From Nodes to Narratives: A Knowledge Graph-based Storytelling Approach

Author 1[1,*], Author 2[2,*] and Author 3[2]

[1]Institution 1, City 1, Country 1
[2]Institution 2, City 2, Country 2

Abstract

Narratives wield a profound influence, shaping perceptions, beliefs, and decision-making processes. Although contemporary pre-trained language models have showcased impressive capabilities in text generation and question-answering tasks, they grapple with inherent limitations in knowledge coverage and exhibit vulnerability to societal biases. This work endeavors to forge a methodology that applies knowledge graphs in narrative construction. Rather than solely focusing on fundamental aspects such as the 4W (who, what, when, where) and general relationships, our approach comprises finely detailed semantic relations, delineating precise type of causality such as an event preventing, intending-to-cause, causing, or enabling another event. Applying state-of-art methods to predict such rich information, we demonstrate that it is possible to obtain automatically generated narratives of better grammatical and semantic accuracy.

Keywords
Narratives, Knowledge Graphs, Information Extraction, Event-centric Knowledge Graphs

1. Introduction

Narratives stand at the heart of our societal fabric, serving our understanding and facilitating the exchange and preservation of knowledge, and cultural heritage. These narratives filter through our everyday lives, appearing in diverse forms such as commercials, political campaigns, news broadcasts, literary works, television shows, and more, each with its unique purpose and significance. What makes narratives truly captivating is their transformative power: they possess the ability to shape our perceptions, instill beliefs, and steer our choices and actions [1]. Consequently, the quest to innovate in the realm of complex narrative generation holds the potential to usher in a new era of AI systems that are intricately attuned to human sensibilities. Building upon the profound role of narratives in our society, it becomes evident that our means of narrative generation and comprehension are intertwined with the capabilities of modern AI. Pre-trained large language models, exemplified by models such as BERT [2], GPT-3 [3], and the more recent ChatGPT (GPT3.5)[1], have showcased remarkable progress in text generation, and conversational tasks. Yet, these models, shaped by training on extensive datasets drawn

---

✉ name.surname@institution1.org (A. 1); name.surname@institution2.org (A. 2); name.surname@institution2.org (A. 3)

[1]https://openai.com/blog/chatgpt/

from undisclosed and diverse sources, bear intrinsic limitations, including knowledge gaps, inaccuracies, and societal biases [3, 4]. Their challenges in maintaining semantic coherence and capturing long-term dependencies within text generation further underscore the need for innovation in narrative crafting [5, 6].

Knowledge Graphs (KGs) are proven to be suitable structures for human knowledge, designed for machine-readability and adaptability, while several experiments of text generation from KGs are present in the literature [7]. Several KGs are available as data sources for the automatic generation of narratives. For example, *EventKG* [8] is a knowledge graph that consolidates and links events extracted from diverse sources, including Wikidata and YAGO [9]. This knowledge graph comprises more than 1.3 million events, each associated with its respective spatial and temporal coordinates. However, EventKG primarily focuses on representing events attributed and relationships between sub-events and super-events. While the value of such a knowledge graph is undeniable, its limitation to specific event properties, notably the sub(super)events or the 4W, results in succinct and somewhat incomplete narratives. To address this limitation, we propose enriching this information by incorporating detailed event relations [10].

The FARO dataset [10, 11] encompasses a broader spectrum of semantically precise relationships. This includes event-related connections such as *Prevention*, *Enabling*, *Causality*, and *Intention*. In this work, we propose to enhance the WebNLG dataset [12], which was originally limited to the 4W relations, by incorporating the FARO dataset. This expansion aims to produce more linguistically sophisticated generated text with richer semantic content.

The remainder of this paper is structured as follows: we first review the prior research pertaining to narratives and the extraction of relevant information from KGs (Section 2). We present datasets in Section 3, and we detail our approach for KG summarization, which encompasses an initial information selection step before text generation in Section 4. We then present both qualitative and quantitative results in Section 5. We conclude and outline some future work in Section 6.

