 2022), which improves the model's reasoning by generating explicit reasoning chains for better multi-step reasoning but may lead to unnecessary inferences.
- ReAct Prompting (Yao et al., 2022), which combines reasoning with interaction, allowing the model to explicitly reason and execute actions during responses, thereby improving environmental perception and decision-making in complex scenarios.

While these prompt techniques improve reasoning capabilities for NER tasks, they still lack effective integration of domain-specific knowledge, limiting their applicability to specialized fields. The development of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) addresses this gap by integrating external knowledge to support reasoning. RAG retrieves relevant documents or vectors from external knowledge bases based on queries and incorporates them into the model's input to support more accurate content generation.

Despite its potential, there are limited frameworks that leverage RAG to enhance KG construction based on LLMs. Chen et al., (2023) proposed AutoKG, a lightweight and efficient automated KG construction framework. It uses pre-trained LLMs to extract keywords from a knowledge base, clusters text blocks via unsupervised algorithms, and calculates text similarity using graph Laplace learning to build graph structures. Lairgi et al., (2024) proposed iText2KG, an incremental, domain-agnostic KG construction method using "zero-shot" learning, avoiding reliance on predefined ontologies or extensive post-processing.

However, these frameworks are not optimized for predefined entity recognition in domain-specific tasks. Direct application to such tasks often yields suboptimal results due to the lack of domain-specific knowledge integration. Combining RAG with a self-improving mechanism for entity recognition in LLMs could address this limitation by maximizing entity and relationship extraction while improving accuracy. For instance, Sheng (2024) introduced a simplified self-improving approach by incorporating an evaluation mechanism to enable LLMs to decide whether further iterations are needed, Similarly, Madaan et al. (2024) proposed Self-Refine, an iterative self-feedback mechanism that allows LLMs to refine their own outputs without additional training, significantly enhancing performance across diverse tasks.

In summary, traditional deep learning methods, while effective, require extensively annotated datasets and are resource-intensive (Hao & Wang, 2020). In contrast, LLMs, leveraging in-context learning (ICL), perform well across diverse NLP tasks without additional training or fine-tuning but lack domain-specific knowledge integration. The advent of RAG offers a promising solution by enabling LLMs to incorporate external knowledge. When combined with the iterative self-improving mechanisms for multi-round entity and relationship extraction, RAG-based approaches can significantly improve the accuracy and scalability of domain-specific KG construction.

3. METHOD DESCRIPTION

We propose a framework that integrates Retrieval-Augmented Generation (RAG) with a self-improving mechanism for constructing KGs from domain-specific news articles. Unlike existing methods that rely solely on LLMs for KG construction, the use of RAG for retrieving entities and relationships allows the inclusion of domain-specific knowledge that is absent in pre-trained LLMs and, hence, enhances the quality and relevance of the resulting KGs. To further improve accuracy and robustness, our framework incorporates several advanced techniques, including prompt engineering and a self-improving mechanism. These components enable iterative refinement of the extracted entities and relationships, reducing redundancy and increasing precision. **Figure 1** shows the scheme design of the framework, in which a stepwise approach is used to process news articles and construct high-quality KGs.

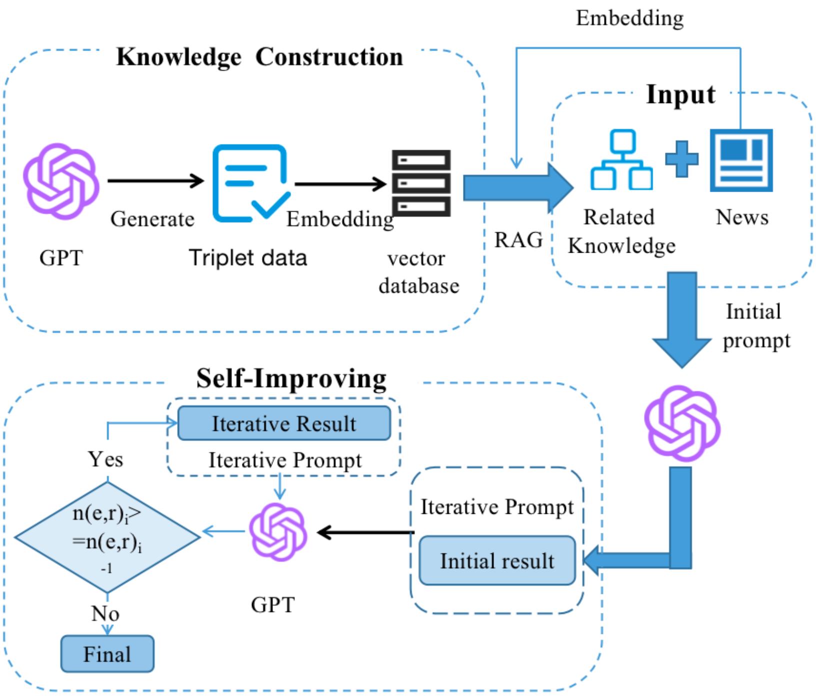

Figure 1. The Structure of Proposed Framework

3.1 Generating A Generic KG Storing Domain Knowledge

The first step involves constructing a high-quality, domain-specific KG containing generic information about climate events. This process is guided by the categorization framework for compound weather events proposed by Zscheischler et al. (2020):

- *Driver*: The direct driving factors causing specific climate events.
- *Impact*: The contextual factors regulating the intensity or distribution of climate events, influencing the severity and/or frequency of events through specific phenomena, events, or processes.
- *Hazard*: The climate hazards triggered by driving factors, leading to negative social or environmental impacts.
- *Modulator*: The potential impacts of climate hazards on nature and society, including the negative consequences for the environment and communities.

