Constructing Domain-Specific Knowledge Graphs From Text: A Case Study on Subprime Mortgage Crisis[1]

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Ali Hur[a,*], and Naeem Janjua[b]
[a]*School of Science Edith Cowan University, Joondalup, WA, Australia*
[b]*Flinders University - Adelaide, South Australia*

Abstract. This research paper details a novel methodology for constructing a domain-specific Knowledge Graph (KG) from unstructured text data, exemplified by a case study on the subprime mortgage crisis. The authors present a five-phase approach – specification, conceptualization, formalization, integration, and augmentation – to transform unstructured financial news articles from the MEANTIME corpus into a structured KG. This framework enables the extraction of valuable insights, revealing trends, correlations, and complex relationships among companies, market movements, and economic indicators. The KG's efficacy is demonstrated through its ability to answer complex queries related to the subprime mortgage crisis, highlighting its potential as a powerful tool for knowledge representation and decision-making in the financial domain.

Keywords: Knowledge Graph, Text Mining, Natural Language Processing, Stock Market, Information Extraction, Semantic Modeling, Subprime Mortgage Crisis

1. Introduction

The proliferation of unstructured textual data across diverse domains has created a significant demand for advanced methodologies to transform this data into structured and actionable knowledge. By leveraging natural language processing (NLP) techniques, knowledge can be extracted by transforming the text into structured representations, such as vectors, tensors, and KGs [1]. KG is a data structure that represents knowledge through interconnected nodes and edges, where nodes describe entities and edges represent relationships. This human and machine-readable structure enables AI applications in question answering [2], semantic search [3], recommendations [4], and more by expressing detailed semantics and facilitating complex reasoning to derive new insights. KGs have become an essential tool for representing and interlinking information semantically as nodes and edges.

Existing methodologies to KG construction are predominantly constrained by their domain-specific focus, which limits their flexibility and generalizability for KG construction for other domains. Therefore, current domain-specific KG construction methodologies face several challenges due to their inherent specificity. Among the most pressing issues are the lack of sufficient training data and the absence of robust domain-specific ontologies, both of which are critical for effective KG

[1] Footnote in title. Please ensure there is a 24 pt blank line before the title!

[*]Corresponding author. E-mail: editorial@iospress.nl. Check if the checkbox in menu *Tools/Options/Compatibility/Lay out footnotes like Word 6.x/95/97* is selected if you make a footnote for the corresponding author.

construction. These limitations often result in systems that are tailored to particular fields, restricting their adaptability to new domains and necessitating extensive modifications for different contexts. This domain-centric methodology's limitation not only hinders the broader applicability for KGs construction, but also incurs significant costs associated with customization and maintenance. As a result, the potential of KGs to serve as versatile tools for various AI applications is significantly undermined.

To address these limitations, this research introduces a novel, domain-agnostic framework for the autonomous construction of Knowledge Graphs from unstructured text. The proposed methodology is designed to develop KGs that can be enriched with domain-specific types through rule-based typification, leveraging the underlying rich linguistic context to enhance semantic accuracy. This approach ensures that the system remains versatile and adaptable to various domains without requiring substantial reconfiguration.

The effectiveness of this framework is demonstrated through a case study focused on the stock market domain. By utilizing a subset of the MEANTIME corpus, which contains news stories related to the stock market, the research showcases the framework's capability to generate domain-specific KGs from a foundational domain-agnostic model. This case study not only illustrates the practical application of the proposed methodology but also highlights its flexibility and scalability in real-world scenarios.

This research presents significant contributions to KG construction and semantic modeling. It introduces a robust stock market ontology that underpins accurate, domain-specific KG creation. The study also advances a formalization method utilizing Labeled Property Graphs (LPGs) to capture complex semantic relationships. A key contribution is the development of an autonomous, domain-agnostic KG construction framework, which can be enhanced with domain-specific types through rule-based typification. The framework's efficacy is demonstrated through a case study on market turbulence caused by the subprime mortgage crisis, utilizing a subset of the MEANTIME corpus. The research also outlines a structured five-phase approach to KG construction and employs question answering-based validation to empirically assess the framework's accuracy and performance.

The rest of this article is organized as follows. Section 2 provide a background about a KG construction. Section 3 reviews related work in Knowledge Graph construction and semantic modeling. Section 4 presents a case study on market turbulence related to the subprime mortgage crisis, providing context for the framework's application. Section 5 offers an overview of the KG construction process, detailing the transition from text extraction to graph formation. Section 6 describes the solution overview, outlining the components and architecture of the proposed system. Section 7 explains the methodology implemented in the system, while Section 8 discusses the implementation and validation processes, including empirical results and analysis. Section 9 concludes the paper with a summary of findings and directions for future research.

2. Background

Big enterprises construct and utilize KGs to fulfill their knowledge requirements and facilitate the development of intelligent downstream applications [5]. KGs are constructed either manually or through automated processes. Manual approaches, such as CYC [6], typically involve the expertise of highly qualified domain experts and knowledge engineers. Alternatively, KGs can be manually constructed by leveraging communities and crowdsourcing efforts, as seen with projects like Conceptnet [7], Wikidata [8], Freebase [9], and others [10]. However, manual construction of KGs is labour-intensive, susceptible to human biases, and prone to errors. Recognizing these challenges, the research community has dedicated significant efforts over the past two decades to devising and inventing techniques [11], [12], [13], aimed at automating various aspects of KG construction.

Due to the vast amount of publicly accessible information on the internet, substantial efforts [14], [15], [16], [17] have been devoted to collecting or harvesting knowledge from these sources. Much of the content available on the internet is unstructured text or semi-structured data, such as tables, trees, spreadsheets, etc., providing a rich source of available knowledge. Over the past decade, there have been significant advancements in knowledge harvesting and information extraction. Techniques like text mining and natural language processing have been developed to extract information from these sources in the form of triples, resulting in extraction graphs [18] or data graphs [13]. While these extraction techniques can gather candidate facts in the form of entities, their attributes, and the relations between them, the

resulting graph often contains redundant, invalid, and inconsistent facts due to the noisy nature of the sources [19]. Moreover, it often lacks necessary background knowledge, semantic descriptions of entities, and relationships, rendering it insufficient to meet the knowledge requirements of enterprises [20]. According to [13], 99% of existing KGs are data graphs.

In [21], it was found that only 19% of the triples in the NELL-995 KG are correct with regards to the NELL schema. Manual removal of these errors is both costly and time-consuming, as exemplified by NELL's use of periodic human supervision, which is prohibitively expensive. Hence, automating these tasks is imperative [22]. Even teams of experts developing and maintaining enterprise-level KGs face similar challenges [23]. To identify and rectify these errors, correctness schemes have been proposed as discussed in [24]. Similarly, there is a significant amount of missing information that needs to be addressed to ensure the completeness of the KG. Coverage and correctness pose major challenges, even for large KGs [25][23]. According to a study mentioned in [26], nearly half of the entities in DBPedia [27] have fewer than five relationships. Another study revealed that the birthplace of over 70% of individuals in Freebase is unknown, and 90% have no mention of ethnicity [28]. In response to such issues, various techniques for KG completion have been proposed in the literature.

3. Related Work

When handling unstructured text, techniques such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging, parsing, and sentiment analysis are employed to convert it into structured data. Unlike structured sources, unstructured text requires advanced information extraction methods to derive valuable insights. These techniques are crucial for enriching and expanding the KG [1], [29], [30], [31], [32], [33]. Unstructured text data is challenging due to ambiguity, linguistic complexities, and noise. Advanced information and knowledge extraction techniques [34], [35], [36], [37] leverage linguistic analysis, pattern recognition, and machine learning to transform this data into a structured format for Knowledge Graph integration.

Knowledge extraction distills knowledge from both structured and unstructured data sources [37]. It involves NLP components such as entity recognition, linking, and relation extraction [36], aiming to convert text into a machine-interpretable format for automated reasoning. This process relies on knowledge representation, which organizes information into entities, concepts, and relations, often structured within an ontology. Knowledge extraction enables the creation of KGs that support advanced reasoning and semantic analysis.

IE involves converting natural language text into structured forms, typically in the form of binary or higher-arity relations. *Extractors* such as [1], [38] are instrumental in retrieving information from various types of sources like text, HTML documents, and human-annotated elements. Techniques employed in information extraction utilize natural language processing and theories of computational linguistics [39], [40], [41], [42], [43], [44] to annotate or label segments of text. A significant portion of information extraction revolves around the identification and characterization of entities, relations, and events [45].

IE entails the extraction of structured information from unstructured data sources, primarily text documents [46]. The core objective of IE is to identify specific entities, relationships, or events within predefined domains or schemas. This process relies on predefined patterns, templates, or machine learning models that are customized to extract particular types of information, such as names of individuals, organizations, and their corresponding relationships. As a result, IE techniques are inherently focused on specific domains or tasks, operating within the constraints of a predefined schema. This stringent adherence to a specific schema characterizes the operational framework of these techniques [47].

In contrast, Open Information Extraction (Open IE) [48] adopts a flexible and schema-agnostic approach, seeking to extract relationships or propositions from text without being bound by predefined schemas or templates. Instead of relying on specific structures, Open IE systems take a more generic approach, extracting basic constituents of sentences by identifying subject, predicate, and object spans. These systems utilize unsupervised or semi-supervised methods, relying on linguistic analysis and statistical techniques to identify and extract relationships or propositions from text without the need for a prior knowledge of specific relations or entities.

Among these techniques, Open IE tools have been pivotal [48], [49] . Open information extraction tools operate using a set of patterns, which are either manually crafted [50], [51] or automatically learned through pattern mining [1], [52], [53] and machine

learning techniques [54], [55], [56]. These tools are tasked with producing propositions from text in the form of tuples. These tuples focus on predicates, subject arguments, and object arguments. However, in their early iterations [32], [48], these techniques were limited to basic tuple extraction, lacking the richness of contextual details such as modality, polarity, factuality, and attribution.

As OpenIE techniques evolved, they encompassed contextual analysis, producing tuples infused with contextual intricacies through frameworks like OLLIE [49]. Recent advancements, exemplified by OpenIE5, have expanded the capabilities further to generate n-ary relations within sentences. Additionally, techniques based on frame semantics assign roles to arguments, enhancing context through specific categorizations. These techniques, associated with semantic role labeling (SRL) (Shi & Lin, 2019) based on schemes such as PropBank [57], FrameNet [58], or VerbNet [59], identify text spans in sentences and assign roles to them.

Despite the diversity of methods employed, a shared limitation is evident—they predominantly focus on individual sentences, confining their semantic scope solely to the sentence level. As a result, these techniques face a challenge in uncovering patterns that encapsulate the discourse structure embedded in multi-sentence or multi-clausal text. A discourse represents a sequential unfolding of eventualities, encompassing both states and events described within a text [60]. In essence, the structure of the discourse is crafted to portray and convey information pertaining to diverse states and events that occur within the contextual framework of the text.

To model interconnected propositions and capture a broader context while addressing various aspects, the outputs produced by these tools necessitate several post-processing steps. The application of advanced NLP techniques in this post-processing phase involves incorporating more complex linguistic features as input. Each extracted feature represents a specific linguistic phenomenon. These post-processing steps encompass activities such as coreference resolution, identification of identical mentions, discourse segmentation, entity recognition, among others. Techniques like PIKES [61] and NEWSREADER [62] strive to consolidate these outputs, channeling results into standardized annotation formats such as NAF [63], [64]. While this harmonizes representation compatibility, it primarily caters to data integration and does not address the intricacies of linguistic analysis.

Graph-based representations play a crucial role in modeling the broader context, particularly when aiming to capture complex structures [65], [66], [67], [68]. Their effectiveness lies in their ability to excellently model topological features i.e., patterns of connections and overall structure of the (sub)graph [68]. Existing graph-based linguistic analysis approaches have garnered significant interest for their focus on graph-based linguistic analysis and representation [69] [70]. However, these approaches often lack comprehensive semantic content that considers the specific domain or task they were tailored for, as they primarily concentrate on modeling the required semantic or syntactic elements. Approaches like PROPS [71], PredPatt [72] and pyBART [73] employ directed graph-based representations, leveraging dependency parse information for sentence portrayal. For instance, pyBART aims to provide a representation that is useful for downstream NLP tasks. The proposed structure consists of labeled, directed multi-graphs, where nodes represent the words of a sentence, and labeled edges indicate the relations between them. This structure is generated from dependency parse information, with additional information added beyond its dependency label. However, these methods are confined to sentence-level extraction of predicate arguments. [53] provides a graph-based text representation that is constrained to sentence level and focuses on word senses and semantic frames, tailored towards detecting causal relations. However, it does not encompass a comprehensive range of other essential semantic elements, and it does not demonstrate linguistic analysis specific to its representation. Likewise, techniques such as those presented by [54], focus on capturing non-local and non-sequential dependencies, but their scope remains limited to exposing these dependencies.