2. Related Work

A *narrative graph* [13] incorporates two main components: the individual representation of events, including the "four W" aspects (*who, what, when, where*) and the interconnection of these events through temporal and causal relationships. The *Simple Event Model* (SEM) [14] provides a foundation for modeling events, but is still insufficient to link disparate events or classes of the same type. To address this limitation, Blin [13] suggests enriching the event relation types: temporal or causal links from Allen [15] and dbo:alongside links between classes of the same type. Furthermore, the FARO ontology[2] [10] covers most of the existing event relations in the literature, from temporal relation to causal and more fine-grained ones such as *prevention*.

KG summarization entails an initial step of information retrieval and selection. To acquire the essential nodes for event description, an effective approach involves ranking techniques that assign significance to nodes based on the relationships they possess. Various methods can be used such as entity ranking, relationship ranking, and semantic document ranking

---

[2]https://anr-kflow.github.io/faro/

[16]. [17] proposes a system that can identify relevant information needed to build a narrative graph, by using an informed graph search traversal strategy. To determine which information is considered 'relevant' the method uses filters to prune the search space with respect to the Simple Event Model (What, Who, Where, When).

On the other hand, different methods for generating texts from knowledge graphs have been proposed. In [18], triples are extracted to fine-tune a GPT-2 model [19], making the model dependent on the input triples. A similar approach is introduced in [20], involving BART [21] and T5 [22]. This approach obtained state-of-the-art performances on the AGENDA dataset [23] but not on the WebNLG dataset. Both found that Pre-trained Language Models (PLM) work well on unordered representations of the graph. JointGT [24] uses BART and T5, and exploits new pre-training methods to explicitly preserve the input graph's structural information. JointGT outperforms the other mentioned technique on WebNLG, which might indicate that including the topology of the graph lead to better results. A different approach [25] uses a transformer encoding structure to encode both the global information and the local topology information, and feeds a transformer to decode and generate text. However, this did not work as well as the previously mentioned technique [20], which used a PLM model without encoding. This might indicate that PLMs obtain better results than self trained transformer models.

3. Dataset

In this section, we present the datasets that will be used to train our method: WebNLG [26], the FARO dataset [11] (Table 1). For evaluation, we use two evaluation datasets: the FARO test set and the ASRAEL KG [27]. ASRAEL is a knowledge graph that includes various event-related articles and their interconnections, including the 4W relations.

Sentence	Trigger1	Trigger2	Tag	Triplets
The government has implemented a series of laws to prevent the abuse of animals.	laws	abuse	prevent	<triplet>laws <subj> abuse <obj>prevent

Table 1
Sample of the FARO dataset

We enhanced the ASRAEL KG with additional relations (similarly to the ones in FARO) within its articles, resulting in a more intricate and comprehensive knowledge graph. To achieve this objective, we used the REBEL model [28] to extract event mentions and event relations. Furthermore, we leverage an existing event co-reference resolution model [29] to perform the task within the KG. This model creates clusters of mentions, computes similarity scores for each cluster, merges those with the highest score, and repeats this process until the score fell below a defined threshold, which we empirically set to 0.95. This clustering process resulted in a graph primarily composed of single mentions. Syntactic matches within these clusters were notably high, indicating the quality of our co-reference resolution. In total, we successfully clustered 45,031 mentions, with 36,057 being unique. The resulting narrative graph[3] provides a RDF representation of event co-references and relationships, enriched with ontologies such

---

[3]https://anonymous.4open.science/r/KG2Narrative/Data/graphs/final_generated/eag_complete_merged.ttl

as NIF (NLP Interchange Format[4]), SEM and FARO to describe the relations between triples, further enhancing the context and meaning of our knowledge graph. More detailed information can be found in Appendix A.

4. Knowledge graph summarization

Knowledge Graph summarization comprises two tasks: the selection of pertinent information from the knowledge graph, and the text generation based on the extracted data.

4.1. Relevant Information Selection

A SPARQL query has been used for the identification of essential nodes for the narrative. This query, prioritizes the selection 4W nodes, with a higher frequency of incoming edges. Mentions are selected similarly; the larger the cluster of mentions is, the higher the priority. Since we face a limitation on the number of input tokens of the text generation model, up to three mentions are selected.