To enhance the LLM's understanding of these entity types, we constructed an initial KG database. This was achieved by designing prompts specifically for KG data generation. The initial prompts included examples of entities and relationships drawn from the literature, helping the LLM generate initial KG data that adhered to domain-specific definitions.

Building on these initial results, we refined the prompts by introducing additional requirements, such as "*expanding the range of entity types to include richer and more diverse examples*". This allowed the LLM to generate examples of entities, such as strong convective weather and typhoons, which align with definitions from literature but were not explicitly listed in the initial output.

We manually evaluated the generated KG to assess and improve their quality, ensuring accuracy and diversity. The evaluation criteria included: 1) correctness of entity types; 2) rationality of relationships; and 3) coverage of examples. After validation, the KG was vectorized and stored in a vectorized knowledge base, enabling efficient retrieval and reuse for subsequent steps of the framework.

3.2 Semantic Retrieval Based on Embedding Similarities

To initiate the knowledge extraction process, we first generate embeddings for each news article. For this purpose, we used OpenAI's "text-embedding-ada-002" model, which has demonstrated state-of-the-art performance in text similarity tasks (Li et al., 2023) and can be easily accessed with OpenAI's API.

Next, we calculate the cosine similarity between the embeddings of the news article and those of pre-vectorized triplets stored in the knowledge base. This step enables the identification of triplets that are contextually relevant to the news content. To filter the most relevant triplets, we applied a similarity threshold of 0.8, which was empirically determined to strike a good balance between precision (minimizing false positives) and recall (maximizing relevant triplet retrieval).

The triplets passing this threshold are considered highly related to the presented news content and serve as the basis for querying detailed information about *entities* and *relations* within the article. **Figure 2** illustrates examples of high-similarity entities retrieved with the filtered triplets.

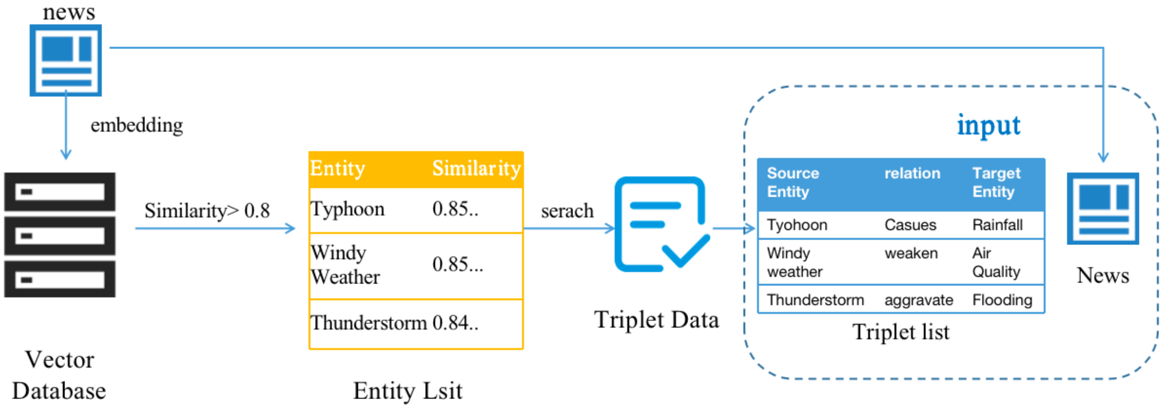

Figure 2. The Process of Retrieval-Augmented Generation.

3.3 Generate Initial Result

To initiate entity recognition and relationship extraction from a news article, we provide an initial prompt that accomplishes two primary objectives: 1) defining and constraining the target entities and relationships, and 2) specifying a precise output format. As shown in **Figure 3** (the "Initial Prompt"), the process begins by providing clear definitions for each entity type and link type. This ensures that the model has an explicit reference for domain-specific concepts. We then define a concise JSON output template to enforce a structured and verifiable format for recognized entities and relationships.

Additionally, the prompt explicitly outlines the following tasks:

1) *Extract domain-specific content*: Identify only the entities and relationships explicitly mentioned in the presented news article.
2) *Validate extracted information*: Ensure every entity or relationship has a supporting sentence from the article; otherwise, mark it for deletion.
3) *Eliminate redundancy*: Merge semantically similar entities to minimize duplication and improve clarity.
4) *Filter external noise*: Exclude entities or relationships that only appear in external knowledge sources but are absent from the article.

