Narrative Extraction from Semantic Graphs

Daniil Lystopadskyi
Faculty of Sciences, University of Porto, Portugal

André Santos
CRACS & INESC TEC, Porto, Portugal
Faculty of Sciences, University of Porto, Portugal

José Paulo Leal
CRACS & INESC TEC, Porto, Portugal
Faculty of Sciences, University of Porto, Portugal

Abstract

This paper proposes an interactive approach for narrative extraction from semantic graphs. The proposed approach extracts events from RDF triples, maps them to their corresponding attributes, and assembles them into a chronological sequence to form narrative graphs. The approach is evaluated on the Wikidata graph and achieves promising results in terms of narrative quality and coherence. The paper also discusses several avenues for future work, including the integration of machine learning, graph embedding methods and the exploration of advanced techniques for attention-based narrative labeling and semantic role labeling. Overall, the proposed method offers a promising approach to narrative extraction from semantic graphs and has the potential to be useful in various applications, including chatbots, conversational agents, and content creation tools.

2012 ACM Subject Classification Information systems → Environment-specific retrieval; Information systems → Information extraction

Keywords and phrases Narratives, Narrative Extraction, Information Retrieval, Knowledge Graphs, Semantic Graphs, Resource Description Framework, Web Ontology

Digital Object Identifier 10.4230/OASIcs.SLATE.2023.9

Category Short Paper

Funding This work is financed by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.
André Santos: Ph. D. Grant SFRH/BD/129225/2017 from Fundação para a Ciência e Tecnologia (FCT), Portugal.

1 Introduction

Narratives are ubiquitous in human communication and understanding them is essential for various applications[8], including information retrieval, summarization, storytelling, question answering and content generation. However, defining narratives and formalizing them for computational processing is a challenging task.

Narrative extraction is the task of automatically identifying, analyzing, and representing narratives from textual or multimedia data[10]. Semantic graphs, which represent entities and their relationships in a structured and rich manner, offer a promising framework for narrative extraction that can capture both local and global coherence in a text[2].

The study of narratives has attracted significant attention in different research fields, including computer science[8]. Most of the work in this field involves extracting narratives from plain text using Natural Language Processing (NLP) techniques[10] or predefined event-centric graphs[12]. However, extracting narratives from semantic graphs is a relatively unexplored area, which is the focus of our work.

Semantic graphs provide a structured source of information that is easier for computers to understand. By defining narratives as graphs, we provide a universal representation that simplifies accessing and transferring information between systems[9]. Despite the potential benefits of semantic graphs for narrative extraction, existing methods face several challenges, such as the complexity of the graph structure, the sparsity of data and the lack of sufficient domain-specific knowledge[7].

The main contributions of this paper are twofold. First, we propose an interactive approach for narrative graph extraction from semantic graphs, which combines string matching and rule-based methods to capture the semantic and structural information of a narrative. Second, we demonstrate the effectiveness and adaptability of our approach across different domains and languages. In the following sections, we give background for common concepts in this field, provide an overview of related work, describe our approach in detail, present experimental results, and discuss future work and conclusions.

2 Background

This section provides an overview of general concepts that are relevant to fully comprehend the context of this field of study.

Semantic graphs, sometimes also referred to as knowledge graphs, store domain/context specific information about concepts and relations between them. The information stored in semantic graphs can also be viewed as a set of triples, where each triple corresponds to (Subject, Predicate, Object). In terms of graphs, these triples are represented as (Node, Edge, Node). Those triples are also called RDF triples.

Resource Description Framework (RDF) is a standard description format used for describing and exchanging metadata and other resources on the web. It forms an important part of the semantic web stack and plays a key role in enabling the interoperability and integration of data across different systems and domains.

The information within semantic graphs is structured according to a web ontology. An ontology, in web semantics context, is a standardized way of defining the hierarchy of concepts and relationships that exist within a particular domain, using a set of classes, properties and constraints.

SPARQL is a query language used to retrieve and manipulate data stored in RDF format. SPARQL allows users to query RDF data by specifying patterns and conditions that the data must match. These patterns can include information about the structure of the data, the types of entities and relationships involved, and constraints on the values of properties.

3 Related Work

Narrative extraction is a challenging task that has attracted significant research interest in recent years. Existing methods can be broadly classified into three categories: NLP methods, rule-based methods and hybrid methods.

NLP-based approaches that have been prominent in the field of narrative extraction can be summarized into five stages: Pre-Processing and Parsing; Identification and Extraction of Narrative Components; Linking Components; Representation of Narratives and Evaluation[10]. An example of this approach is narrative extraction from administrative records[4].