The quality of the output depends largely on the quality the output of previous steps (relation extraction and co-reference resolution). Future work aims to enhance the accuracy of both these tasks and explore methods for identifying indirectly linked relevant nodes to selected events.

4.2. Text Generation from Knowledge Graphs

As anticipated in Section 2, using a PLM instead of training a language model from scratch can lead to better results. Furthermore, incorporating the graph's topology into the model has been shown to generate better natural text. The JointGT model [24] incorporates both these characteristics, hence, we adopted this method. The authors pre-trained this model on the KGText dataset [30], consisting of 7 million graph-text pairs extracted from the English Wikidump.[5] It includes around 1.8 million entities and 1,210 relations.

The WebNLG dataset does not contain any of the FARO relations. Therefore, we fine-tuned the model on a merged dataset, combining the WebNLG and FARO, as in Table 2.

Table 2
Sizes of the datasets used for training and evaluating the JointGT model

Dataset	Train	Val	Test
WebNLG	12,876	1,619	1,600
FARO	1,800	201	108
Combined	14,676	1,820	1,708

The model undergoes fine-tuning on the WebNLG dataset. We refer to the original model as *base model*, and the model fine-tuned on the combined dataset as *combined model*.[6]

---

[4]https://persistence.uni-leipzig.org/nlp2rdf/

[5]https://dumps.wikimedia.org/

[6]The model was replicated using the same parameters from the original paper, except for the batch size lowered due to memory constraints. The parameters are *Learning rate*: 0.000025, *Batch size*: 4, *Epochs*: 10, *Optimizer*: Adam. *Early stopping*: 10 epochs

5. Results

5.1. Quantitative analysis

Table 3
The performance metrics of the best performing model on their corresponding validation and test set – either WebNLG or the combined set. Both models are evaluated also on the FARO test set

Model	Dataset	BLEU	METEOR	ROUGE	Step	Epoch
Base (WebNLG)	Val	0.6642	0.4727	0.7558	22400	6
	Test	0.6529	0.4681	0.7535	-	-
	FARO test	0.0	0.0565	0.1299	-	-
Combined	Val	0.6368	0.4543	0.7468	36000	9
	Test	0.6101	0.4409	0.7260	-	-
	FARO test	0.0477	0.0877	0.1949	-	-

Table 3, reveals important insights into the model's performance. The ROUGE metric indicates a relatively high level of performance, suggesting that the model generates text that closely aligns with the reference texts in terms of information. In contrast, the BLEU metric exhibits slightly lower performance, implying that the model's output words differ slightly from those in the reference texts. The relatively low METEOR score could be attributed to how the predicted and reference texts are aligned when calculating the score. It is worth noting that the base model's performance on the test set aligns closely with the results reported in the original JointGT paper [24].

The model that was trained on the combined dataset performed slightly worse for all three metrics than the model that was trained on the base WebNLG data. This can be explained by two considerations. First, it is evident in Table 3 that tests on FARO have very low performances. Secondly, the FARO dataset only accounts for a relatively small proportion in the combined dataset (Table 2). To better understand the reasons, a qualitative analysis is proposed in the next section.

5.2. Qualitative analysis.

We select some instances from both WebNLG and FARO datasets, to investigate the performance of the base and the combined model. Looking at these instances from Tables 4 and 5, we see that the generated text from the model trained on the combined dataset is semantically richer. The generated text of the base model on the FARO triples (Table 4, column *Base generated*) is very short, almost exclusively being a transposition of the triples. In addition, most of the examined sentences are semantically not correct. On the same dataset, the combined model generates more articulated and correct sentences, in which the direction of the triples is preserved (column *Combined generated*). It is worth to mention that the generated content, even if it respects the ordering of the triples and it is semantically correct, still has some limitations in terms of changing the content of the original label.