By limiting the extraction scope to the provided news content and requiring a "*supporting sentence*" for validation, the process mitigates the risk of hallucinations and enhances the reliability of the results. These measures ensure that the model relies strictly on the input text for generating outputs. The initial prompt captures the majority of relevant entities and relationships, establishing a robust baseline for subsequent refinement steps. This approach standardizes domain-specific entity recognition and aligns the model's output with downstream tasks, such as KG construction.

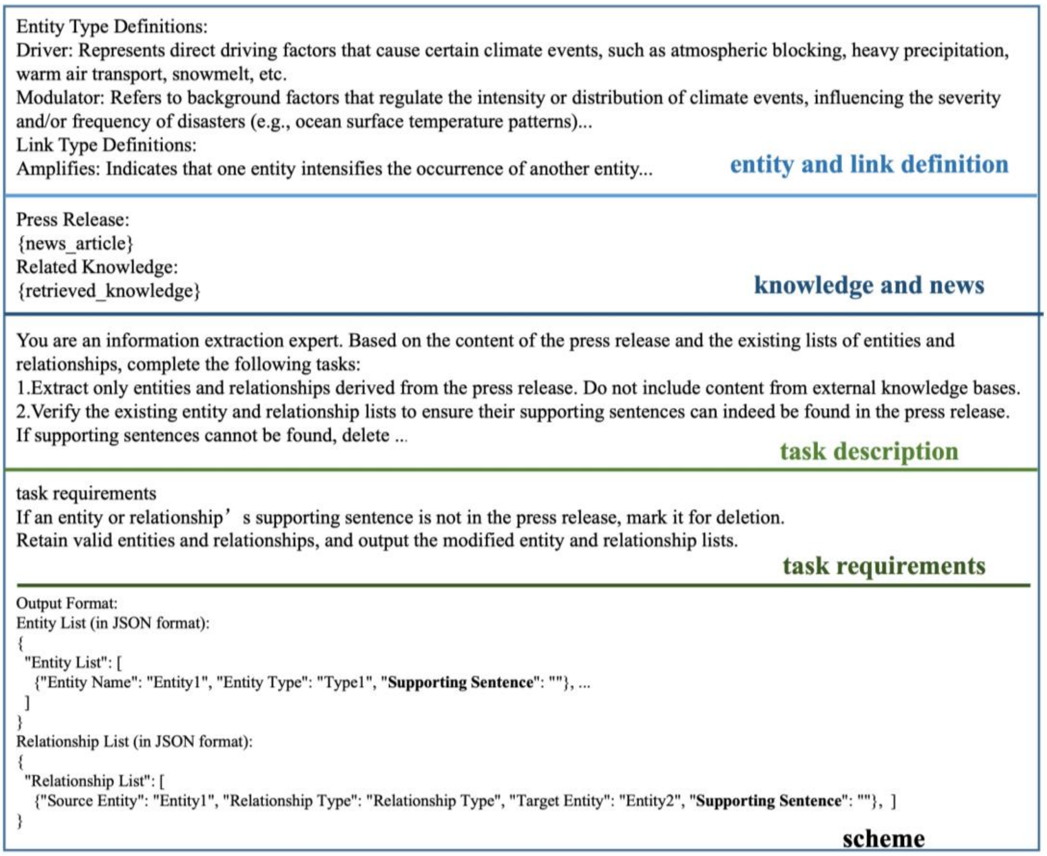

Figure 3. The Structure of Initial Prompt (translated in English)

3.4 Refining Retrieved Results

Although the initial prompt effectively identifies numerous entities and relationships, it is not immune to common challenges such as hallucinations (fabricated entities or relationships not present in the news) or omissions (overlooked subtle, domain-specific entities). To address these issues, we employ an iterative refinement process, as illustrated in **Figure 4**.

This iterative prompt enhances the initial output by systematically revisiting the press release and the previously generated entity and relationship lists. The refinement process focuses on the following key tasks:

1) *Identify missed or newly detected entities/relationships*: Scans for the identified entities and relationships that may have been overlooked in the initial pass.
2) *Eliminate hallucinations:* Any entities or relationships lacking supporting evidence in the press release are flagged and removed.
3) *Correct relationships:* Verifies the accuracy of relationships, ensuring the proper alignment of source and target entities, and resolves any misclassification.