Furthermore, [54] showcase graph-based linguistic analysis tailored to specific tasks like entity recognition and relation extraction. However, their applicability is constrained by a lack of comprehensive context, rendering them unfit for domain-agnostic and task-agnostic utilization.

Additionally, it's important to note that most existing approaches [62], [74], [75], [76] rely on RDF for graph data representation. RDF, while widely used, presents several limitations [77], [78] in the context of complex textual data. It often results in sparse graphs with limited structural detail, particularly when handling intricate relationships within diverse real-world scenarios. This can lead to

challenges in accurately representing and analyzing complex text-based knowledge. Another issue faced by RDF is the challenge of reified statements. While they provide additional information, they can result in slower graph traversal and a substantial increase in the serialization size of the graph [77]. In contrast, the current literature didn't explore the LPG for text representation, which offers superior flexibility and adaptability, mitigating these RDF limitations for advanced knowledge representation. Moreover, LPG excels in executing scalable graph analytical tasks such as sub-graph matching, network alignment, and real-time KG querying. It distinguishes itself with efficient storage, rapid traversal capabilities, and the versatility to model various real-world domains [77].

4. Case Study: Market Turbulence Due to the Subprime Mortgage Crisis

The news story titled Markets Dragged Down by Credit Crisis was published on August 10, 2007. It provides an account of the market downturn triggered by the subprime mortgage crisis.

4.1. Story and Character: Bob, the Financial Analyst

Let's assume, Bob is a seasoned financial analyst at Alpha Investments, a firm specializing in portfolio management and investment advisory services, was tasked with advising his firm on investment strategies during the turbulence caused by the subprime mortgage crisis. As the news of the credit crisis broke on August 10, 2007, Bob needed to quickly assess the situation and provide actionable insights to his clients. Alpha Investments managed assets worth billions of dollars, and its clients relied on timely and accurate advice to make informed investment decisions. In such a high-stakes environment, the ability to interpret market conditions and news events quickly and accurately was crucial. Bob had previously relied on manually examining news articles and financial reports to gather the necessary information, but this approach was becoming increasingly impractical as the volume of information exploded.

The announcement of the credit crisis came as a shock to the financial world. Major news outlets were flooded with reports, analyses, and opinions. Bob faced the daunting task of sifting through a massive amount of unstructured text to extract relevant information about the market turmoil. The sheer volume of information made it impossible for him to manually process and analyze all the data in a timely manner. Bob needed a system that could automatically interpret and structure this unstructured data to provide clear and actionable insights. Without the KG, Bob would have faced several challenges:

a) Volume of Information: Manually sifting through countless news articles and reports would be time-consuming and prone to errors.
b) Speed: Timely decision-making is crucial in financial markets. Delays in information processing could result in missed opportunities or increased risks.
c) Comprehensive Analysis: Manually correlating information from various sources to understand relationships and trends would be nearly impossible within a short time frame.
d) Contextual Understanding: Extracting contextual relevance and connecting the dots between different entities, events, and actions is challenging without an automated system.

This case study highlights the need for an autonomous, pipeline-based framework for constructing domain-specific KGs, especially in domains where timely and accurate information is critical for decision-making.

5. Overview of Knowledge Graph Construction

The process of constructing a knowledge graph comprises a systematic arrangement of sequential components, each contributing crucial information to subsequent stages, thereby delineating a sequential progression in the construction pipeline. This intricate procedure can be visualized as a series of construction stages, where the output of each stage acts as foundational input for the subsequent phases, as illustrated in Figure 1. The last two phases, integration and augmentation, are typically automated. The augmentation phase consists of two sub-phases: enrichment and refinement. The enrichment phase employs knowledge extraction strategies to expand the current knowledge graph, which includes an information extraction pipeline.

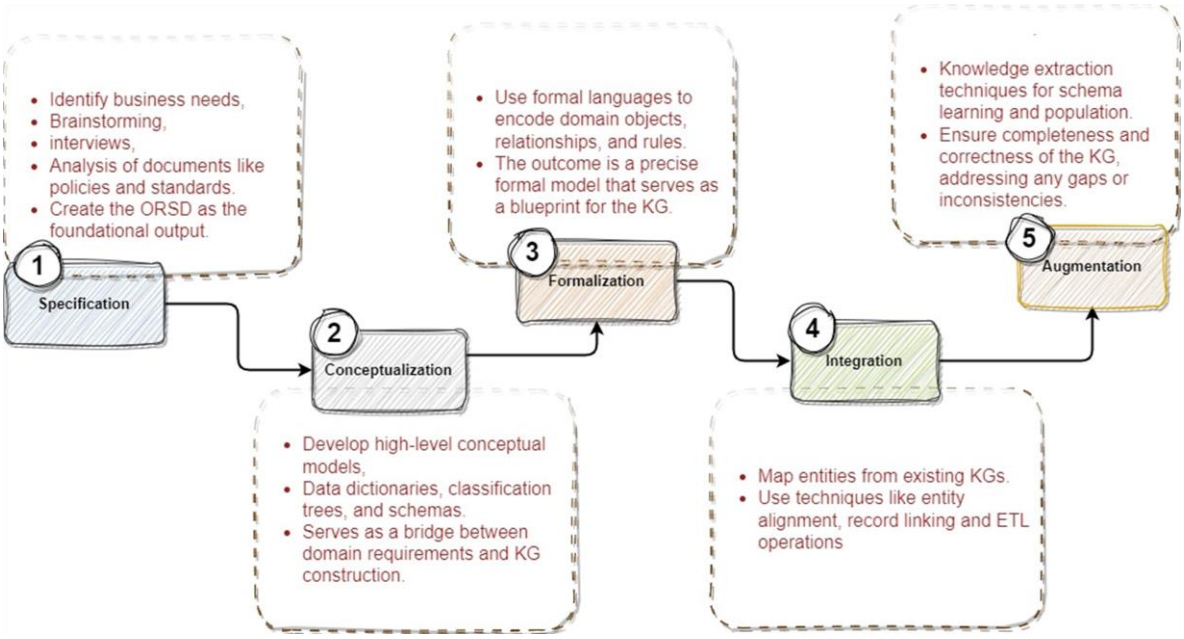

Figure 1: Knowledge Graph Construction Process

5.1. Specification Phase

The specification phase serves as the foundational bedrock, defining the purpose, scope, and intricacies governing the construction of a KG [79], [80]. It involves defining the problem that the Knowledge graph aims to solve. It encapsulates a comprehensive elucidation of the overarching motivations driving the KG's creation, portraying its intended domain specifications and a meticulous set of both functional and non-functional requirements [81], [82], [83].

The success of the knowledge graph heavily relies on the thoroughness and accuracy of the specifications [84], [85]. A well-defined specification document ensures that the resulting KG aligns closely with the requirements of the chosen domain, leading to a high-quality representation that accurately reflects the underlying information. The specification phase is, therefore, a foundational step in achieving the construction of a meaningful and effective KG from a specification document [86].

This critical phase commences with an exhaustive exploration of the business requirements, necessitating a multifaceted approach toward *knowledge acquisition*. Various methodologies are employed, ranging from intensive brainstorming sessions [87] and targeted interviews to comprehensive analyses of textual documents such as policy documents and user manuals. Moreover, a wide array of information sources is tapped into, leveraging existing datasets, ontologies, standards documents, dictionaries, taxonomies, and legal frameworks to holistically gather domain-specific insights. This diverse amalgamation of sources enriches the understanding of the domain landscape and helps in delineating the multifaceted aspects that contribute to the KG's structure.

The process entails a series of steps to transform the gathered requirements into coherent and structured documentation. A pivotal artifact emerging from this phase is the *Ontology Requirement Specification Document (ORSD)* a comprehensive compendium particularly crafted in natural language [85]. The ORSD serves as a foundational repository, encapsulating various facets such as explicit description of business constraints and requirements, vivid use case descriptions, an all-inclusive explanation of the project's overall scope, an exhaustive glossary comprising domain-specific terms, categorization of related terms, and a set of competency questions. The inclusion of intermediate representations aids in crystallizing these requirements, offering a structured view of the intricate domain facets and their interrelationships.

However, a deeper exploration into how these requirements are prioritized, and refined, and how conflicts or inconsistencies are mitigated within the ORSD would further enrich the comprehension of this

foundational phase in the construction of a knowledge graph.

5.2. Conceptualization Phase

The conceptualization phase follows the foundational specification phase and plays a pivotal role in structuring and formalizing the amassed domain knowledge into a coherent and structured framework [88]. The conceptualization phase involves the creation of an abstract representation of the domain of interest, achieved by defining a cohesive set of interconnected concepts [89]. This phase is primarily dedicated to translating the gathered domain knowledge, elucidated during the specification phase, into tangible and structured *intermediate representations* [90].

Central to the objective of the conceptualization phase is the creation of a structured framework that encapsulates domain-specific knowledge in a manner that facilitates subsequent stages of the KG construction. The conceptualization phase encompasses the creation of a data dictionary, concept classification trees, attribute classification trees, and other key elements essential for structuring the Knowledge Graph [88], [91].

This phase essentially functions as a bridge between the identified domain requirements and the eventual construction of the KG and often referred to be part of *Ontology Design* phase [83]. Its core activities involve a meticulous examination and articulation of the problem statement and its corresponding solution, predominantly structured around domain-specific vocabularies and terminologies. Key activities within this phase encompass the identification and delineation of data schemata, attributes, classes, relationships, and the interconnections between various data schemas.

The primary outcome of the conceptualization phase is the creation of a high-level conceptual representation that serves as the foundational backbone of the ensuing KG construction. This representation comprises essential concepts, their interrelationships, and connections, thereby laying the groundwork for the subsequent expansion and enrichment of the KG. The resulting high-level conceptual framework serves as a guiding blueprint, providing a structured and minimalistic yet comprehensive foundation for the KG. It encapsulates the core concepts essential for the KG's architecture, facilitating subsequent growth, expansion, and refinement.

5.3. Formalization Phase

The subsequent phase following the conceptualization stage is the formalization phase[83], [88], [92], which represents a crucial step in the KG construction process. This phase is centered on the conversion of the conceptual model, derived from the preceding phases, into a formal and machine-interpretable representation [91]. Central to the formalization phase is the transformation of the high-level conceptual model, delineated in the earlier stages, into a structured and precise machine-interpretable model. Various formal languages such as OWL (Web Ontology Language), Description Logic, Framenets, RDFS (Resource Description Framework Schema), among others, serve as instrumental tools for this transformation. These formal languages facilitate the precise encoding of the conceptual domain model, enabling a machine-understandable representation of domain objects, their relationships, and associated rules and constraints. Knowledge engineers and domain experts collaborate in this phase to orchestrate the translation of the conceptual model into a formal representation. Their expertise and qualifications are pivotal in ensuring the accuracy, completeness, and adherence to domain-specific nuances during this transformation process.

The outcome of the formalization phase is a particularly crafted formal model that encapsulates domain objects, their interrelationships, and a set of defined rules and constraints. This formal model serves as a machine-interpretable blueprint of the domain, facilitating the computational processing and manipulation of domain-specific knowledge within the KG framework. The formal model generated in this phase acts as the foundation for the subsequent stages of the KG construction process. Its precision and adherence to formal languages enable seamless interoperability, semantic structuring, and scalability within the KG.

5.4. The Integration Phase

The subsequent phase in the KG construction pipeline involves extending the formal model by integrating a diverse array of existing knowledge sources [91]. These sources, identified during the specification stage, encompass a spectrum of established knowledge graphs such as Dbpedia [27], LinkedGeoData [93], Freebase [9], Yago [94], Cyc [6], Babelnet [95], WordNet [96], Wiktionary [97], and Conceptnet [98]. The integration process typically

involves mapping entities from these external sources into the formalized model of the evolving knowledge graph. The integration of existing knowledge sources serves as a cornerstone for enriching the comprehensive knowledge representation within the constructed graph. Various techniques, prominently including entity and schema alignment, are deployed to seamlessly fuse the relevant segments of these external knowledge repositories with the developing KG.

The diverse nature of existing knowledge sources entails different integration approaches. For instance, if the data within a source is structured as a graph, the objective revolves around linking the pertinent sections of this graph with the evolving knowledge graph. Conversely, when dealing with relational database management system (RDBMS) sources, the focus lies in establishing connections between specific nodes/entities within the KG and corresponding records within the RDBMS. This process, known as record linking, entails establishing direct linkages between specific entities or nodes within the knowledge graph and the precise records within the RDBMS. To facilitate the integration process, Extract, Transform, Load (ETL) operations are often employed, particularly when converting existing data into RDF (Resource Description Framework), a format conducive to KG representation. ETL operations serve as a pivotal mechanism for converting and harmonizing diverse data formats into a uniform RDF format, thereby enabling seamless integration into the evolving KG. The formalization and integration phase of the KG construction pipeline involves the complex yet essential process of extending the formal model by integrating diverse external knowledge sources.