We also get a sight why the quantitative results are slightly worst for the combined model. The WebNLG data (Table 5) contains multiple triples per instance, giving more information about the text, and contains multiple labels. The FARO data (Table 4) contains only one triple per

Triple	Label	Base generated	Combined generated
(offer, cause, reimburse)	(The directors said if Messrs. Drabinsky and Gottlieb mail an offer to shareholders by Nov. 22, it will reimburse them a maximum of C$8.5 million for expenses related to a bid.)	The cause of the offer is to reimburse .	The company has also announced that it will offer a new credit facility to small businesses, in an effort to reimburse them for the cost of capital expenditures.

Table 4: Sample of the FARO test-set and the generated output of the base- and combined model

Triple	Label	Generated
(3Arena, owner, Live Nation Entertainment), (Dublin, is part of, Republic of Ireland), (3Arena, location, Dublin), (Dublin, is part of, Leinster)	(The owner of 3Arena , Dublin , Leinster , Republic of Ireland is Live Nation Entertainment .), (Dublin is part of Leinster and a city in the Republic of Ireland . Dublin is also home to the 3Arena which is currently owned byLive Nation Entertainment.)	3Arena is located in Dublin , Leinster , Republic of Ireland and is owned by Live Nation Entertainment .

Table 5: Sample of the WebNLG test-set and the generated output of the base model.

instance, together with one target sentence (label). Therefore, the model has less information about what to generate, and less chances to match the target label. Looking at the FARO input triples and the target label, it can be seen that the relationship (predicate) is often not explicitly represented by a particular word in the target sentence (implicit relation), making the evaluation with matching words harder. We provide additional insights in Appendix A.

5.3. User Evaluation on ASRAEL

To evaluate the system's performance, seven events from the ASRAEL dataset have been selected based on several criteria: values for the 4w properties, a minimal number of articles, etc. The two largest events from this group are selected for evaluation: *"Operation Breaking Dawn"*, and *"2021 storming of the United States Capitol"*. Among the remaining events that include information about the place and time, five additional events are selected.

The information selection method is used to select time, place, actor, and up to three mentions from the seven selected events. The base and combined models are used to generate text from the selected information. This information per event can be found in Appendix A, together with the generated text. A manual evaluation is realised because of the lack of reference text, precluding the use of automatic metrics.

Per event, three annotators have annotated which of the two models have generated the best text, by using either: "win", "lose", or "tie". In doing so, a distinction is made between *fluency* (whether a sentence is grammatically fluent) and *adequacy* (whether a sentence incorporates the triples correctly). This follows the same approach used in [24]. Majority voting is applied to determine the winning model, or if the models perform equally. Afterwards, the non-parametric "sign test" is performed with a significance level α of 0.05 to decide if one model is better than the other. The results of this annotation are available in Table 6.

The combined model produces better fluent text than the base model in 71.4% of the cases. The non-parametric "signed test" was performed to measure a significant difference in the fluency of the text. With a p-value of 0.11, no significant difference was found. The same was done to measure a significant difference in the adequacy of the text. With a p-value of 0.25, no significant difference was found.

Table 6
Fleiss' Kappa (κ) indicates perfect, and moderate agreement between annotators. The wins, losses, and ties when comparing the combined model against the base model are indicated in percentages. No model was significantly better than another with a significance level of 0.05.

Model	Fluency			κ	Adequacy			κ
	Win %	Lose %	Tie %		Win %	Lose %	Tie %	
Combined vs Base	71.4	14.3	14.3	1.0	28.6	0.0	71.4	0.6

5.4. User Evaluation on an Manually Annotated Event

To demonstrate whether the obtained results are consistent independently from the quality of the information extraction output, we decided to perform a user evaluation on a single article (sample), which has been manually annotated by users, which handcrafted the resulting

subgraph. This subgraph has been processed with both the combined and base model, and then evaluated using either "win", "lose", or "tie", in the same way as described in the previous section. The percentage of wins, losses and ties for the combined model, together with the Fleiss' kappa are reported in Table 7. The combined model has been assigned more wins for producing fluent and adequate text. The non-parametric "signed test" is applied to test if this is significant, again, with a significance level of 0.05. With a p-value of 0.34, no significant difference is found in generating more fluent texts between models. With a p-value of 0.04, a significant difference is found in generating more adequate sentences by the combined model, compared to the base model.