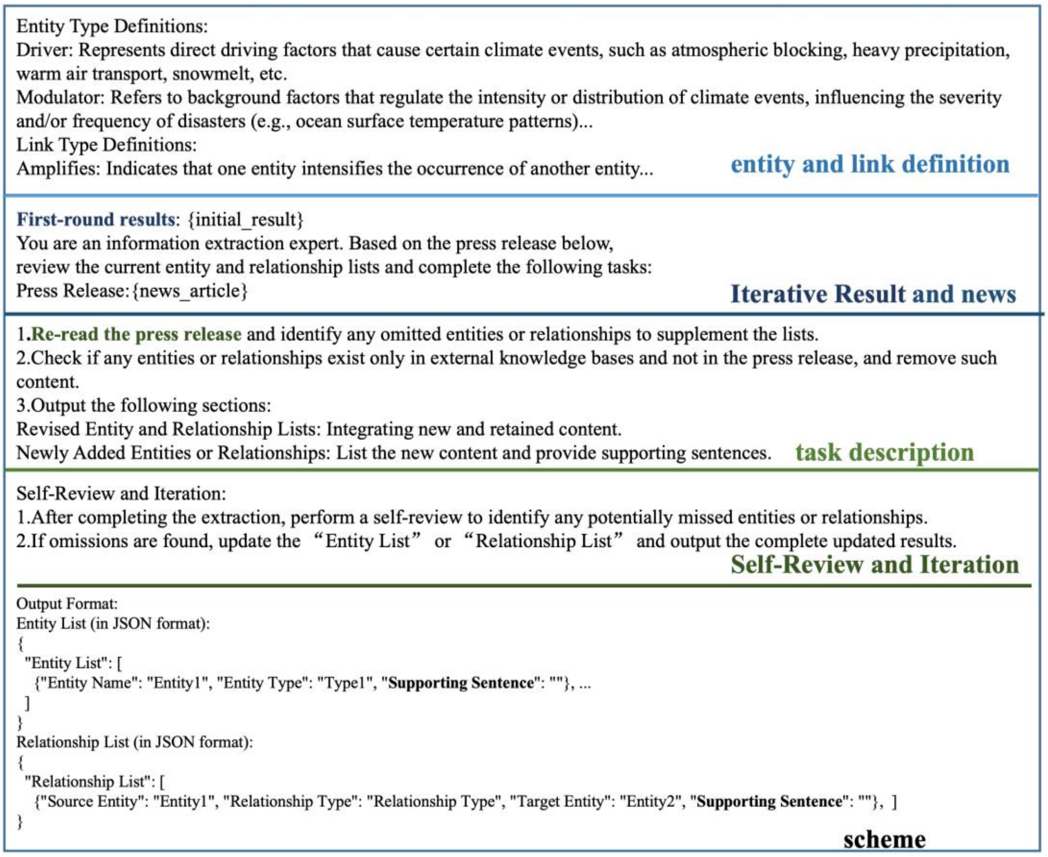

Figure 4. The Structure of Iterative Prompt (translated in English)

After each iteration, the model produces a revised list of entities and relationships. A simple counter mechanism is employed to determine whether further refinement is needed. This mechanism evaluates the output based on two criteria: 1) the total number of valid entities and relationships stops increasing or 2) a predefined maximum iteration count is reached. This iterative loop design incrementally removes fabricated entities and relationships while identifying those initially missed.

Algorithm 1 Iterative Extraction

Input: news_article (Text), retrieved_knowledge (Text), max_iterations (Integer)

Output: FinalEntityAndRelationList (JSON)

```
function IterativeExtraction(news_article, retrieved_knowledge, max_iterations)
  InitialPrompt ← GenerateInitialPrompt(news_article, retrieved_knowledge)
  InitialResult ← QueryModel(InitialPrompt)

  PrevTotal ← 0
  Total ← 0
  LoopCounter ← 1

while LoopCounter ≤ max_iterations do
     ParsedData ← ParseEntitiesAndRelationships(InitialResult)
     NumEntities ← CountEntities(ParsedData)
     NumRelationships ← CountRelationships(ParsedData)
     Total ← NumEntities + NumRelationships

   if Total ≤ PrevTotal then
          Print("No further improvement, stopping iteration.")
          break

     PrevTotal ← Total
     FollowUpPrompt ← GenerateFollowUpPrompt(ParsedData, news_article)
     FollowUpResult ← QueryModel(FollowUpPrompt)
     InitialResult ← FollowUpResult

     Print("Iteration:", LoopCounter, "Entities:", NumEntities, "Relationships:", NumRelationships, "Total:", Total)

     LoopCounter ← LoopCounter + 1

  FinalEntityAndRelationList ← ParseEntitiesAndRelationships(InitialResult)
  return FinalEntityAndRelationList
end function
```

4. EXPERIMENT

We utilized a dataset comprising 100 weather-related news articles from Shenzhen, written in Chinese. These articles were sourced from various weather-focused public accounts and news media. The dataset includes diverse content such as daily weather forecasts, reports on weather phenomena, and predictions of weather hazards. This richness in content introduces significant challenges for the LLM due to the presence of numerous semantically similar entities and specialized weather-related terminology. **Figure 5** shows an example of KG generated with the proposed framework.