5.5. Augmentation Phase

The preceding phases lay the groundwork by establishing domain specifications, schemas, standard vocabularies, and integrating existing datasets to form the foundational bedrock for constructing a factual base within the knowledge graph. After these preparatory stages, the subsequent phase, termed as augmentation becomes paramount. This phase encompasses two pivotal aspects which are as follow:

e) Enrichment and Expansion: Involves activities related to knowledge extraction [36], [38], [99], [100], [101], encompassing schema learning and population.

f) Refinement: Emphasizes ensuring the completeness [102] and correctness [21], [103] of the KG.

6. The Solution Overview

The proposed solution, as illustrated in Figure 2, systematically outlines the autonomous construction of domain-specific knowledge graphs from unstructured text. This approach, divided into two key phases, is tailored for advanced AI applications.

6.1. Phase 1: Pipeline-Based Comprehensive Text Representation

The initial phase, aptly named the *linguistic analysis phase*, stands as the bedrock for the creation of high-quality text representations. Within this phase, we embark on an exhaustive exploration of linguistic knowledge, traversing multiple layers that encompass document structure, morphology, syntax, semantics, and pragmatics.

Now, let's further delve into the layers that constitute this phase, as they work in concert to craft the foundation for the proposed advanced AI-driven knowledge graphs.

6.1.1. Document Preprocessing

At the forefront of this phase, the journey begins with document preprocessing components. These components fulfill a multitude of vital tasks, encompassing grammatical error correction and spelling error rectification. They also record invaluable document-level metadata, such as creation date, authorship, and URL sources. Furthermore, they model and segment the document's layout, providing critical insights into the document's structure, including paragraphs, sentences, sections, and more.

6.1.2. Syntactical, Morphological, and Lexical Analysis

The second component of this phase delves into the intricacies of syntactical, morphological, and lexical analysis. This includes fundamental tasks like tokenization, part-of-speech tagging, morphological attribute extraction, word sense disambiguation, and the generation of dependency and constituency parses. Additional responsibilities in this component involve

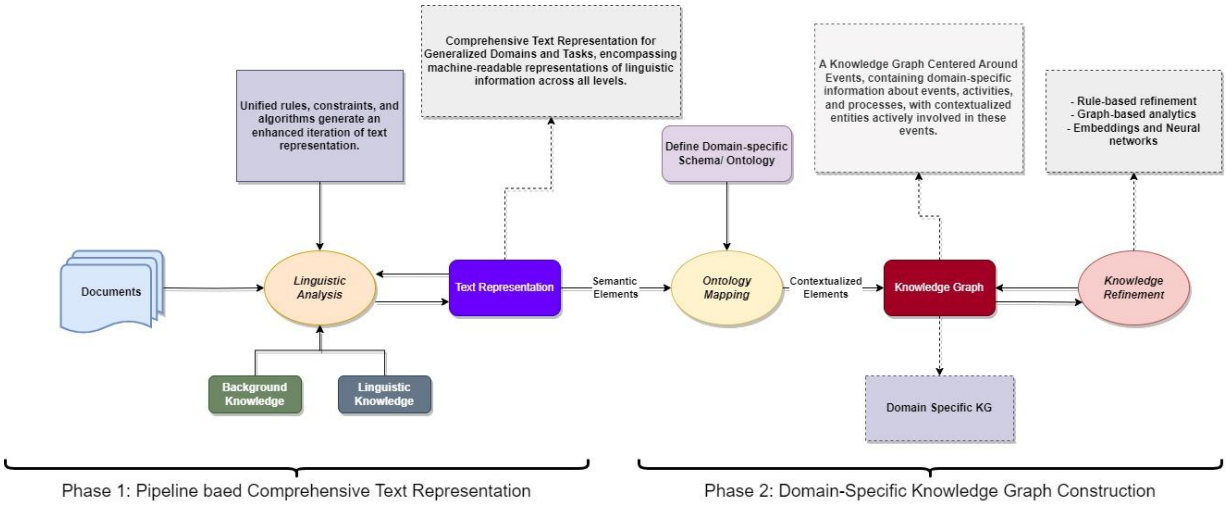

Figure 2: Illustration of the Two-Phase Approach for Constructing Domain-Specific Knowledge Graphs

chunking and headword identification, adding depth and structure to the text representation.

6.1.3. Semantic Analysis

As we ascend to the third component, the semantic analysis layer comes into focus. Building upon the insights derived from the previous components, this segment of our pipeline delves into discourse analysis and coreference resolution, an essential task to discern when pronouns or references point to the same entity. The layer also excels in event detection and extraction, identifying pivotal events within the text and differentiating between them. Furthermore, it undertakes the creation of entity instances and the disambiguation and linking of entities, ensuring that each entity is accurately represented in the knowledge graph. The temporal dimension is addressed through temporal expression recognition and normalization (TERN), a crucial task to understand the temporal context of events and entities. Semantic role labeling and event participant extraction ensure that entities and their roles are accurately captured. Nominal mention detection, named entity recognition and classification (NERC), and pronominal mention detection further enrich the semantic analysis layer.

6.2. Phase 2: Domain-Specific Knowledge Graph Construction

In this phase, we focus on constructing domain-specific knowledge graphs by building on the comprehensive text representation generated in Phase 1. Key activities include:

– Integrating the text representation with a specified schema aligned with semantic elements.
– Applying domain-specific ontology to ensure the knowledge graph's relevance and accuracy.
– Conducting ontology mapping to align the text representation with the domain ontology, ensuring the knowledge graph reflects domain intricacies.
– Enriching the process with linguistic and common-sense knowledge resources like WordNet, ConceptNet, and DBpedia, which enhance the graph's depth and understanding.

This methodical approach produces knowledge graphs that are not merely data structures but robust, domain-specific resources. In essence, the proposed solution represents a thorough orchestration of linguistic analysis, semantic insight, and domain-specific ontology mapping.

6.3. Solution Architecture

The proposed solution overcomes existing challenges in semantic relation extraction and KG construction by introducing a multi-layered architecture focused on advanced semantic models and comprehensive linguistic analysis. It integrates semantic relations within a unified context and employs flexible graph formats like semantic property graphs for scalability and precision. Figure 3 illustrates the architecture of the proposed system, which lays the groundwork for more accurate knowledge extraction from text. Detailed descriptions of each layer are provided in the following sections.

6.3.1. Text Layer

This layer focuses on handling unstructured and semi-structured text. It encompasses different types or genres of text, including document collections, and involves processes such as indexing and metadata extraction, covering aspects like author, date/time, URL, topic, and genre.

6.3.2. Linguistic Analysis Framework

The Linguistic Analysis Framework, residing within the Context Construction Layer, orchestrates linguistic analysis components. It performs various levels and types of linguistic analysis, emphasizing graph-based representation using LPG. This layer involves orchestration, data normalization, quality assurance, and integration, facilitating advanced querying, navigation, and traversal operations within the context layer.

6.3.3. Context Layer

This layer handles the storage and retrieval of graph-based contextual information using a graph database, specifically in LPG format. The Context Layer ensures a unified, cohesive, and comprehensive representation, spanning all linguistic levels extracted during the analysis.

6.3.4. KG Construction Layer

Comprising both domain-agnostic and domain-specific KG construction, this layer disintegrates semantic elements in the former and enriches them with domain-specific ontology in the latter. The two-phase KG construction approach ensures a robust and accurate representation of semantic relationships.

This layered architecture enables the seamless transition from raw textual data to a sophisticated Knowledge Graph, overcoming deficiencies in existing methods. The incorporation of graph-based representation, linguistic analysis orchestration, and a dedicated KG construction layer ensures a holistic solution for extracting and representing semantic relations in textual data.

7. Methodology

This section outlines the approach for constructing a stock market KG using the MEANTIME corpus, following the five-phase methodology described in Section 4, and tailored to the challenges of financial news data.

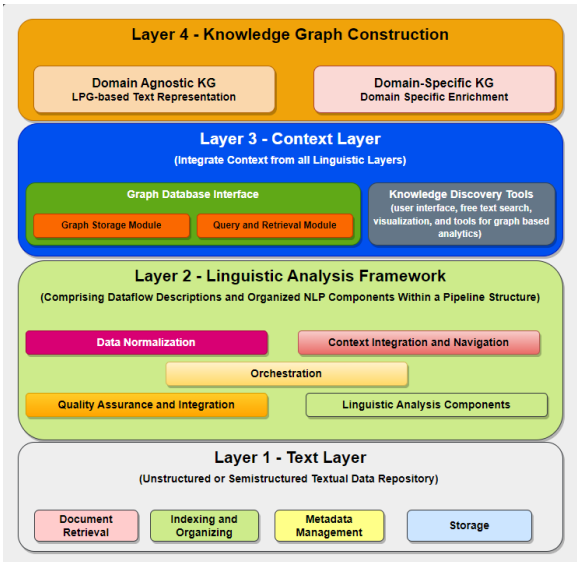

Figure 3: Architectural Overview for Automated Knowledge Graph Construction from Textual Data

7.1. Specifications Phase

The specification phase stands as the genesis in constructing a purpose driven KG from the MEANTIME corpus, a repository abundant with news articles cantered on the stock market domain. This phase refrains from detailed methodological exploration focusing instead on defining pivotal elements and objectives. This phase includes problem statement, definition of competency questions, and use cases.

7.1.1. Problem Statement

The challenge in the financial domain is to develop a comprehensive KG that can effectively synthesize and organize extensive stock market news data. The KG will facilitate the extraction of valuable insights, identification of patterns, and prediction of market dynamics by capturing key entities such as companies, stocks, indices, and events like mergers and IPOs.

7.1.2. Competency Questions

Constructing a domain-specific KG from stock market news involves capturing detailed financial information. Competency questions ensure the KG meets user needs, covering various types and dimensions, including:

General Financial Queries:

– Market Trends: Understanding the overarching movements in various financial markets like stock, bond, or commodity markets over a

specific period. This involves identifying patterns, shifts, or trends within these markets.
- Indices Performance: Examining the performance of major stock indices such as the S&P 500, Dow Jones, or NASDAQ. This includes tracking their movements, fluctuations, and overall performance.
- Major Financial Events: Keeping abreast of significant events that impact the financial world. This might involve major economic policy changes, geopolitical shifts, or industry-specific occurrences.

Company-Specific Inquiries:

- Company Reports: Analyzing and summarizing a company's financial statements, including its income statement, balance sheet, and cash flow statement. These reports offer insights into a company's financial health and performance.
- Mergers and Acquisitions: Monitoring announcements and impacts of mergers, acquisitions, or divestitures within the corporate landscape. Understanding how these actions affect the companies involved and their industries.
- Investor Sentiments: Gauging the feelings and attitudes of investors toward a particular company. This sentiment analysis could be positive, negative, or neutral, impacting stock prices and market perceptions.

Impact Analysis:

- Effects of Financial Policies: Assessing the implications of fiscal and monetary policy changes on various sectors, industries, and the economy as a whole. This includes analyzing how interest rate changes, tax policies, or government stimulus affect markets.
- Bailouts and Market Fluctuations: Examining the impacts of financial bailouts on specific industries or markets. Understanding how economic crises or market fluctuations influence companies and investors.

Predictive Analytics:

- Forecasting Market Trends: Predicting future market movements, industry shifts, and potential trends based on historical data, economic indicators, and current market conditions.
- Investment Opportunities: Identifying potentially profitable investment avenues or areas showing growth potential based on market analysis and economic projections.
- Risk Assessment: Evaluating potential risks associated with investment decisions, economic shifts, or policy changes. This involves understanding and quantifying risks to make informed investment choices.

Table 1 outlines the competency questions based on the identified types and dimensions, guiding the next phase focused on actionable scenarios in the finance domain.

Table 1: Exhaustive List of Competency Questions for a Stock Market based Knowledge Graph

Category	Inquiry
General Financial Information	What are the current stock prices of [Company X]?
	How did the stock market indices perform today?
	What are the trending financial news articles for the week?
Market Trends and Analysis	Which sectors experienced the most growth last month?
	What were the top-performing stocks in [Industry Y] last quarter?
	How has the market responded to recent economic policy changes?
Company-Specific Inquiries	What are the recent financial reports of [Company Z]?
	Has [Company A] announced any mergers or acquisitions recently?
	What is the market sentiment towards [Company B]?
Financial Events and Announcements	Are there any upcoming IPOs in the tech industry?
	Have there been any significant layoffs in the banking sector?
	What are the major events affecting the energy market this month?
Comparative and Statistical Analysis	How does the performance of [Company C] compare to its competitors?
	Can you provide a comparative analysis of stock prices between two specific dates?

	What is the correlation between interest rate changes and stock market performance?
Investment Insights and Risk Assessment	Which stocks are recommended for long-term investment?
	What are the risk factors associated with investing in [Industry D]?
	Can you provide historical volatility data for a specific stock?
Regulatory and Economic Inquiries	What are the implications of recent fiscal policy changes on the housing market?
	How has the trade war impacted international markets?
	What are the regulatory changes affecting the banking sector?
Predictive Analysis and Forecasting	What are the predicted market trends for the next quarter?
	Can you forecast the potential impact of an interest rate hike?
	Based on historical data, what is the projected growth rate for a particular industry?
Geopolitical and Global Financial Trends	How have global economic events affected the local stock market?
	What are the investment opportunities arising from geopolitical shifts?
	Can you provide insights into the economic impact of natural disasters on financial markets?
News and Sentiment Analysis	What is the sentiment analysis of recent news articles related to [Market X]?
	How do financial news sentiments correlate with stock price movements?
	Can you summarize the sentiment of articles regarding a specific company's performance?