Rule-based methods for narrative extraction rely on manually crafted rules or heuristics to identify the narrative structure or content. These methods often require domain-specific knowledge and are limited in their adaptability to new domains or languages. Those types of approaches are sparse, the most recent one being [2], which uses pre-defined mappings between event components and properties in Wikidata to extract events.

Hybrid methods for narrative extraction combine NLP-based and rule-based approaches to leverage their respective strengths. For example, [5] proposed a method that uses both NLP-based and rule-based techniques to extract narrative events and their temporal relations from Wikipedia biographies.

4 Approach

In relation to the existing work in this specific field, our contribution consists of expanding the concept of events by removing constraints such as narrative genres and event classes, allowing for more flexible narratives, as well as providing a web interface for deeper levels of interactability and, consequently, higher generalizability. This approach can be described by the pipeline in Figure 1. The next subsections describe, in detail, each step of the pipeline.

4.1 Narrative Definition and Ontology Specification

In this paper, narratives will be defined as ordered sequences of events. Although there is a significant body of research focused on conceptualizing and defining narratives, this simplified definition is sufficient for the purposes of this study. For the purposes of this study, an event is considered an occurrence linked to a specific point in time. The 5W1H method, suggested by [2], which involves answering questions related to "What", "Where", "Who", "When", "Why" and "How", will be used to describe events, with the "How" and "Why" questions excluded due to their complexity, leaving only 4W's: "What", "Where", "Who" and "When".

To ensure consistency in the description of narratives, a web ontology was created. The ontology has two primary classes: one to define events and the other to define narratives. The event class includes four properties, one for each of the 4W categories mentioned above, as well as a property to establish relationships between events and other entities that do not fit into any of the attributes, although still fundamental for narrative flow. The narrative class, as of time of writing this paper, has one property that links the narrative to its events. However, the properties for the narrative class are still under development and will be determined in due course. An example of an event in this ontology depicting the "Word War I Ottoman Southern Front" can be seen in Figure 2.

4.2 Narrative and Graph Configuration

One of the primary challenges that we face in applying the rule-based approaches universally is the inconsistency in graph structures across different domains[5]. This inconsistency affects how the graph is traversed and how information is retrieved, making it difficult to find a common ground between different graphs. Other inconsistencies, such as variations in schema and uri prefixes for entities and predicates, add further complexity to the problem. Furthermore, it is challenging to infer the meaning of ontology classes and predicates without a clear understanding of the ontology itself.

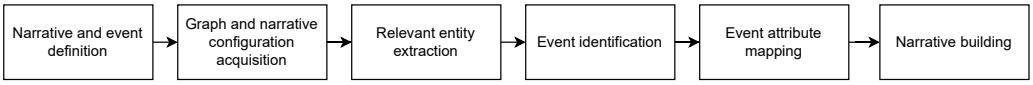

Figure 1 Approach Pipeline.

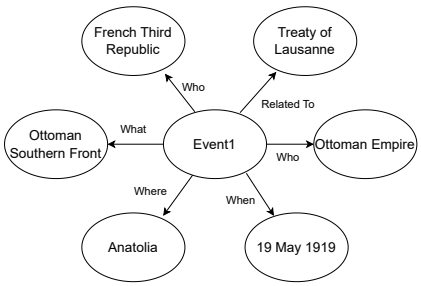

Figure 2 Ontology Event Representation.

Designing an algorithm that can work with any semantic graph, without prior knowledge of the specific graph, is a formidable task [6]. Our approach achieves adaptability by prompting users to provide specific configuration parameters for their graphs, such as the property that could replace "rdf:type".

Graphs are not the only aspect that requires configuration. Narratives, being a subjective topic, require user input on how they should look and behave. Primarily, there needs to be a parameter that defines how "deep" a narrative should be, seeing as narratives can span out infinitely due to event inter-connectivity. This is also commonly known as the "butterfly effect", which the parameter "depth" attempts to control.

Some graphs are too large for the algorithm to handle, such as Wikidata. Thus, it is necessary to specify a parameter that controls how much information is retrieved from the graph in order to avoid long execution times or even timeouts. We will refer to this parameter as the "size".

4.3 Relevant Entity Extraction

The first step of this algorithm involves identifying the entity that represents the input topic. The topic is given by the user as a string. The algorithm queries the graph for this string and returns all entities that are labeled as such, of which there can be more than one. The user then needs to specify which entity they want the narrative for. This entity is then marked as the main entity, and event extraction can begin using it as the central point.