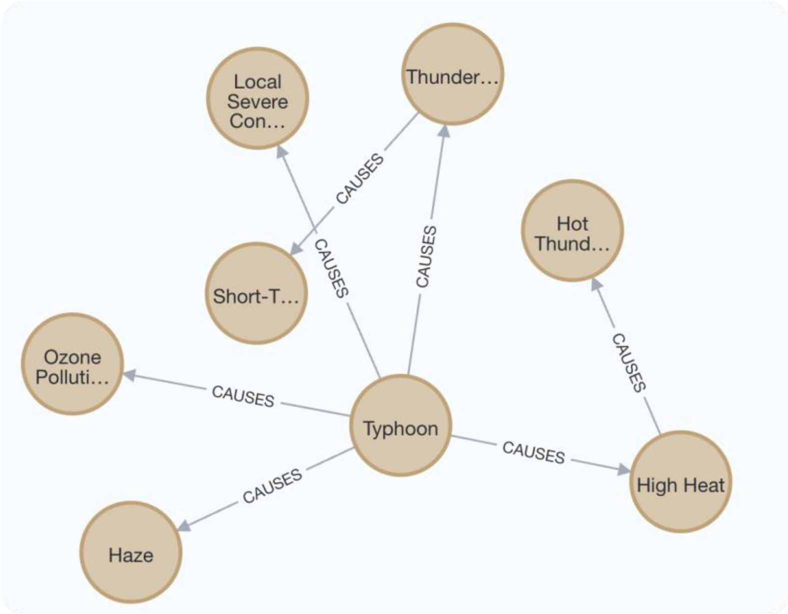

Figure 5. An example of a Knowledge Graph Generated with Our Method Plotted with Neo4j (translated in English)

To evaluate our framework, we compared its performance with three widely-used prompting techniques: *Few-Shot Prompt*, *CoT prompt*, and *ReAct prompt*. Each technique was adapted for entity and relationship extraction tasks as follows:

- *Few-Shot Prompt*: This method involves providing a few examples to help the LLM understand the task. For this study, we designed multiple examples of entities and relationships to guide the LLM in performing the extraction tasks effectively.
- *Chain-of-Thought (CoT) Prompt*: This method encourages the model to think through problems step-by-step. In this study, we designed a reasoning chain to guide the LLM to think progressively through identifying entities and relationships, improving its knowledge extraction performance.
- *Synergize Reasoning+Acting (ReAct) Prompt*: This technique combines reasoning and action in a multi-step process, including understanding the problem, reasoning, executing tasks, verifying results, and iterating. We adapted this approach by designing prompts aligned with this structured methodology.

The definitions of entities and relationships remained consistent across all prompt types. The primary goal of this evaluation was to analyze the differences in outcomes generated by each prompting design, highlighting their respective strengths and weaknesses.

We experimented with the different types of prompt techniques with ChatGPT-4o (Hurst et al., 2024), the latest model from OpenAI. This model matches GPT-4 Turbo's performance for English text and code while demonstrating significant improvement in handling non-English languages such as Chinese. **Figure 6** presents the comparative outputs of different methods, highlighting both the correct and false extractions.

There are currently three typhoons active in the South China Sea and the northwest Pacific, namely "Chapaca" (tropical depression level) located in the Beibu Gulf, "Yanhua" (typhoon level) located in the sea east of Zhejiang, and "Nibert" located in the northwest Pacific (Tropical storm level, about 3,600 kilometers away from Shenzhen). "Fireworks" is expected to make landfall in Zhejiang from the afternoon to night of the 25th. The downdraft outside it has brought high temperatures and local strong convection weather to Shenzhen for two days; "Chapaca" may enter the South China Sea again, bringing thunderstorms around the 26th; "Nibert" tends to Japan and has no impact…… **News**

Entity Name	Entity Type	Supporting Sentence
Typhoon	Driver	There are currently three typhoons ……
Extreme Heat	Hazard	Its surrounding subsidence airflow has ……
Localized Severe Convective Weather	Hazard	Its surrounding subsidence airflow has ……
Thunderstorm Weather	Hazard	"Chapaca" may enter the South China Sea ……
Ozone Pollution	Impact	The subsidence airflow on the outskirts ……
Haze	Impact	The subsidence airflow on the outskirts ……
Heat Thunderstorm	Hazard	Localized heat thunderstorm are likely……
Short Gusts	Hazard	Rainfall is accompanied by short gusts of……

Source Entity	Relation	Target Entity	Supporting Sentence
Typhoon	Causes	Extreme Heat	Its surrounding……
Typhoon	Causes	Localized Severe Convective Weather	Its surrounding……
Typhoon	Causes	Thunderstorm Weather	"Chapaca" may ……
Typhoon	Causes	Ozone Pollution	The subsidence……
Typhoon	Causes	Haze	The subsidence……
Extreme Heat	Causes	Heat Thunderstorm	Localized heat ……
Thunderstorm Weather	Causes	Short Gusts	Rainfall is ……