7.1.3. Use Cases

The following section highlights practical use cases derived from the competency questions, showcasing real-world applications and scenarios that demonstrate the relevance and utility of the identified areas of expertise. Below are the high-level use cases:

Investment Analysis

Linked to “Company-Specific Inquiries” and “Predictive Analytics”: Assessing stock performance falls under company-specific inquiries, where understanding a company's financial reports aids in evaluating its potential as an investment. Identifying investment opportunities is part of predictive analytics, where forecasts and trends hint at promising investment avenues.

Market Trends Evaluation

Correlated with “General Financial Queries”: Analyzing market movements aligns with the broader understanding of market trends, connecting closely with general financial queries related to tracking indices' performance and identifying major financial events. The impact of these events on stocks falls within this purview.

Risk Management

Tied to "Investment Insights and Risk Assessment": Assessing risks associated with investments based on market conditions directly relates to the risk assessment aspect of investment insights. Understanding market fluctuations, policy impacts, and potential market trends aids in evaluating and managing risks associated with investment decisions.

Connecting these aspects provides a holistic view of how these competencies intersect and contribute to different categories of competency questions within the domain.

7.2. Conceptualization Phase

In the stock market KG construction, the conceptualization phase involves cataloging terms and entities to form a coherent semantic structure. High-level semantic categories are identified and organized into domain-specific clusters, with relationships such as "ACQUISITION," "SELL," and "PARTNERSHIP" defining the domain's structure. The schema is refined through cross-validation with established sources like FIBO and input from domain experts, ensuring flexibility and adaptability through a "middle-out" approach for iterative refinement.

7.2.1. Entities within the Financial Domain

The foundation of our schema design is a precise categorization of entities within the financial domain. These entities span a multifaceted spectrum:

Financial Entities

The schema includes a diverse array of financial institutions, such as banks and investment firms, which form the core of the financial landscape.

Additionally, it encompasses an expansive suite of financial instruments—stocks, bonds, and commodities—and considers the diverse personas of market participants, including investors and traders, who actively influence market dynamics.

Economic Indicators

Key economic indicators are integrated into the schema to reflect the broader economic health. These indicators include GDP, inflation rates, employment statistics, and trade balances, which serve as crucial metrics influencing decision-making across the financial domain.

Market Events

The schema also accounts for significant market events that impact financial landscapes, such as market crashes, regulatory shifts, IPOs, and mergers and acquisitions. These events are critical inflection points that shape market trajectories.

7.2.2. Relationships and Associations

Beyond entities, the schema captures the intricate relationships and associations that characterize the dynamic financial ecosystem:

Financial Relationships

The schema delineates complex relationships within the financial domain, including ownership structures, investment portfolios, strategic partnerships, and transactional linkages. These relationships illuminate the interconnected nature of financial interactions.

Economic Interconnections

It highlights the interdependencies between economic indicators and financial entities, showcasing how shifts in economic metrics influence financial entities and market events, emphasizing the symbiotic relationship between economic health and financial vigor.

Temporal Relationships

Temporal associations are integral to the schema, capturing the chronology, duration, and recurring patterns of market events. These relationships provide insights into the periodic occurrences and their impact on the financial landscape.

7.2.3. Attributes and Properties

The schema design goes beyond entities and relationships—it meticulously delineates their inherent properties and defining characteristics.

Financial Attributes

This aspect includes specific characteristics of financial entities, such as market capitalization, asset valuations, and profit margins. These attributes offer a quantitative perspective on financial entities, highlighting their economic significance.

Economic Metrics

The schema integrates properties associated with economic indicators, such as GDP growth rates and inflation percentages. These metrics are vital for understanding economic trends and their implications for the financial domain.

Event Properties

The schema also catalogs the properties of market events, including event timestamps, levels of impact, and other qualitative and quantitative characteristics. These properties capture the essence and repercussions of pivotal market occurrences.

7.2.4. Overview of Conceptual Elements

The conceptualization phase culminates in the structured representation of diverse entities, relationships, and attributes, as detailed in the following tables:

- Table 2 provides a comprehensive overview of the diverse range of entities and relationships crucial to the stock market landscape. It categorizes financial entities, market indicators, market events, ownership structures, market transactions, market influences, and temporal aspects, offering a granular view of the domain.
- Table 3 lists related terms for each label within the financial domain, serving as a reference for assigning domain-specific labels to semantic elements in the KG. This exhaustive list ensures consistency and accuracy in the representation of financial concepts.
- Table 4 outlines domain-specific labels within the financial domain, providing definitions and relevant examples. These labels are critical for the precise categorization of entities and events, ensuring that the KG accurately reflects the details of the financial landscape.

Table 2: Diverse Range of Entities and Relationships Crucial in the Stock Market Landscape

Schema Element	Subcategory	Description	Examples

Entities within the Stock Market Domain	Financial Entities	Encompasses entities specifically within the stock market, including publicly traded companies (Apple Inc., Microsoft), stock exchanges (NYSE - New York Stock Exchange, NASDAQ), financial institutions (Goldman Sachs, Morgan Stanley), and market participants such as investors, traders.	NYSE (New York Stock Exchange), Apple Inc., Warren Buffet
	Market Indicators	Includes key market indicators such as stock prices, market indices (S&P 500, Dow Jones Industrial Average), trading volumes, market capitalization, and sector-specific indicators.	S&P 500 index reaching 4,000 points, Apple's stock price at $150 per share
	Market Events	Spans significant market occurrences: earnings reports (quarterly financial results), IPOs (Initial Public Offerings), mergers, acquisitions, and stock splits.	Apple's quarterly earnings announcement, Tesla's IPO, Disney's acquisition of Fox
Relationships and Associations	Ownership Structures	Illustrates ownership connections: majority/minority stakes held by one company in another, parent-subsidiary relationships.	Berkshire Hathaway's ownership in Coca-Cola, Google's acquisition of YouTube
	Market Transactions	Portrays trading transactions: stock purchases, sales, block trades, and institutional buying/selling patterns.	Block trade of Amazon stocks, Institutional buying of Tesla shares
	Market Influences	Indicates factors influencing stock prices: economic indicators, geopolitical events, regulatory changes, and corporate actions.	Interest rate changes impacting stock markets, Apple's product launch affecting its stock price
Temporal Aspects and Events	Earnings Releases	Chronicles chronological releases: quarterly earnings announcements, annual reports, investor meetings.	Apple's Q4 earnings release, Tesla's annual report presentation
	Market Movement Patterns	Describes movement trends: daily price fluctuations, market volatility, trading patterns (bearish, bullish trends).	Stock market volatility during economic downturns, Bullish trend in tech stocks
	Event-Specific Details	Provides details about specific market events: event dates, impact on stock prices, announcement details.	Amazon's Prime Day impact on stock prices, Google's new product announcement

Table 3: Exhaustive List of Related Terms for each Label within the Financial Domain.

Label	Related Terms or Words
FinancialActivity	Stock Trading, Investment, Portfolio Management, Asset Allocation, Capital Allocation, Trading Strategy, Equity Investment, Bonds, Derivatives Trading, Forex Trading, Asset Management, Share Trading, Investment Strategy, Wealth Management, Commodities Trading, Futures Contracts, Securities Trading, Algorithmic Trading, High-Frequency Trading, Options Trading, Risk Management.
FinancialIndicator	Stock Price, Market Index, Interest Rate, Dividend Yield, Bond Yield, Volatility Index, Price-to-Earnings Ratio, Consumer Price Index, Gross Domestic Product (GDP), Unemployment Rate, Stock Market Index, Bond Rating, Inflation Rate, Exchange Rate, Treasury Yield, Credit Spread, Mortgage Rate, Yield Curve, Commodity Price Index, Leading Economic Index, Retail Sales Index

EconomicActivity	Trade Relations, GDP Growth, Consumer Spending, Industrial Production, Business Investment, Foreign Direct Investment, Trade Deficit, Trade Surplus, Economic Development, Employment Rate, Housing Starts, Business Sentiment, Consumer Confidence, Retail Sales, Business Investment, Factory Orders, Trade Balance, Export-Import Volume, Manufacturing PMI.
EconomicPolicy	Monetary Policy, Fiscal Policy, Interest Rate Policy, Tax Policy, Budgetary Policy, Regulatory Policy, Economic Stimulus, Inflation Targeting, Interest Rate Decision, Quantitative Easing, Fiscal Stimulus, Central Bank Intervention, Austerity Measures, Tax Reform, Tariff Policy, Trade Agreement, Regulatory Reform, Budget Deficit Reduction
EconomicSituation	Recession, Inflation, Deflation, Market Volatility, Economic Downturn, Economic Recovery, Stagflation, Hyperinflation, Economic Stability, Economic Indicators, Economic Recession, Economic Recovery, Economic Slowdown, Boom-Bust Cycle, Deflationary Pressures, Economic Expansion, Economic Stagnation, Economic Resilience, Fiscal Imbalance, Debt Crisis.
EconomicEntity	Financial Institution, Multinational Corporation, Investment Bank, Commercial Bank, Hedge Fund, Sovereign Wealth Fund, Credit Rating Agency, Central Bank, Pension Fund, Investment Firm, Venture Capitalist, Angel Investor, Mutual Fund, Credit Union, Brokerage Firm, Insurance Company, Financial Regulator, Sovereign Debt Holder, Corporate Bondholder, Institutional Investor.
GeographicRegion	Developed Economies, Emerging Markets, Developed Countries, Developing Nations, Global Regions (North America, Asia-Pacific, Europe, etc.), Economic Zones, Free Trade Zones, Emerging Markets, Developing Economies, Global Economies, Regional Blocs (EU, ASEAN, NAFTA), Economic Zones, Special Economic Zones, Economic Corridors, Economic Blocs, Economic Alliances

Table 4: List of Domain-Specific Labels within the Financial Domain

Domain Label	**Definition**	**Example**
FinancialActivity		
StockTrading	The buying, selling, or exchange of stocks on financial markets.	Purchasing shares of a publicly listed company on the New York Stock Exchange.
Investment	Allocating capital into financial assets or securities for future returns.	Acquiring bonds or mutual funds for long-term profit.
PortfolioManagement	Supervising and adjusting an investment portfolio to achieve financial objectives.	Balancing asset allocation to mitigate risks in an investment portfolio.
FinancialIndicator		
StockPrice	The current or historical value of a share in a company.	Apple Inc.'s stock price surged by 10% in a single trading day.
MarketIndex	A statistical measure representing the overall performance of a group of stocks.	The S&P 500 index tracks the stock performance of 500 large companies listed on U.S. stock exchanges.
InterestRate	The cost of borrowing money or the return on investment.	Central banks may adjust interest rates to influence economic growth.
EconomicActivity		
TradeRelations	The commercial interactions and agreements between nations or entities.	Negotiating trade agreements between the U.S. and China.

GDPGrowth	The increase in a country's Gross Domestic Product (GDP) over time.	India's GDP growth rate reached 7% in the last fiscal year.
ConsumerSpending	The total expenditure by individuals on goods and services.	A surge in holiday season consumer spending boosted retail sales.
EconomicPolicy		
MonetaryPolicy	Governmental control of money supply and interest rates.	Central banks may adjust interest rates to manage inflation.
FiscalPolicy	Government decisions on taxation, spending, and borrowing.	Implementing tax cuts to stimulate economic growth.
EconomicSituation		
Recession	A significant decline in economic activity for a sustained period.	The 2008 financial crisis led to a global recession.
Inflation	The rise in prices of goods and services over time.	High inflation rates reduce the purchasing power of a currency.
MarketVolatility	Rapid or unpredictable changes in market prices.	The stock market experienced heightened volatility during the pandemic.
EconomicEntity		
FinancialInstitution	Organizations providing financial services.	Banks, investment firms, and insurance companies.
MultinationalCorporation	Companies operating in multiple countries.	Coca-Cola and IBM are multinational corporations.
GeographicRegion		
DevelopedEconomies	Countries with well-established industrial and economic infrastructure.	The United States, Germany, and Japan are developed economies.
EmergingMarkets	Nations undergoing rapid economic growth and industrialization.	China and India are often cited as emerging market economies.

7.3. Formalizing Conceptual Representation into Labeled Property Graphs

In constructing the stock market KG as an LPG, we assigned node labels and edge types based on the classes and relationships defined during the conceptualization phase. Each node represents an entity or event, with properties capturing their attributes. While we focus on straightforward label assignments, schema-based reasoning can be enabled by defining explicit relationships like IS-A. Constraints and rules are implemented using the graph database and query language to ensure accurate representation and retrieval.