Once the main entity is acquired, the identification and extraction of all entities relevant to the narrative can commence. Relevant entities are the ones linked to the main entity through some property and constitute the narrative space.

The entities are collected through graph traversing using Breadth-first search. The depth of the search is set by the equivalently called parameter which defines how far should we search from the main entity. The higher the depth, the more specific or irrelevant events get in relation to the main entity.

4.4 Event Identification

To identify events within a semantic graph, our approach looks for timestamps associated with graph resources. Two types of events can be derived by manually analyzing semantic graphs and their contents: entity events and property events. Entity events are represented by entities that depict real-world events, such as "World War 2" or "1952 Swiss Mount Everest expedition", while property events are represented by properties whose range is a literal with datatype "date" and/or "time", such as "date of birth". While it is true that there could be entities that define time instances, for the sake of simplicity, those are ignored.

Property events are extracted from all relevant entities that are not classified as events. The final set of events corresponds to extracted property events, as well as all relevant entities classified as events. This set is then used for attribute mapping for each event.

4.5 Attribute Mapping

The process of event attribute mapping requires a deep understanding of entity and predicate types. The assignment involves mapping all relevant data to each W of the 4W's (Who, What, Where, When) for every identified event, which is a challenging task. For this purpose, we ask user to, manually, assign one of the 4 classes ("Person/Group", "Location", "Event" or "Other") to all extracted entities and properties for further processing. As a result, this assignment can substantially influence the resulting narrative structure. The main advantage of this approach is high adaptability to different graphs, since it does not rely on hardcoded assignments.

For entity events, the attributes are mapped according to the following rules:

- **Who**: All entities linked to the event entity that are classified as "Person/Group".
- **What**: Label of the event entity, concatenated with the property label that contains the timestamp of the event, e.g. "start time".
- **Where**: All entities linked to the event entity that are classified as "Location".
- **When**: Value assigned to the property that contains the timestamp of the event.
- **Related to**: All entities linked to the event entity that are classified as "Event".

On the other hand, for event properties, the key difference is that not all objects of type "Person" and "Location" that belong to the same entity are relevant to the event. Also, we assume that all the necessary data is contained within the RDF triples of the same entity, meaning that all incoming links to this entity are ignored. The first step is to find all triples related to the triple containing the event property. This is done by clustering RDF triples of the same entity based on Ratcliff-Obershelp similarity between each triple's property labels and only match properties that show similarity scores greater or equal than a certain threshold, which, at this point of the development, has to be configured manually, through trial and error.

Once we have the set of all related triples, the attributes are mapped according to the following rules:

- **Who**: All objects from clustered triples that are classified as "Person/Group".
- **What**: Label of the event property.
- **Where**: All objects from clustered triples that are classified as "Location".
- **When**: Value assigned to the event property.
- **Related to**: All objects from clustered triples that are classified as "Event".

The summary of this process can be visualized in Figure 3 for entity events, and in Figure 4 for property events.

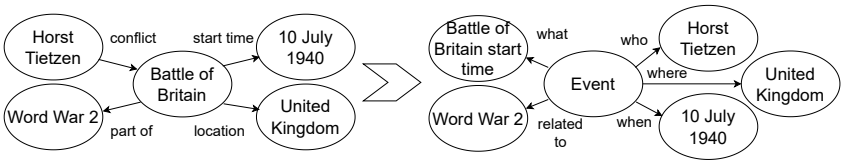

Figure 3 Entity Event Example.

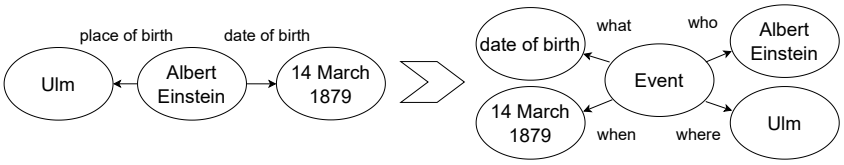

Figure 4 Property Event Example.

4.6 Narrative Building

After gathering all of the relevant data, the next step in the event extraction process involves assembling the narrative, which is, perhaps, both the simplest and the hardest step of the process. For the purpose of this paper and for the sake of clarity, we will be using a simple approach.