Our Method

Few-Shot Prompt

Entity Name	Entity Type
Chapaka	Driver
Fireworks	Driver
Epsilon	Driver
Extreme Heat	Hazard
Localized Severe Convective Weather	Hazard
Thunderstorm Weather	Hazard
Ozone Pollution	Hazard
Haze	Hazard
Subsidence Flow	Modulator
Extreme Temperature Forecast	Impact

Source Entity	Relation	Target Entity
Firework	Amplifies	Extreme Heat
Firework	Amplifies	Localized Severe Convective Weather
Chapaks	Amplifies	Thunderstorm Weather
Subsidence Flow	Amplifies	Ozone Pollution
Subsidence Flow	Amplifies	Haze

CoT Prompt

Entity Name	Entity Type
Typhoon	Driver
Extreme Heat	Hazard
Localized Severe Convective Weather	Hazard
Thunderstorm Weather	Hazard
Ozone Pollution	Impact
Haze	Impact
Subsidence Flow	Modulator
Extreme Temperature Forecast	Modulator

Source Entity	Relation	Target Entity
Typhoon	Amplifies	Extreme Heat
Typhoon	Amplifies	Localized Severe Convective Weather
Typhoon	Amplifies	Thunderstorm Weather
Extreme Heat	Amplifies	Localized Severe Convective Weather
Subsidence Flow	Weakens	Pollutant Diffusion
Subsidence Flow	Amplifies	Ozone Pollution
Subsidence Flow	Amplifies	Haze
Extreme Temperature Forecast	Exacerbate	Extreme Heat

ReAct Prompt

Entity Name	Entity Type
Chapaka	Driver
Fireworks	Driver
Epsilon	Driver
Extreme Heat	Hazard
Localized Severe Convective Weather	Hazard
Thunderstorm Weather	Hazard
Ozone Pollution	Hazard
Haze	Hazard
Subsidence Flow	Modulator
Extreme Temperature Forecast	Modulator
Heat Thunderstorm	Hazard
Short Gusts	Hazard

Source Entity	Relation	Target Entity
Fireworks	Amplifies	Extreme Heat
Fireworks	Amplifies	Localized Severe Convective Weather
Subsidence Flow	Amplifies	Thunderstorm Weather
Subsidence Flow	Amplifies	Extreme Heat
Subsidence Flow	Amplifies	Ozone Pollution
Chapaka	Amplifies	Haze
Localized Severe Convective Weather	Amplifies	Heat Thunderstorm
Heat Thunderstorm	Amplifies	Short Gusts

Figure 6. Output examples of ours, Few Shot, CoT, and ReAct, with green indicating correct recognition and red indicating incorrect recognition (translated in English).

Comparison between Our Method and Few-Shot Prompt

Entity Accuracy: Our framework outperformed the few-shot prompt in entity recognition by accurately generalizing specific typhoon names such as "Cempaka" and "Nepartak" into a unified entity, "Typhoon." This capability reduces redundancy and noise, showcasing the framework's strength in merging semantically similar entities. In contrast, the few-shot prompt treated each typhoon name as a separate entity, which introduced unnecessary complexity and reduced the clarity of the results.

Relationship Recognition: Our framework successfully identified the impacts associated with typhoons and correctly linked them back to the overarching "Typhoon" entity. Conversely, the few-shot prompt struggled to establish these relationships due to its fragmented entity recognition, leading to misaligned relationships and increased data noise. Moreover, our framework demonstrated superior semantic understanding by recognizing "Typhoon" as the root cause of certain impacts, whereas the few-shot prompt erroneously identified "subsiding airflow" as the initiating entity.

Comparison between Our Method and CoT Prompt

Entity Accuracy: Both our framework and the CoT prompt performed well in entity recognition, but our framework demonstrated a slight edge. For instance:

- The CoT prompt mistakenly recognized "High-Temperature Orange Alert" as an entity, despite it not representing a complex weather phenomenon.
- The CoT prompt failed to classify "subsiding airflow" as part of the "Typhoon" entity, a mistake that our

framework avoided.

These errors demonstrate CoT's limitation in semantically grouping related phenomena, a task our framework handles effectively through its iterative refinement and semantic merging.

Relationship Recognition: While the CoT prompt avoided errors related to specific typhoon names, it incorrectly identified "subsiding airflow" as the initiating entity in relationships, failing to trace the causality back to "Typhoon". In contrast, our framework accurately linked impacts to the appropriate source entity, providing a more coherent and meaningful relationship structure.

Comparison between Our Method and ReAct Prompt

Entity Accuracy: Similar to CoT, the ReAct prompt mistakenly identified specific typhoon names and "High-Temperature Orange Alert" as distinct entities. These misclassifications highlighted ReAct's limited ability to merge semantically similar entities in recognizing meaningful groupings.

Relationship Recognition: While ReAct effectively identified impacts caused by specific typhoons, its erroneous entity recognition rendered these relationships largely meaningless. Like CoT and few-shot, ReAct incorrectly identified "subsiding airflow" as the initiating entity in relationships, failing to trace the root cause to "Typhoon". Our framework avoided such errors and produced relationships that were both accurate and semantically coherent.