Figure 6: Ontology Illustration for Stock Market Domain: A Hierarchical Representation of Entities, Aspects, and Economic Factors illustrates the OWL-based ontology, which provides a hierarchical schema for stock market entities, market aspects, and economic factors. This ontology enhances the LPG-based formalization by adding semantic precision and formal structure. It facilitates semantic interoperability and supports schema-based reasoning when required.

7.4. Integration Phase

In Knowledge Graph construction, the integration phase traditionally involves incorporating structured knowledge sources—such as relational databases or RDF datasets—into the developing KG. This process utilizes techniques like R2ML or ETL operations to enhance the KG with structured information.

However, in our research, we deliberately exclude the integration of external structured resources. Our focus is on constructing the Knowledge Graph exclusively from unstructured textual content. This approach underscores the capability of our methodology to independently generate a structured KG directly from unstructured data. While this choice does not negate the importance of conventional integration practices, it illustrates the robustness and self-sufficiency of our approach in effectively

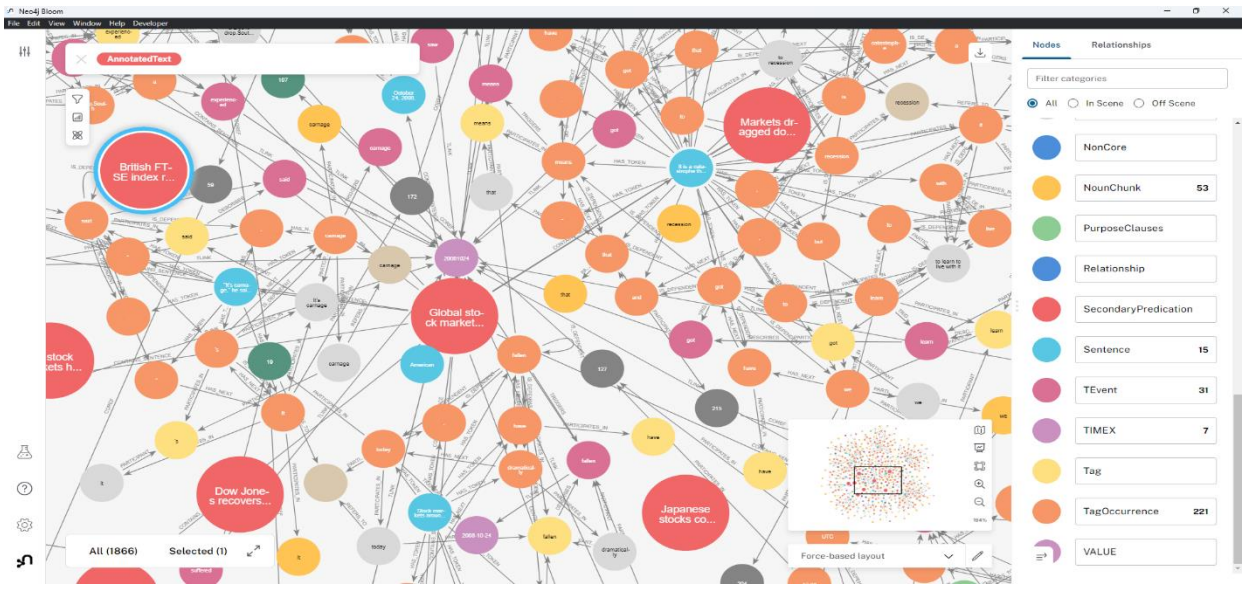

Figure 4: Graph based Text Representation of News Stories Generated by The Proposed System

transforming unstructured text into a coherent Knowledge Graph.

7.5. Augmentation Phase – Extraction and Refinement

The Augmentation Phase concludes our Knowledge Graph construction by evolving the initial domain-agnostic KG into a domain-specific model tailored to the stock market. This phase builds upon the extraction and refinement processes, as described in [104], to enhance the generalized KG with domain-specific details.

Initially, the Extraction phase creates a domain-agnostic KG from unstructured stock market news using an LPG-based text representation. This representation includes entities, events, temporal elements, and relationships like TLINKs and CLINKs.

7.5.1. Enriching Semantic Elements in the LPG-Based Text Representation

In this phase, the objective is to transition the domain-agnostic Knowledge Graph, initially established through the LPG-based text representation, into a refined, domain-specific model tailored for the stock market. This process involves enhancing the graph's semantic elements—entities, events, and relationships—with contextually relevant information specific to the financial domain.

Entity Labelling and Classification

The initial domain-agnostic graph contains various entities, including nominal, pronominal, and named entities. This phase involves matching these entities with terms from a domain-specific dictionary developed during the Specification and Conceptualization phases. This matching process assigns domain-specific labels and categories to each entity, refining their representation and aligning them with the financial context. Attributes associated with these entities are similarly matched to ensure precise classification and enhance the granularity of the entity representation.

Event Typing

Events within the Knowledge Graph are classified according to their relevance to the stock market domain. This involves analyzing event descriptions and predicates and matching them with entries in the domain-specific dictionary. By attributing domain-specific types to these events, the process ensures their contextual alignment with financial activities, thereby improving the interpretability of event-related information within the Knowledge Graph.

Relationship Labeling

In this phase, relationships between entities and events are identified and labeled to reflect real-world financial associations, such as investor-company relationships and economic indicator impacts. The relationships are matched against terms in the domain-specific dictionary, ensuring that they accurately

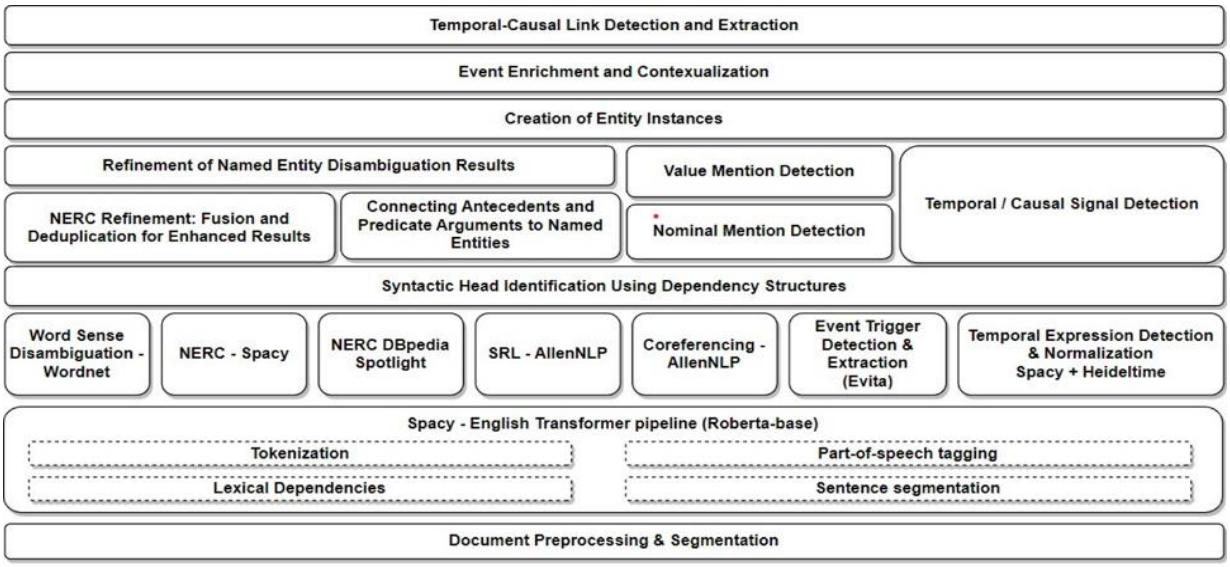

Figure 5:Graph-based NLP Framework for Syntactic and Semantic Graph Generation from Text

represent connections within the stock market domain. This labeling process enhances the relevance and accuracy of the relationships in the Knowledge Graph, emphasizing their significance within the financial context.

8. Implementation and Validation

This section presents the implementation and validation of the proposed framework. The source code of the implementation can be found on the github[2] .

8.1. Phase 1: Construction of Text Representation

The proposed LPG-based text representation system is developed following a pipeline-based approach, integrating various components and techniques to generate a comprehensive and semantically rich graph representation. The system was implemented in Python 3 and utilized the Neo4j graph database for efficient storage and querying of the LPG graph. Figure 5 presents the layered architecture of the proposed framework implementation. The implementation of our proposed approach employs cutting-edge third-party NLP components for diverse tasks, as indicated in Figure 5. Figure 4 illustrates the text graph, showcasing the nodes and edges representing entities and their relationships extracted from the unstructured text data.

8.1.1. Validation of LPG based Text Representations

The validation results of the LPG-based text representations from Phase 1 of the proposed system indicate a generally strong performance in extracting and normalizing various textual elements. The system demonstrates effective entity extraction, with an F1 score of 0.694, and excels in recognizing numeric values, achieving an impressive F1 score of 0.940. Temporal expression detection is also highly accurate, with an F1 score of 0.943, showcasing the system's strength in handling time-related information. However, the system shows room for improvement in certain areas, such as event recall (F1 score of 0.719) and precision in entity instances (F1 score of 0.550), where enhancing precision and recall would be beneficial. The performance in extracting event participants is moderate (F1 score of 0.634), and the system's ability to capture temporal links between events or expressions is notably weaker, with an F1 score of 0.339, highlighting the need for further refinement in these aspects. Overall, while the system is effective in many key areas, targeted improvements could enhance its overall accuracy and reliability.

[2] https://github.com/neostrange/text2graphs

8.2. Phase 2: Construction of Domain-Specific KG

The objective of the implementation phase is to apply the established methodology to demonstrate the construction and enrichment of the KG. This involves showcasing how the domain-specific KG functions in answering relevant questions for financial analysts during market turbulence.

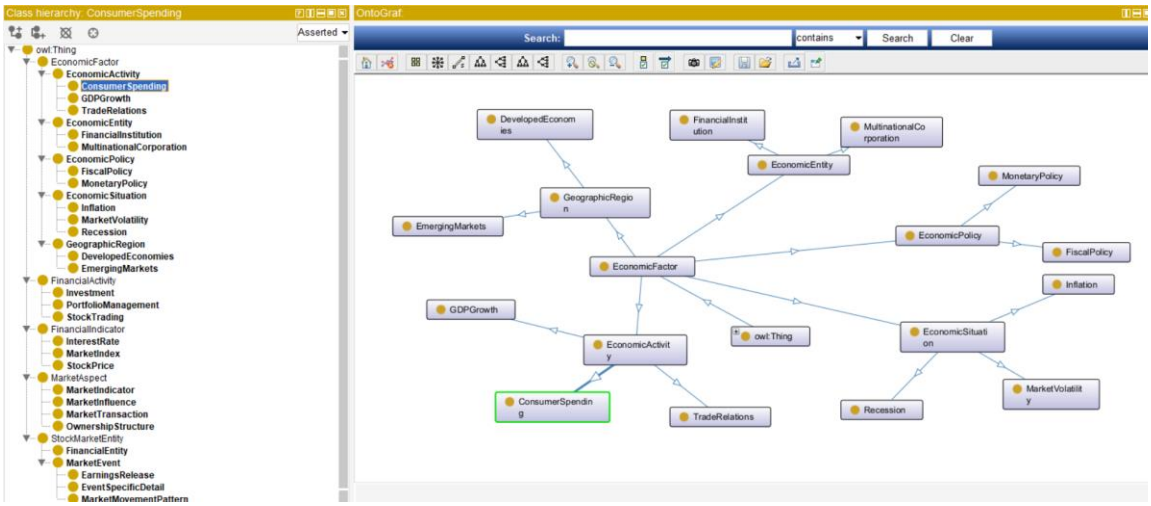

Figure 6: Ontology Illustration for Stock Market Domain: A Hierarchical Representation of Entities, Aspects, and Economic Factors

8.3. Process

8.3.1. Formulate Competency Questions

Competency questions are crafted to ensure the KG meets the needs of financial analysts. Examples include identifying key entities, relationships, and events relevant to stock market news.

8.3.2. Query the KG

Once the KG is built, it is queried using competency questions to extract relevant information. This step verifies the KG's ability to provide accurate and timely responses.

8.3.3. Evaluate Responses

Responses from the KG are assessed for accuracy, completeness, and relevance, confirming that the KG effectively captures and represents the necessary domain knowledge.

8.4. Implementation of Semantic Enrichment Scheme

8.4.1. Overview of Semantic Enrichment Process

Phase 2 involves transforming the domain-agnostic text representation into a domain-specific KG by labeling entities and events with relevant types. This phase uses a rule-based approach to enhance the KG with domain-specific semantics. The rules leverage contextual cues from the text to assign accurate labels, demonstrating the application of domain knowledge.

8.4.2. Enrichment of Entities

Entities identified in the text are labeled with domain-specific types using a Domain Dictionary. For example, "mortgage" is classified as "Loan," and "Jim Cramer" is labeled as "Person."

Entities are matched with dictionary terms, and labels are assigned based on context. For instance, "CNBC" is categorized as "Media" due to its role in financial reporting.