The narrative is essentially an ordered sequence of events that occurred within a certain period of time. As the extracted events all contain timestamps, building the narrative becomes a straightforward task of arranging the events in chronological order. This chronological order will serve as the backbone of the narrative, providing a clear timeline for the sequence of events. Since the goal is to build semantic graphs out of extracted narratives, each extracted event becomes a node in the graph. These events are linked to their attribute nodes, which can be either literals or other entities. A narrative is a node itself, linked to every event node that constitutes it. The final narrative graph can be embedded into original semantic graph.

A more complex approach would involve incorporating other important concepts such as the perspective of the narrator, interconnected narratives and character roles. However, these more complex narrative-building options will be explored only after the primary method has been fully developed and refined.

5 Preliminary Results

To evaluate the quality of our approach, we conducted experiments on the Wikidata graph, a benchmark knowledge graph for information retrieval. The results were extracted in form of a table, with each row corresponding to an event and each column representing an event attribute. All results were obtained for parameters “depth” equals one and “size” equals thirty.

For the first example, “Marie Curie” was used as the topic for the algorithm. Once the correct type was selected for the topic, in this case a person, and all entity and property classes were assigned, the program returned results seen in Figure 5. In total, we managed to obtain forty-eight events, which had to be truncated to the first ten in order to fit in this paper. In those ten example events, we can deduct the general flow of the narrative, starting from birth and progressing towards marriage and winning an award. As can be seen in the table, some attributes are optional, since those were not found. Overall, the “What” attribute requires more inference from the part of the reader, since it is not always obvious what it is referring to. This is one of the possible quality-of-life adjustments that could be introduced.

Another topic used was “Battle of Greece”, a historical event, for which seven events were extracted and can be found in Figure 6. Once again, events that were too large were excluded due to space limitations. Just as in the first example, some items appear duplicated due to errors in the implementation of the algorithm, which can be resolved easily. In this short example, we can see the main entity, Battle of Greece, which is then linked to other

Event ID	When	Who	What	Where	Related To
0	1867-01-01T00:00:00Z	Marie Curie	residence start time	Warsaw	
1	1867-11-07T00:00:00Z	Marie Curie	date of birth	Warsaw,Warsaw	
2	1891-01-01T00:00:00Z	Marie Curie	residence end time	Warsaw	
3	1891-01-01T00:00:00Z	Marie Curie	residence start time	Paris	
4	1891-01-01T00:00:00Z	Marie Curie	educated at start time	University of Paris	
5	1893-01-01T00:00:00Z	Marie Curie	educated at end time	University of Paris	
6	1894-01-01T00:00:00Z	Marie Curie	educated at end time	University of Paris	
7	1895-01-01T00:00:00Z	Marie Curie	country of citizenship start time	France	
8	1895-07-26T00:00:00Z	Pierre Curie,Marie Curie	spouse start time		
9	1898-01-01T00:00:00Z	Marie Curie	prix Gegner winner point in time		Marie Curie

Figure 5 Section of Extracted Narrative for Marie Curie.

3	1941-04-01T00:00:00Z	United Kingdom,Wehrmacht,Australian Army,Hellenic Armed Forces	Battle of Greece	Greece,Kingdom of Greece	part of World War II
4	1941-04-09T00:00:00Z	United Kingdom	Battle of Metaxas Line		part of Battle of Greece,part of Battle of Greece
5	1941-04-13T00:00:00Z	United Kingdom	Battle of Ptolemaida	Ptolemaida	part of Battle of Greece,part of Battle of Greece
6	1941-04-15T00:00:00Z		Battle of Lake Kastoria	Kastoria	part of Battle of Greece,part of Battle of Greece

Figure 6 Section of Extracted Narrative for Battle of Greece.

events, such as World War 2, Battle of Metaxas Line, Battle of Ptolemaida, and so on... Once again, some attributes are omitted. The "Who" attributes in this case refers to participants of respective battles.

Overall, despite there being a lot of room for improvement, our preliminary results demonstrate the potential of our proposed approach for narrative extraction from semantic graphs. Seeing as this is a work in progress, a better validation method is required once the algorithm is finished to further evaluate the quality of produced narratives.

6 Conclusions

In this paper, we presented an interactive approach for extracting narratives from semantic graphs. Our approach utilizes matching methods, rule-based techniques and string comparison algorithms to analyse semantic graphs and extract narrative structures by converting them into their own semantic graphs. Our approach was evaluated on the Wikidata graph and showed promising results for a small sample size.