We summarized the advantages of our proposed method, as compared to the three baseline prompting techniques, in **Table 1**.

Table 1. Comparison between the proposed method and other prompt-based methods for domain-specific KG construction.

Aspect	Baseline Prompts	Proposed Framework
NER Ability		
Entity Coverage	Often identified redundant entities, limiting comprehensive event chain description.	Broader and more accurate, generalizes specific mentions to avoid redundancy.
Accurate Classification	Struggled with complex categorization and lacked verification mechanisms.	Handles complex categorization with high accuracy, using RAG for references.
Relationship Recognition	Prone to hallucinations, incorrectly identifying relationships between distant entities.	Identifies both positive and negative relationships accurately, avoiding hallucinations.
Output Reliability	Lacked context for verifying output accuracy.	High confidence and verifiability, with supporting context sentences.
Framework Scalability		
Prompt Customizability	Limited adaptability to different fields without extensive model redesign.	Modular design allows easy adaptation to different fields without complex model design.
Knowledge Base Replaceability	Not designed for easy cross-domain migration.	Supports cross-domain migration by replacing knowledge bases without major adjustments.
Self-Improving Mechanism	Lacked mechanisms for iterative improvement and domain-specific optimization.	Domain-independent, adaptable with evaluation indicators for performance optimization.

5. DISCUSSION AND CONCLUSION

5.1 Methodological Contributions

Despite the progress in KG construction, previous studies face several limitations. First, fine-tuned BERT models require a substantial amount of annotated data for domain-specific fine-tuning, but the high cost and inefficiency of manual annotation render this approach impractical for many domains (Webersinke et al., 2022). Second, high semantic similarity between entities often results in redundancy, inflating knowledge graphs with significant noise and reducing their usability. Third, traditional one-shot extraction methods struggle to accommodate the complexity of domain-specific terminologies. These limitations hinder the construction of effective knowledge graphs for climate event news, restricting progress in automated analysis.

To address these challenges, this study presents an innovative framework that integrates RAG with a self-improving mechanism to enhance domain-specific entity and relationship extraction in climate-related event news. Unlike traditional methods, the proposed framework eliminates the need for additional fine-tuning. Instead, it leverages LLMs along with semantic context from existing knowledge graphs, to enhance recognition efficiency and accuracy. For this study, we constructed a high-quality knowledge graph incorporating four key types of climate events or factors defined in the literature: *hazards*, *modulators*, *drivers*, and *impacts*. By introducing an iterative counter, the framework refines the entity and relationship extraction results through iterations, resulting in a highly efficient and precise KG in climate event contexts.

5.2 Practical Implications in Crisis Management

Domain-specific knowledge graphs (KGs) offer powerful support for crisis management through both visual and analytical capabilities. Visually, KGs help decision-makers grasp event dynamics—such as key drivers and causal relationships—enabling more informed responses. Analytically, KGs facilitate downstream applications like real-time information querying, domain-specific AI model development, and decision support systems (DSS) (e.g., Chasseray et al., 2024; Janzen et al., 2024). They also aid in simulating complex, compound crisis scenarios, equipping local governments to navigate multifaceted challenges (Hao et al., 2024). Our future work will expand entity types and integrate spatiotemporal attributes (e.g., location, time) to create generalized climate knowledge bases. Such enhancements will improve the detection of compound climate phenomena, where interconnected hazards or drivers amplify risks across regions.

5.3 Limitations and Future Work

Despite its advantages, the framework is subject to several limitations that merit discussion. First, the current experimental evaluation was conducted on a relatively small dataset consisting of only 100 climate-related news articles. In future work, we plan to extend our evaluation using a more diverse and comprehensive dataset—including sources from different cities and media outlets—to further validate the robustness and generalizability of the framework. Second, while we made an initial attempt to quantify extraction accuracy by recruiting and training volunteers to manually annotate entities and relations for comparison, we encountered issues related to annotation consistency. In some cases, annotators failed to identify certain entities that were correctly extracted by the LLM, indicating the difficulty of establishing a reliable gold-standard dataset. Developing a standardized and accurate annotation pipeline will be an important direction for future work. Third, the current framework lacks memory functionality, preventing the retention of intermediate reasoning results across iterations and limiting its contextual understanding. Integrating memory components could improve coherence and reasoning depth. Additionally, although prompt engineering and iterative refinement help mitigate hallucinations, the framework still occasionally generates outputs influenced by retrieved knowledge rather than grounded news content. Addressing this limitation will also be part of our ongoing research.

5.4 Conclusion

In conclusion, this study introduces a scalable, accurate, and traceable entity recognition and relationship extraction framework using LLMs. The framework is well-suited for tasks involving predefined entity recognition and relationship extraction in specialized domains. It eliminates the need for large-scale annotated data for fine-tuning, making it highly cost-effective and adaptable to diverse fields. As foundational LLMs continue advance, the framework's performance is expected to improve further. Further optimizations, such as incorporating a memory module or enabling multi-document iteration, hold the potential to unlock even greater capabilities for the presented framework.