8.4.3. Enrichment of Events

Event enrichment involves identifying triggers (e.g., "fell," "surged") and integrating these with the pre-assigned entity labels. Domain-specific rules are used to label events based on context and associated entities.

Table 5: Rules for Domain-Specific Entity Enrichment

Generic Entity	Domain-Specific Label

Mortgage	Loan
American Home Mortgage	Financial Institution
CNBC	Media
Federal Reserve	Regulatory Body
Dow Jones Industrial Average	Financial Indicator

Table 6: Rules for Domain-Specific Event Enrichment

Rule Type	**Verbs**	**Contexts**	**Arguments**
MarketMovement	fell, surged, plummeted	Global Stock Markets, Dow Jones	ARG0: Economic Crisis, ARG1: Market
FinancialRescue	injecting, adding	Federal Reserve, European Central Bank	ARG0: Regulatory Body, ARG1: Funds, ARG2: Amount
RegulatoryAction	authorized, decided	Federal Reserve, European Central Bank	ARG0: Regulatory Body, ARG1: Policy
InterestRate	raised, adjusted	Federal Funds Rate	ARG1: Monetary Policy Indicator
CorporateEvent	announced, filed	Bear Stearns, Lehman Brothers	ARG0: Financial Institution, ARG1: Corporate Action
EconomicSituation	declining, worsening	US Housing Market	ARG1: Real Estate Market Condition

Example Complex Rule:

Type: MarketMovement
Verbs: ['fell', 'surged']
Contexts: ['GlobalStockMarkets', 'DowJones']
Arguments:
ARG0: ['EconomicCrisis']
ARG1: ['Market', 'StockMarketIndexMovement']

In this rule, a "MarketMovement" event is identified by verbs indicating market fluctuations and contexts related to stock market indices. The arguments specify the roles of entities involved, such as the economic crisis and market indices.

8.5. Validation of Domain Specific Enrichment Process for Stock Market Analysis

To validate the domain-specific enrichment process for stock market analysis, a detailed validation has been conducted on a case study (Table 7) based on a historical news story concerning the global market downturn triggered by the subprime mortgage crisis in the United States. The aim is to demonstrate that the constructed KG can effectively answer questions posed by stock market experts, investors, financiers, and other stakeholders who wish to analyse the stock market dynamics during crises. This section details the validation methodology, queries, and results to demonstrate the efficacy of our KG in addressing queries pertinent to stock market experts, investors, and other stakeholders.

Table 7: Case Study Overview: Market Turbulence Due to the Subprime Mortgage Crisis

Title	Markets Dragged Down by Credit Crisis
Creation Date	August 10, 2007
URL	*https://en.wikinews.org/wiki/Markets_dragged_down_by_credit_crisis*

The validation process involved querying the KG to test its ability to extract and interpret information from the news story. Specific queries were designed to evaluate the KG's proficiency in identifying key entities, relationships, events, and their contextual relevance. The following section lists the queries used to validate the KG constructed after domain-specific enrichment.

8.5.1. Examples Queries Used in a Case Study:

g) Retrieve all financial institutions mentioned in the news.
h) Find the specific categories of financial activities performed by central banks.
i) Identify all events triggered by economic crises.
j) List all market movements and their triggers.

k) Find out all the public statements made by corporate entities.
l) Retrieve all the events related to loan repayment.
m) Get the names of all persons making public statements.
n) List the financial events associated with bailouts.
o) Retrieve all events indicating market condition changes.
p) Find all events involving the Federal Reserve.

8.5.2. Case Study Analysis

The case study demonstrates the KG's ability to extract and interpret crucial information during the subprime mortgage crisis, aiding financial analysts like Bob in making informed decisions.

The constructed KG offered a solution. By transforming unstructured news articles into a structured, semantically rich KG, Bob could efficiently query and analyze the information. Bob's immediate concerns included understanding which financial institutions were most affected, what actions central banks were taking, and how these actions were influencing market movements. Here is how the KG helped Bob:

- Financial Institutions: Bob identified key financial institutions such as Bear Stearns and Lehman Brothers mentioned in the news. Knowing which institutions were involved allowed Bob to assess the potential impact on related stocks and financial instruments.
- Financial Activities by Central Banks: The KG revealed central banks' activities, including liquidity injections and interest rate adjustments. This information was crucial for Bob to understand the broader economic implications and to predict potential market responses.
- Events Triggered by Economic Crises: Bob found significant events like stock market drops and credit tightening. This helped him identify patterns and anticipate future market movements.
- Market Movements and Triggers: The KG provided a detailed list of market movements and their triggers, such as the announcement of the credit crisis causing a significant drop in stock prices. Bob used this information to advise clients on potential buying or selling opportunities.
- Public Statements by Corporate Entities: Bob accessed statements from CEOs and other corporate officials, gaining insights into market sentiment and potential future actions by these companies. For example, knowing that a major bank announced a write-down of subprime assets helped Bob predict further declines in the financial sector.
- Events Related to Loan Repayment: The KG identified events such as mortgage defaults and foreclosure rates, providing Bob with a clearer picture of the underlying issues in the housing market and their impact on financial markets.
- Persons Making Public Statements: Bob retrieved names of influential figures, including Federal Reserve officials and financial analysts, allowing him to follow key voices and interpret their statements' implications.
- Financial Events Associated with Bailouts: The KG highlighted bailout events and government interventions, which were critical for Bob to understand government actions to stabilize the market.
- Market Condition Changes: Bob reviewed events indicating changes in market conditions, such as volatility spikes and trading halts, helping him advise clients on risk management strategies.
- Events Involving the Federal Reserve: The KG detailed Federal Reserve's actions, providing Bob with a comprehensive understanding of policy responses and their likely effects on the market.

8.5.3. Validation Against the Case Study Queries

To validate the KG, a set of competency questions was used to simulate the kinds of queries financial analysts might perform. The responses generated by the KG were then compared against manually annotated data to assess their accuracy and completeness. Table 7-3 presents the queries along with the system’s responses for each, highlighting the effectiveness of the KG in delivering relevant and meaningful information.

8.5.4. Evaluation Metrics

The performance of the Knowledge Graph was evaluated using the following metrics:

Table 8: Precision, Recall, and F-Score Results of the Domain-Specific Knowledge Graph Enrichment Process.

Precision	**Recall**	**F-Score**
0.97	0.87	0.91

These results demonstrate that the KG is highly precise, meaning that most of the retrieved results are relevant and accurately reflect the information present

in the dataset. The recall score indicates that the KG is effective at retrieving a significant portion of relevant information, although there is some room for improvement in capturing all relevant data. The F1-score, which balances precision and recall, confirms the overall effectiveness of the KG in handling complex queries related to the stock market domain.

Table 9: Query based Evaluation for Stock Market Knowledge Graph

Query	**Cypher Version**	**Response**
Retrieve all financial institutions mentioned in the news.	MATCH (fi:FinancialInstitution) RETURN fi.id	The response includes the list of financial institutions mentioned in the news article which include BNP Paribas, Deutsche Bank, American Home Mortgage, Bank of America Home Loans, Washington Mutual, Hedge Fund, Bear Stearns, American Home Mortgage Investment Corp, The Bank of Japan
Find the specific categories of financial activities performed by central banks.	MATCH p= (cb:RegulatoryBody)--(fa:FinancialActivity) WHERE cb.id = 'Central banks across the world' RETURN fa.form, fa.generalCategory, fa.specificCategory	The response shows the list of financial activities performed by Central banks, which includes funds injection, and addition of market liquidity.
Identify all events triggered by economic crises.	MATCH (ec:EconomicCrisis)-[r {type:'ARG0'}]->(e:TEvent) RETURN ec.id, e.form, e.generalCategory, e.specificCategory, r.type	The response lists the events triggered by economic crisis which states that the markets have been dragged down by the credit crisis.
List all market movements and their triggers.	MATCH (mm:MarketMovement) RETURN mm.form, mm.generalCategory, mm.specificCategory	The response lists various events that are indicative of market movements such as dragged, fell, rebounded, dragged, failing, ending, tumbling, etc. Different organizations such as Bear Stearns, Global Stock Market, The UK's FTSE-100 index, Nikkei, and Dow Jones.
Find out all the public statements made by corporate entities.	MATCH g= (fi:FinancialInstitution)-[:PARTICIPANT {type:'ARG0'}]->(e:TEvent {specificCategory:'PublicStatement'})<-[s:PARTICIPANT]-(:TEvent) WHERE s.type = 'ARG1' or s.type= 'ARG2' //RETURN g RETURN fi.id, e.form, e.specificCategory	The response shows the public statement made by Bank of America Home Loans, saying they will be forced to retain a greater proportion of mortgage.
Retrieve all the events related to loan repayment.	MATCH g=(mc:MarketCondition)-[r:PARTICIPANT]-(p) RETURN mc.form, p.id, r.type	The response shows the repay event which indicates a financial activity corresponding to loan repayment.
Get the names of all persons making public statements.	MATCH g = (p:Person)-[:PARTICIPANT]->(e:PublicStatement) RETURN p.id, e.form, e.generalCategory, e.specificCategory	The response shows that the only person making public statements is Jim Cramer. The response list three events, Called 'Feds to lower rates immediately.' Remarked 'that as many as seven million peoples will lose their homes from bad mortgage'.

		Saying ‘that the Fed was asleep’.
List the financial events associated with bailouts.	MATCH p= (fe:FinancialRescue)--(e:TEvent) WHERE fe.id CONTAINS 'bail-out' RETURN fe.id as name, e.form as event, e.generalCategory as GeneralCategory, e.specificCategory as SpecificCategory	The response shows the bail-out event which corresponds to the category of Financial Rescue activity.
Retrieve all events indicating market condition changes.	MATCH g=(mc:MarketCondition)-[r:PARTICIPANT]-(p) RETURN mc.form, p.id, r.type	The response shows the list of events that indicate market condition changes.
Find all events involving the Federal Reserve.	MATCH g= (p WHERE p.id CONTAINS 'Federal_Reserve')--(e:TEvent) RETURN p.id, e.form, e.specificCategory, e.generalCategory	The response shows all the events involving Federal Reserve, and includes: Transferred US$ 24 Billion Raised Interest rates. Entered repurchase agreement. Decided to Maintain its target rate 5.25%

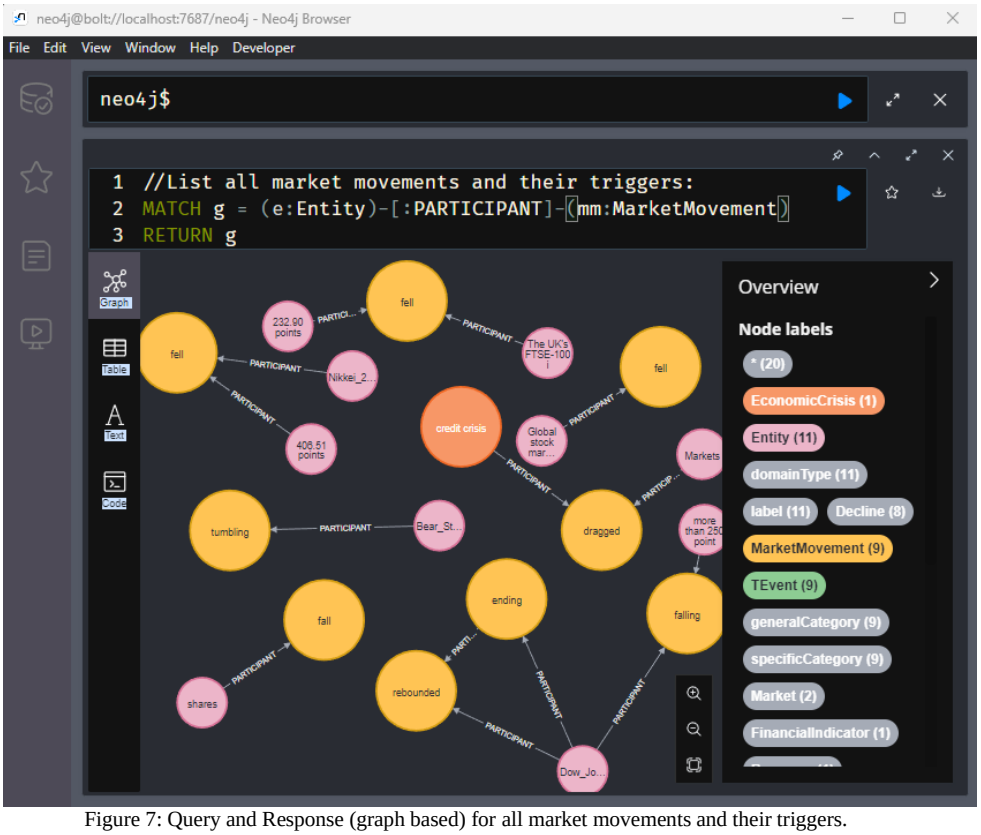

Figure 7: Query and Response (graph based) for all market movements and their triggers.