Table 7
Fleiss' Kappa (κ) indicates substantial agreement between annotators. The wins, losses, and ties when comparing the combined model against the base model are indicated in percentages. The combined model was significantly better than the base model in generating adequate sentences.

Model	Fluency			κ	Adequacy			κ
	Win %	Lose %	Tie %		Win %	Lose %	Tie %	
Combined vs Base	33.3	16.7	50.0	0.73	58.3	8.3	33.3	0.61

BLEU, METEOR, and ROUGE metrics have been computed using the sentences from the article as "reference label". These scores are detailed in Table 8. The information from the table illustrates that the base model performs slightly better than the model that was trained on the combined data. A reason for this could be formulated by looking at the generated texts, which can be found in Appendix. The base model will, more often than the combined model, output parts of the triple without taking the relationship between them into account. This will result in a badly formed sentence, but higher metrics, since more triples are incorporated. This is also reflected in the scores in Table 7, where the combined model is more often stated as generating more fluent texts.

Table 8 shows the metric scores for the article. These scores are much lower then those computed on the WebNLG test (Table 3). This outcome could be expected, considering that some of the triples extracted from the article are not, or to a limited extend, present in the original WebNLG data used to pre-train the JointGT model.

6. Conclusion and Future Work

The primary goal of this research is to investigate how to build complex narratives in the form of graphs of events, generating text with good level of complexity and semantic richness, expecting the system to generate answers beyond only *What* (the event), *Who* (the actor), *Where* (the location), and *When* (the time).

We enhanced the WebNLG dataset through the incorporation of the FARO dataset, aimed at refining the semantic depth of event relations. The expanded dataset now encompasses intricate relations including causality, prevention, intention, and enabling. From qualitative analysis, we can state that training on precise event relations produces more complete generated sentences, while no statistically significant difference was observed on fluency. Future work will experiment on more data to draw final conclusions. Our information selection from the graph

Table 8
BLEU, METEOR, and ROUGE scores per model on the generated text from the article

Model	BLEU	METEOR	ROUGE
Combined	0.1681	0.2081	0.3622
Base	0.1874	0.2273	0.3738

focuses solely on the main event, disregarding pertinent details from interconnected events. Additionally, the data used for fine-tuning differs from the original dataset in terms of triple counts and instances, potentially impacting model evaluation. Future research could explore selectively extracting sub-events and relations at the document level to enhance clustering. Moreover, augmenting the dataset through NLP techniques could significantly improve its quality and comprehensiveness.

References

[1] M. C. Green, T. C. Brock, G. F. Kaufman, Understanding media enjoyment: The role of transportation into narrative worlds, Communication theory 14 (2004) 311–327.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, ACL, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[4] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, in: 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2021, pp. 1286–1305. doi:10.18653/v1/2021.emnlp-main.98.

[5] J. Li, T. Tang, W. X. Zhao, J.-R. Wen, Pretrained Language Model for Text Generation: A Survey, in: Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), Survey Track, IJCAI Organization, 2021, pp. 4492–4499. doi:10.24963/ijcai.2021/612.

[6] N. Malkin, Z. Wang, N. Jojic, Coherence boosting: When your pretrained language model is not paying enough attention, in: 60th Annual Meeting of the Association for Computational Linguistics, 2022.

[7] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition, and Applications, IEEE Transactions on Neural Networks and Learning Systems 33 (2022) 494–514. doi:10.1109/TNNLS.2021.3070843.

[8] S. Gottschalk, E. Demidova, EventKG – the hub of event knowledge on the web – and biographical timeline generation, Semantic Web 10 (2019) 1039–1070. 6.

[9] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum, YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, Artificial Intelligence 194 (2013) 28–61.

[10] Y. Rebboud, P. Lisena, R. Troncy, Beyond Causality: Representing Event Relations in Knowledge Graphs, in: Knowledge Engineering and Knowledge Management, Springer, 2022, pp. 121–135. doi:10.1007/978-3-031-17105-5_9.