REFERENCE

Chasseray, Y., Conges, A., Fertier, A., & Barthe-Delanoë, A. M. (2024, May). Knowledge on demand: the future of decision support systems?. In *ISCRAM 2024-21th-Information Systems for Crisis Response and Management.* (Vol. 21).

Chen, B., & Bertozzi, A. L. (2023). AutoKG: Efficient automated knowledge graph generation for language models. In *2023 IEEE International Conference on Big Data (BigData)*, (pp. 3117–3126). IEEE

Deng, S., Rangwala, H., & Ning, Y. (2020, August). Dynamic knowledge graph based multi-event forecasting. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1585-1595).

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681–694.

Hao, H., & Wang, Y. (2020, May). Hurricane Damage Assessment with Multi-, Crowd-Sourced Image Data: A Case Study of Hurricane Irma in the City of Miami. In *ISCRAM* (pp. 825-837).

Hao, H., Wang, Y., & Chen, J. (2024). Empowering Scenario Planning with Artificial Intelligence: A Perspective on Building Smart and Resilient Cities. *Engineering*.

He, J., Wang, L., Hu, Y., Liu, N., Liu, H., Xu, X., & Shen, H. T. (2023). ICL-D3IE: In-context learning with diverse demonstrations updating for document information extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 19485–19494). https://doi.org/10.1109/ICCV51070.2023.01785

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... & Kivlichan, I. (2024). Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Janzen, S., Gdanitz, N., Kirchhöfer, M., Spanke, T., & Maaß, W. (2024, May). From Data to Action: A Graph-Based Approach for Decision Support in Civil Protection Operations Planning. In *Proceedings of the International ISCRAM Conference*.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019).

Kumar, A., & Starly, B. (2021). “Fabner”: Information extraction from manufacturing process science domain literature using named entity recognition. Journal of Intelligent Manufacturing, 33(8), 2393–2407. https://doi.org/10.1007/s10845-021-01807-x

Lairgi, Y., et al. (2024). iText2KG: Incremental knowledge graphs construction using large language models.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 260–270). San Diego, California: Association for Computational Linguistics.

Le-Phuoc, D., Mau Quoc, H. N., Quoc, H. N., Tran Nhat, T., & Hauswirth, M. (2016). The graph of things: A step towards the live knowledge graph of connected things. Journal of Web Semantics,37–38,25–35.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems,* 33, 9459-9474.

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1), 50-70.

Li, X., Henriksson, A., Duneld, M., Nouri, J., & Wu, Y. (2023). Evaluating embeddings from pre-trained language models and knowledge graphs for educational content recommendation. Future Internet, 16(1), 12.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... & Clark, P. (2024). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3–26.

Oettinger, M. A., Schatz, D. G., Gorka, C., & Baltimore, D. (1990). RAG-1 and RAG-2, adjacent genes that synergistically activate V (D) J recombination. Science, 248(4962), 1517–1523.

Opdahl, A. L., Al-Moslmi, T., Dang-Nguyen, D.-T., Gallofré Ocaña, M., Tessem, B., & Veres, C. (2022). Semantic knowledge graphs for the news: A review. ACM Computing Surveys, 55(7), Article 140, 38 pages. https://doi.org/10.1145/3543508

Perez, E., Kiela, D., & Cho, K. (2021). True few-shot learning with language models. Advances in Neural Information Processing Systems, 34, 11054–11070.

Peters, M. E., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1756–1765). Vancouver, Canada: Association for Computational Linguistics.

Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. *In The Semantic Web–ISWC 2013: 12th International Semantic Web Conference*, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I 12 (pp. 542-557). Springer Berlin Heidelberg.

Sheng, A. (2024). From language models to practical self-improving computer agents. ArXiv, abs/2404.11964. https://arxiv.org/abs/2404.11964

Shrivastava, M., Seri, K., & Wagatsuma, H. (2022). A named entity recognition model for manufacturing process based on the BERT language model scheme. In Lecture Notes in Computer Science (pp. 576–587). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-24667-8_50

Tamla, P., Freund, F., & Hemmje, M. L. (2020). Supporting named entity recognition and document classification in a knowledge management system for applied gaming. In Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) (pp. 108–121). https://doi.org/10.5220/0010145001080121

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. ArXiv preprint arXiv:2210.03629. https://arxiv.org/abs/2210.03629

Zou, Y., Huang, Y., Wang, Y., Zhou, F., Xia, Y., & Shen, Z. (2024). The Construction of Urban Rainstorm Disaster Event Knowledge Graph Considering Evolutionary Processes. *Water*, 16(7), 942. https://doi.org/10.3390/w16070942

Zscheischler, Jakob & Martius, Olivia & Westra, ... & Edoardo. (2020). A typology of compound weather and climate events. Nature Reviews Earth & Environment. 1. 1-15. 10.1038/s43017-020-0060-z.