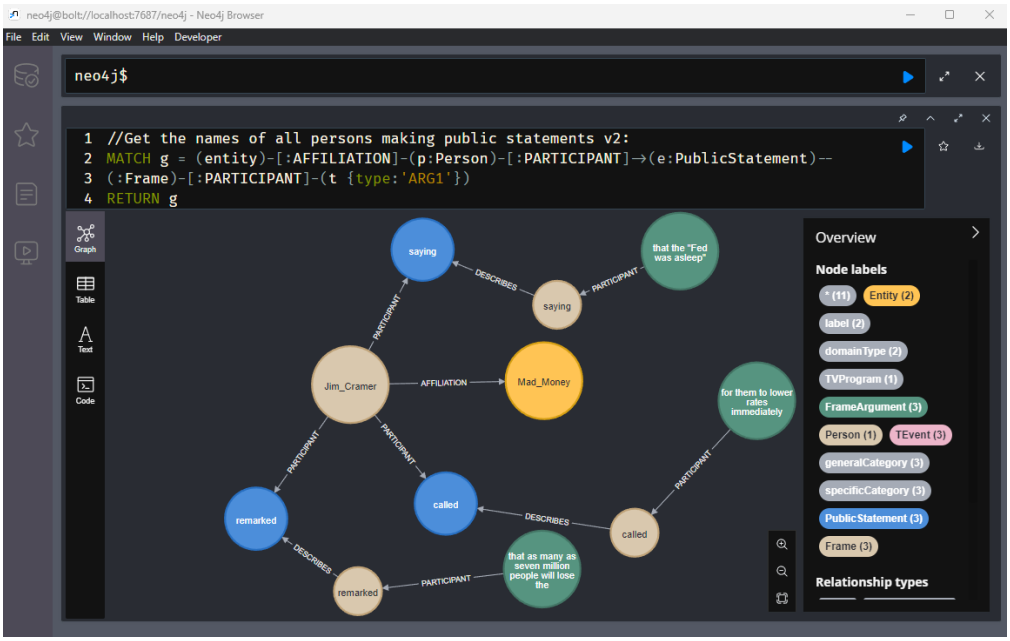

Figure 8: Query and Response (graph based) for "the names of all persons making public statements."

9. Conclusion

This research article has introduced a robust framework for the construction of domain-specific Knowledge Graphs (KGs), with an emphasis on the stock market domain. The methodology, articulated through a structured five-phase approach, has been demonstrated to effectively capture and represent the complex relationships inherent in stock market data. By integrating linguistic, semantic, and formal techniques, the framework ensures a high degree of accuracy in entity recognition, event extraction, and semantic enrichment, facilitating a comprehensive representation of domain-specific knowledge. The case study utilizing the MEANTIME corpus underscores the practical utility and efficacy of the constructed KG, revealing its potential to support advanced AI-driven applications such as semantic search, question answering, and predictive analytics within the stock market context. The evaluation metrics indicate that the proposed system performs commendably in recognizing and extracting entities, events, and temporal expressions, with strong precision and recall scores. However, the analysis also highlights areas where the system can be further refined, particularly in the extraction of temporal links and event participants. These findings point to the necessity for continued refinement of the KG construction process to enhance the precision and completeness of the knowledge graph.

9.1. Future Work

Future research will focus on addressing the identified limitations in the current framework, particularly the enhancement of temporal link detection and event participant extraction. To achieve these objectives, we plan to incorporate advanced machine learning models and deep learning techniques that can more effectively discern the complex relationships between temporal elements and event participants. Furthermore, the integration of more diverse and extensive datasets will be pursued to improve the generalizability and robustness of the constructed KG across various stock market scenarios.

Another critical direction for future work is the adaptation of the proposed system to real-time data processing, enabling dynamic updates of the KG as new information becomes available. This capability will be particularly vital for applications requiring real-time decision-making and analytics, such as automated trading systems and financial news analysis tools.

Lastly, extending the framework to other domains beyond the stock market is a key objective. By refining the domain-agnostic components of the system, the methodology can be adapted to construct

KGs across diverse fields such as healthcare, finance, and e-commerce. This extension will broaden the applicability and impact of the research, positioning the proposed framework as a versatile tool for domain-specific knowledge representation and reasoning.

References

[1] X. Dong *et al.*, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 601–610, 2014, doi: 10.1145/2623330.2623623.

[2] X. Lu, A. Abujabal, S. Pramanik, Y. Wang, R. S. Roy, and G. Weikum, "Answering complex questions by joining multi-document evidence with quasi knowledge graphs," *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, no. iv, pp. 105–114, 2019, doi: 10.1145/3331184.3331252.

[3] E. Lupiani-Ruiz *et al.*, "Financial news semantic search engine," *Expert Syst Appl*, vol. 38, no. 12, pp. 15565–15572, Nov. 2011, doi: 10.1016/J.ESWA.2011.06.003.

[4] Y. Xian, Z. Fu, S. Muthukrishnan, G. De Melo, and Y. Zhang, "Reinforcement knowledge graph reasoning for explainable recommendation," *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 285–294, 2019, doi: 10.1145/3331184.3331203.

[5] P. Haase, D. M. Herzig, A. Kozlov, A. Nikolov, and J. Trame, "Metaphactory: A platform for knowledge graph management," *Semant Web*, vol. 10, no. 6, pp. 1109–1125, 2019, doi: 10.3233/SW-190360.

[6] B. L. Douglas, "CYC: A Large-Scale Investment in Knowledge Infrastructure," *Commun ACM*, vol. 38, no. 11, pp. 33–38, 1995, doi: 10.1145/219717.219745.

[7] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge," *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, no. Singh 2002, pp. 4444–4451, Dec. 2017.

[8] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledge base," *Commun ACM*, vol. 57, no. 10, pp. 78–85, 2014, doi: 10.1145/2629489.

[9] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1247–1249, 2008, doi: 10.1145/1376616.1376746.

[10] J. Weng, Y. Gao, J. Qiu, G. Ding, and H. Zheng, "Construction and Application of Teaching System Based on Crowdsourcing Knowledge Graph," *Communications in Computer and Information Science*, vol. 1134 CCIS, pp. 25–37, 2019, doi: 10.1007/978-981-15-1956-7_3.

[11] N. KERTKEIDKACHORN and R. ICHISE, "An Automatic Knowledge Graph Creation Framework from Natural Language Text," *IEICE Trans Inf Syst*, vol. E101.D, no. 1, pp. 90–98, 2018, doi: 10.1587/transinf.2017SWP0006.

[12] G. Wu *et al.*, "An Automatic and Rapid Knowledge Graph Construction Method of SG-CIM Model," *Proceedings - 2020 IEEE International Conference on Smart Cloud, SmartCloud 2020*, pp. 193–198, 2020, doi: 10.1109/SmartCloud49737.2020.00044.

[13] X. Wu, J. Wu, X. Fu, J. Li, P. Zhou, and X. Jiang, "Automatic knowledge graph construction: A report on the 2019 ICDM/ICBK Contest," *Proceedings - IEEE International Conference on Data Mining, ICDM*, vol. 2019-Novem, no. Icdm, pp. 1540–1545, 2019, doi: 10.1109/ICDM.2019.00204.

[14] M. Stewart, M. Enkhsaikhan, and W. Liu, "ICDM 2019 knowledge graph contest: Team UWA," *Proceedings - IEEE International Conference on Data Mining, ICDM*, vol. 2019-Novem, pp. 1546–1551, 2019, doi: 10.1109/ICDM.2019.00205.

[15] N. Kertkeidkachorn and R. Ichise, "T2KG: An end-to-end system for creating knowledge graph from unstructured text," *AAAI Workshop - Technical Report*, vol. WS-17-01-, pp. 743–749, 2017.

[16] F. L. Li *et al.*, "AliMe KG: Domain knowledge graph construction and application in e-commerce," *ArXiv*, 2020.

[17] D. Buscaldi, D. Dessì, E. Motta, F. Osborne, and D. R. Recupero, “Mining scholarly data for fine-grained knowledge graph construction,” *CEUR Workshop Proc*, vol. 2377, pp. 21–30, 2019.

[18] J. Pujara, H. Miao, L. Getoor, and W. W. Cohen, “Using semantics and statistics to turn data into knowledge,” *AI Mag*, vol. 36, no. 1, pp. 65–74, 2015, doi: 10.1609/aimag.v36i1.2568.

[19] P. Cimiano, “Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods,” *Semant Web*, vol. 8, no. 3, pp. 489–508, 2017, [Online]. Available: http://www.semantic-web-journal.net/content/knowledge-graph-refinement-survey-approaches-and-evaluation-methods

[20] S. Razniewski and P. Das, “Structured Knowledge: Have we made progress? An extrinsic study of KB coverage over 19 years,” *International Conference on Information and Knowledge Management, Proceedings*, no. 2, pp. 3317–3320, 2020, doi: 10.1145/3340531.3417447.

[21] K. Wiharja, J. Z. Pan, M. J. Kollingbaum, and Y. Deng, “Schema aware iterative Knowledge Graph completion,” *Journal of Web Semantics*, vol. 65, p. 100616, 2020, doi: 10.1016/j.websem.2020.100616.

[22] X. Chen, S. Jia, and Y. Xiang, “A review: Knowledge reasoning over knowledge graph,” *Expert Syst Appl*, vol. 141, 2020, doi: 10.1016/j.eswa.2019.112948.

[23] N. Noy, Y. Gao, A. Jain, A. Patterson, A. Narayanan, and J. Taylor, “Industry-scale knowledge graphs lessons and challenges,” *Queue*, vol. 17, no. 2, pp. 1–28, 2019, doi: 10.1145/3329781.3332266.

[24] A. Hur, N. Janjua, and M. Ahmed, “A Survey on State-of-the-art Techniques for Knowledge Graphs Construction and Challenges ahead,” 2021, [Online]. Available: https://www.seagate.com/files/www-

[25] G. Weikum, S. Razniewski, L. Dong, and F. Suchanek, “Machine knowledge: Creation and curation of comprehensive knowledge bases,” *ArXiv*, 2020.

[26] B. Shi and T. Weninger, “Open-World Knowledge Graph Completion,” *ArXiv*, pp. 1957–1964, 2017.

[27] J. Lehmann *et al.*, “DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia,” *Semant Web*, vol. 6, no. 2, pp. 167–195, 2015, doi: 10.3233/SW-140134.

[28] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin, “Knowledge base completion via search-based question answering,” *WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web*, pp. 515–525, 2014, doi: 10.1145/2566486.2568032.

[29] F. Wu and D. S. Weld, “Open Information Extraction Using Wikipedia,” 2010, *Association for Computational Linguistics*. Accessed: Nov. 10, 2022. [Online]. Available: https://aclanthology.org/P10-1013

[30] N. Bhutani, H. V. Jagadish, and D. Radev, “Nested propositions in open information extraction,” in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016. doi: 10.18653/v1/d16-1006.

[31] A. Fader, S. Soderland, and O. Etzioni, “Identifying relations for Open Information Extraction,” in *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2011.

[32] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland, “TextRunner: Open Information Extraction on the Web,” in *NAACL-Demonstrations ’07: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Apr. 2007, pp. 25–26. doi: 10.5555/1614164.1614177.

[33] A. Akbik and A. Löser, “KRAKEN: N-ary Facts in Open Information Extraction,” in *AKBC-WEKEX ’12: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, NAACL-HLT, 2012, pp. 52–56. doi: 10.5555/2391200.

[34] K. Gashteovski, R. Gemulla, B. Kotnis, S. Hertling, and C. Meilicke, “On Aligning OpenIE Extractions with Knowledge Bases: A Case Study,” pp. 143–154, 2020, doi: 10.18653/v1/2020.eval4nlp-1.14.

[35] H. Chen, S. Deng, W. Zhang, Z. Xu, J. Li, and E. Kharlamov, “Neural Symbolic Reasoning

with Knowledge Graphs: Knowledge Extraction, Relational Reasoning, and Inconsistency Checking," *Fundamental Research*, no. August, 2021, doi: 10.1016/j.fmre.2021.08.013.

[36] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, D. Garigliotti, and R. Navigli, "Open Knowledge Extraction Challenge," 2015, pp. 3–15. doi: 10.1007/978-3-319-25518-7_1.

[37] J. Unbehauen, S. Hellmann, S. Auer, and C. Stadler, "Knowledge Extraction from Structured Sources," 2012, pp. 34–52. doi: 10.1007/978-3-642-34213-4_3.

[38] D. C. Wimalasuriya and D. Dou, "Components for Information Extraction: Ontology-based information extractors and generic platforms," *International Conference on Information and Knowledge Management, Proceedings*, no. April, pp. 9–18, 2010, doi: 10.1145/1871437.1871444.

[39] A. Kappagoda, "The Use of Systemic-Functional Linguistics in Automated Text Mining," p. 71, 2009.

[40] C. Paradis, "Lexical Semantics," in *The Encyclopedia of Applied Linguistics*, Oxford, UK: Wiley, 2012, pp. 189–200. doi: 10.1002/9781405198431.wbeal0695.