With that being said, there are several avenues for future work to further improve and extend our method. One direction is the integration of graph-based algorithms, such as graph embedding and graph neural networks[11], to enhance the model's understanding of the narrative. Another direction is the development of machine learning models for property and entity classification, considering the vast amount of data provided by knowledge graphs[3]. Additionally, exploring more advanced techniques for attention-based narrative labeling and semantic role labeling can further improve the quality of the results[1]. Finally, it is important to evaluate our approach on other benchmark graphs and real-world datasets to assess its adaptability and practical usefulness.

In conclusion, our proposed approach for narrative extraction from semantic graphs represents a step forward in the field of event extraction and narrative understanding. We believe that our approach has the potential to be applied to a wide range of real-world

applications, including chatbots, conversational agents, and automated story generation. We hope that our work will inspire future research in this exciting field and lead to further advancements in narrative extraction and semantic graph analysis.

References

1 Nandini Anantharama, Simon D. Angus, and Lachlan O'Neill. Canarex: Contextually aware narrative extraction for semantically rich text-as-data applications. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3551–3564. Association for Computational Linguistics, 2022. URL: https://aclanthology.org/2022.findings-emnlp.260.

2 Inès Blin. Building narrative structures from knowledge graphs. In Paul Groth, Anisa Rula, Jodi Schneider, Ilaria Tiddi, Elena Simperl, Panos Alexopoulos, Rinke Hoekstra, Mehwish Alam, Anastasia Dimou, and Minna Tamper, editors, *The Semantic Web: ESWC 2022 Satellite Events - Hersonissos, Crete, Greece, May 29 - June 2, 2022, Proceedings*, volume 13384 of *Lecture Notes in Computer Science*, pages 234–251. Springer, 2022. doi:10.1007/978-3-031-11609-4_38.

3 Victor de Boer. Knowledge graphs for impactful data science (keynote). In Umutcan Simsek, David Chaves-Fraga, Tassilo Pellegrini, and Sahar Vahdat, editors, *Proceedings of Poster and Demo Track and Workshop Track of the 18th International Conference on Semantic Systems co-located with 18th International Conference on Semantic Systems (SEMANTiCS 2022), Vienna, Austria, September 13th to 15th, 2022*, volume 3235 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3235/keynote1.pdf.

4 Karine Megerdoomian, Karl Branting, Charles Horowitz, Amy Marsh, Stacy Petersen, and Eric Scott. Automated narrative extraction from administrative records. In Luther Karl Branting, editor, *Proceedings of the Workshop on Artificial Intelligence and the Administrative State co-located with 17th International Conference on AI and Law (ICAIL 2019), Montreal, QC, Canada, June 17, 2019*, volume 2471 of *CEUR Workshop Proceedings*, pages 38–48. CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2471/paper7.pdf.

5 Daniele Metilli. *Enhancing the Computational Representation of Narrative and Its Extraction from Text*. PhD thesis, University of Pisa, Italy, 2021. URL: https://etd.adm.unipi.it/theses/available/etd-10222021-095519/.

6 Thiloshon Nagarajah, Filip Ilievski, and Jay Pujara. Understanding narratives through dimensions of analogy. *CoRR*, abs/2206.07167, 2022. doi:10.48550/arXiv.2206.07167.

7 Emetis Niazmand, Gezim Sejdiu, Damien Graux, and Maria-Esther Vidal. Efficient semantic summary graphs for querying large knowledge graphs. *Int. J. Inf. Manag. Data Insights*, 2(1):100082, 2022. doi:10.1016/j.jjimei.2022.100082.

8 Priyanka Ranade, Sanorita Dey, Anupam Joshi, and Tim Finin. Computational understanding of narratives: A survey. *IEEE Access*, 10:101575–101594, 2022. doi:10.1109/ACCESS.2022.3205314.

9 Vetle Ryen, Ahmet Soylu, and Dumitru Roman. Building semantic knowledge graphs from (semi-)structured data: A review. *Future Internet*, 14(5):129, 2022. doi:10.3390/fi14050129.

10 Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, January 2023. doi:10.1007/s10462-022-10338-7.

11 Daniil Sorokin. *Knowledge Graphs and Graph Neural Networks for Semantic Parsing*. PhD thesis, Technical University of Darmstadt, Germany, 2021. URL: http://tuprints.ulb.tu-darmstadt.de/19187/.

12 Zhihua Yan and Xijin Tang. Narrative graph: Telling evolving stories based on event-centric temporal knowledge graph. *Journal of Systems Science and Systems Engineering*, 32(2):206–221, April 2023. doi:10.1007/s11518-023-5561-0.