[41] R. D. Van Valin, "Role and reference grammar," *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session*, vol. 37, no. 1, Jan. 1993, doi: 10.31356/silwp.vol37.05.

[42] E. M. Bender, *Linguistic Fundamentals for Natural Language Processing II*, vol. 59. 2022.

[43] V. Ng, E. E. Rees, J. Niu, and A. Zaghlool, "Application of natural language processing algorithms for extracting information from news articles in event-based surveillance," *Canada Communicable Disease Report*, vol. 46, no. 6, pp. 186–191, 2020, doi: 10.14745/ccdr.v46i06a06.

[44] D. Jurafsky, "Representing and integrating linguistic knowledge," pp. 199–204, 1990, doi: 10.3115/997939.997974.

[45] J. Eisenstein, *Introduction to Natural Language Processing. Adaptive Computation and Machine Learning serie.* MIT Press, 2019. Accessed: May 17, 2023. [Online]. Available: https://mitpress.mit.edu/books/introduction-natural-language-processing

[46] J. Piskorski and R. Yangarber, "Information extraction: Past, present and future," *Multi-source, multilingual information extraction and summarization*, pp. 23–49, 2013.

[47] J. R. Hobbs and E. Riloff, "Information Extraction.," *Handbook of natural language processing*, vol. 15, p. 16, 2010.

[48] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A Survey on Open Information Extraction," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.05599

[49] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "Open language learning for information extraction," in *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference*, 2012.

[50] J. Strötgen and M. Gertz, "HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 321–324. [Online]. Available: https://aclanthology.org/S10-1071

[51] L. T. Wu, J. R. Lin, S. Leng, J. L. Li, and Z. Z. Hu, "Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web," Mar. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.autcon.2021.104108.

[52] R. Snow, D. Jurafsky, and A. Y. Ng, "Learning syntactic patterns for automatic hypernym discovery," *Adv Neural Inf Process Syst*, 2005.

[53] B. Rink, C. A. Bejan, and S. Harabagiu, "Learning textual graph patterns to detect causal event relations," *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23*, no. Flairs, pp. 265–270, 2010.

[54] Y. Luan, D. Wadden, L. He, A. Shah, M. Ostendorf, and H. Hajishirzi, "A General Framework for Information Extraction using Dynamic Span Graphs," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.03296

[55] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *2016 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, pp. 260–270, 2016, doi: 10.18653/v1/n16-1030.

[56] R. Agerri and G. Rigau, "Robust multilingual Named Entity Recognition with shallow semi-supervised features," *Artif Intell*, vol. 238, pp. 63–82, Sep. 2016, doi: 10.1016/j.artint.2016.05.003.

[57] M. Palmer, P. Kingsbury, and D. Gildea, "The Proposition Bank: An Annotated Corpus of Semantic Roles," *Computational Linguistics*, vol. 31, no. 1, pp. 71–106, Mar. 2005, doi: 10.1162/0891201053630264.

[58] C. Baker, "FrameNet: A Knowledge Base for Natural Language Processing," in *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1–5. doi: 10.3115/v1/W14-3001.

[59] K. K. Schuler, *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, 2005.

[60] B. Webber, M. Egg, and V. Kordoni, "Discourse structure and language technology," *Nat Lang Eng*, vol. 18, no. 4, pp. 437–490, 2012.

[61] F. Corcoglioniti, M. Rospocher, and A. P. Aprosio, "Extracting Knowledge from Text with PIKES." [Online]. Available: http://pikes.fbk.eu/

[62] P. Vossen *et al.*, "NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news," *Knowl Based Syst*, vol. 110, pp. 60–85, 2016, doi: 10.1016/j.knosys.2016.07.013.

[63] A. Fokkens *et al.*, "NAF and GAF: Linking linguistic annotations," *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, May 26, 2014, Reykjavik, Iceland*, no. January, pp. 9–16, 2014.

[64] N. A. F. Deliverable, "NAF : NLP Annotation Format .," no. 1, pp. 1–49.

[65] I. Ali and A. Melton, "Graph-Based Semantic Learning, Representation and Growth from Text: A Systematic Review," *Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019*, pp. 118–123, Mar. 2019, doi: 10.1109/ICOSC.2019.8665592.

[66] Y. Qian, E. Santus, Z. Jin, J. Guo, and R. Barzilay, "GraphIE: A Graph-Based Framework for Information Extraction," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.13083

[67] M. Vazirgiannis, F. D. Malliaros, and G. Nikolentzos, "Graphrep: Boosting text mining, NLP and information retrieval with graphs," *International Conference on Information and Knowledge Management, Proceedings*, pp. 2295–2296, 2018, doi: 10.1145/3269206.3274273.

[68] A. Broekman and L. Marshall, "Linguistic Inspired Graph Analysis," May 2021, [Online]. Available: http://arxiv.org/abs/2105.06216

[69] V. Nastase, R. Mihalcea, and D. R. Radev, "A survey of graphs in natural language processing," *Nat Lang Eng*, vol. 21, no. 5, pp. 665–698, 2015, doi: 10.1017/S1351324915000340.

[70] A. H. Osman and O. M. Barukub, "Graph-Based Text Representation and Matching: A Review of the State of the Art and Future Challenges," *IEEE Access*, vol. 8, pp. 87562–87583, 2020, doi: 10.1109/ACCESS.2020.2993191.

[71] G. Stanovsky, J. Ficler, I. Dagan, and Y. Goldberg, "Getting more out of syntax with PROPS," *InProceedings of the 2016 Conf erence on Empirical Methods in NaturalL anguage Processing (EMNLP), pages 892–898. Association for Computational Linguistics, November.*.

[72] S. Zhang, R. Rudinger, and B. van Durme, "An evaluation of PredPatt and open IE via stage 1 semantic role labeling," in *12th International Conference on Computational Semantics, IWCS 2017 - Short Papers*, 2017.

[73] A. Tiktinsky, Y. Goldberg, and R. Tsarfaty, "pybart: Evidence-based syntactic transformations for ie," *arXiv preprint arXiv:2005.01306*, 2020.

[74] F. Corcoglioniti, M. Rospocher, and A. P. Aprosio, "Frame-Based Ontology Population with PIKES," *IEEE Trans Knowl Data Eng*, vol. 28, no. 12, pp. 3261–3275, 2016, doi: 10.1109/TKDE.2016.2602206.

[75] J. L. Martinez-Rodriguez, I. Lopez-Arevalo, and A. B. Rios-Alvarado, "OpenIE-based

approach for Knowledge Graph construction from text," *Expert Syst Appl*, vol. 113, pp. 339–355, 2018, doi: 10.1016/j.eswa.2018.07.017.

[76] Z. Moteshakker Arani, A. Abdollahzadeh Barforoush, and H. Shirazi, "Representing unstructured text semantics for reasoning purpose," *J Intell Inf Syst*, vol. 56, no. 2, pp. 303–325, Apr. 2021, doi: 10.1007/s10844-020-00621-w.

[77] S. Purohit, N. Van, and G. Chin, "Semantic Property Graph for Scalable Knowledge Graph Analytics," *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, pp. 2672–2677, 2021, doi: 10.1109/BIGDATA52589.2021.9671547.

[78] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke, and E. Rahm, "Construction of knowledge graphs: State and challenges," *arXiv preprint arXiv:2302.11509*, 2023.

[79] M. C. Suárez-Figueroa, A. Gómez-Pérez, and B. Villazón-Terrazas, "How to Write and Use the Ontology Requirements Specification Document," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5871 LNCS, no. PART 2, 2009, pp. 966–982. doi: 10.1007/978-3-642-05151-7_16.

[80] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," *Computing*, vol. 102, no. 2, pp. 393–411, Feb. 2001, doi: 10.1007/s00607-018-0687-5.

[81] A. Umber, M. S. Naweed, T. Bashir, and I. S. Bajwa, "Requirements elicitation methods," *Adv Mat Res*, vol. 433, pp. 6000–6006, 2012.

[82] F. Ciroku, "Supporting requirement elicitation and ontology testing in knowledge graph engineering," 2023.

[83] M. C. Bravo Contreras, L. F. Hoyos Reyes, and J. A. Reyes Ortiz, "Methodology for ontology design and construction," *Contaduría y Administración*, vol. 64, no. 4, p. 134, Mar. 2019, doi: 10.22201/fca.24488410e.2020.2368.

[84] I. Dubielewicz, B. Hnatkowska, Z. Huzar, and L. Tuzinkiewicz, "Domain modeling based on requirements specification and ontology," in *Software Engineering: Challenges and Solutions: Results of the XVIII KKIO 2016 Software Engineering Conference 2016 held at September 15-17 2016 in Wroclaw, Poland*, 2017, pp. 31–45.

[85] M. C. Suárez-Figueroa, A. Gómez-Pérez, and B. Villazón-Terrazas, "How to write and use the ontology requirements specification document," in *On the Move to Meaningful Internet Systems: OTM 2009: Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009, Vilamoura, Portugal, November 1-6, 2009, Proceedings, Part II*, 2009, pp. 966–982.

[86] A. Brack, A. Hoppe, M. Stocker, S. Auer, and R. Ewerth, "Analysing the requirements for an Open Research Knowledge Graph: use cases, quality requirements, and construction strategies," *International Journal on Digital Libraries*, vol. 23, no. 1, pp. 33–55, Mar. 2022, doi: 10.1007/s00799-021-00306-x.

[87] P. Weichbroth, "Facing the brainstorming theory. A case of requirements elicitation," 2016.

[88] R. de A. Falbo, C. S. de Menezes, and A. R. C. da Rocha, "A Systematic Approach for Building Ontologies," 1998, pp. 349–360. doi: 10.1007/3-540-49795-1_31.

[89] F. Ameri, B. Kulvatunyou, N. Ivezic, and K. Kaikhah, "Ontological conceptualization based on the SKOS," *J Comput Inf Sci Eng*, vol. 14, no. 3, 2014, doi: 10.1115/1.4027582.

[90] C. M. Z. Jaramillo, A. Gelbukh, and F. A. Isaza, "Pre-conceptual schema: A conceptual-graph-like knowledge representation for requirements elicitation," in *Mexican International Conference on Artificial Intelligence*, 2006, pp. 27–37.

[91] M. Ferndndez, A. G6mez-P~rez, and N. Juristo, "METHONTOLOGY: From Ontological Art Towards Ontological Engineering," 1997. doi: 10.1109/AXMEDIS.2007.19.

[92] S. Moiseyenko and V. Ermolayev, "Conceptualizing and Formalizing Requirements for Ontology Engineering.," in *PhD@ ICTERI*, 2018, pp. 35–44.

[93] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "Linkedgeodata: A core for a web of spatial open data," *Semant Web*, vol. 3, no. 4, pp. 333–354, 2012.

[94] M. Fabian, K. Gjergji, and W. Gerhard, "YAGO: A core of semantic knowledge unifying wordnet and wikipedia," *16th International World Wide Web Conference*,

..., pp. 697–706, 2007, doi: 10.1145/1242572.1242667.

[95] R. Navigli and S. P. Ponzetto, "BabelNet: Building a very large multilingual semantic network," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 216–225.

[96] G. A. Miller, "WordNet," *Commun ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[97] C. M. Meyer and I. Gurevych, *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012.

[98] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge," *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, no. Singh 2002, pp. 4444–4451, Dec. 2017, [Online]. Available: http://arxiv.org/abs/1612.03975

[99] M. Alam, A. Gangemi, V. Presutti, and D. Reforgiato Recupero, "Semantic role labeling for knowledge graph extraction from text," *Progress in Artificial Intelligence*, vol. 10, no. 3, pp. 309–320, 2021, doi: 10.1007/s13748-021-00241-7.

[100] H. Yu, H. Li, D. Mao, and Q. Cai, "A relationship extraction method for domain knowledge graph construction," *World Wide Web*, vol. 23, no. 2, pp. 735–753, 2020, doi: 10.1007/s11280-019-00765-y.

[101] F. Corcoglioniti, M. Rospocher, and A. P. Aprosio, "A 2-phase frame-based knowledge extraction framework," in *Proceedings of the ACM Symposium on Applied Computing*, Association for Computing Machinery, Apr. 2016, pp. 354–361. doi: 10.1145/2851613.2851845.

[102] S. Issa, O. Adekunle, F. Hamdi, S. S.-S. Cherfi, M. Dumontier, and A. Zaveri, "Knowledge Graph Completeness: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 31322–31339, 2021, doi: 10.1109/ACCESS.2021.3056622.

[103] L. B. B, E. Laurenza, and E. Sallinger, *Reasoning under uncertainity in Knowledge Graphs*, vol. 1. Springer International Publishing, 2020. doi: 10.1007/978-3-030-57977-7.

[104] A. Hur, N. Janjua, and M. Ahmed, "Unifying context with labeled property graph: A pipeline-based system for comprehensive text representation in NLP," *Expert Syst Appl*, vol. 239, p. 122269, Apr. 2024, doi: 10.1016/J.ESWA.2023.122269.