[11] Y. Rebboud, P. Lisena, R. Troncy, Prompt-based Data Augmentation for Semantically-Precise Event Relation Classification, in: Semantic Methods for Events and Stories workshop (SEMMES), Hersonissos, Greece, 2023.

[12] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, Creating Training Corpora for NLG Micro-Planners, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, Vancouver, Canada, 2017, pp. 179–188. doi:10.18653/v1/P17-1017.

[13] I. Blin, Building Narrative Structures from Knowledge Graphs, in: The Semantic Web: ESWC 2022 Satellite Events, Springer, Germany, 2022, pp. 234–251.

[14] W. R. van Hage, V. Malaisé, R. Segers, L. Hollink, G. Schreiber, Design and use of the Simple Event Model (SEM), Journal of Web Semantics 9 (2011) 128–136.

[15] J. F. Allen, Maintaining Knowledge about Temporal Intervals, Commun. ACM 26 (1983) 832–843. doi:10.1145/182.358434.

[16] V. Jindal, S. Bawa, S. Batra, A review of ranking approaches for semantic search on Web, Information Processing & Management 50 (2014) 416–425.

[17] I. Blin, I. Tiddi, R. van Trijp, A. ten Teije, Identifying graph traversal strategies to build narrative graphs, Under review (2023).

[18] X. Yang, I. Tiddi, Creative Storytelling with Language Models and Knowledge Graphs, in: International Conference on Information and Knowledge Management, 2020.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, OpenAI blog 1.8 (2019).

[20] L. F. R. Ribeiro, M. Schmitt, H. Schütze, I. Gurevych, Investigating Pretrained Language Models for Graph-to-Text Generation, in: 3rd Workshop on Natural Language Processing for Conversational AI, ACL, Online, 2021, pp. 211–227.

[21] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, in: Annual Meeting of the Association for Computational Linguistics, 2019.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, J. Mach. Learn. Res. 21 (2020).

[23] R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, H. Hajishirzi, Text Generation from Knowledge Graphs with Graph Transformers, in: NAACL, 2019.

[24] P. Ke, H. Ji, Y. Ran, X. Cui, L. Wang, L. Song, X. Zhu, M. Huang, JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, ACL, 2021, pp. 2526–2538.

[25] L. Liu, M. He, G. Xu, M. Tan, Q. Wu, How to Train Your Agent to Read and Write, in: AAAI Conference on Artificial Intelligence, 2021, pp. 13397–13405.

[26] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, The WebNLG Challenge:

Generating Text from RDF Data, in: 10th International Conference on Natural Language Generation, ACL, Santiago de Compostela, Spain, 2017, pp. 124–133.

[27] C. Rudnik, T. Ehrhart, O. Ferret, D. Teyssou, R. Troncy, X. Tannier, Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata, in: 2019 World Wide Web Conference, WWW, ACL, New York, NY, USA, 2019, p. 1232–1239.

[28] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation Extraction By End-to-end Language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, ACL, 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204.

[29] S. Barhom, V. Shwartz, A. Eirew, M. Bugert, N. Reimers, I. Dagan, Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution, in: 57th Annual Meeting of the Association for Computational Linguistics, ACL, Florence, Italy, 2019, pp. 4179–4189.

[30] W. Chen, Y. Su, X. Yan, W. Y. Wang, KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation, in: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, Online, 2020, pp. 8635–8648.

A. Appendix

A.1. Results REBEL Model

| Relation	Precision	Recall	F1	Support	Micro F1	Macro F1
Cause	80.00	96.00	87.27	25		
Enable	94.55	94.55	94.55	55	94.03	93.13
Prevent	98.21	91.67	94.83	60		
Intend	96.67	95.08	95.87	61		

Table 9
Detailed result for the validation dataset using the REBEL model

Relation	Precision	Recall	F1	Support	Micro F1	Macro F1
Cause	85.19	100.00	92.00	46		
Enable	84.62	61.11	70.97	18	85.71	82.12
Prevent	84.62	68.75	75.86	16		
Intend	92.86	86.67	89.66	15		

Table 10
Detailed result for the test dataset using the REBEL model
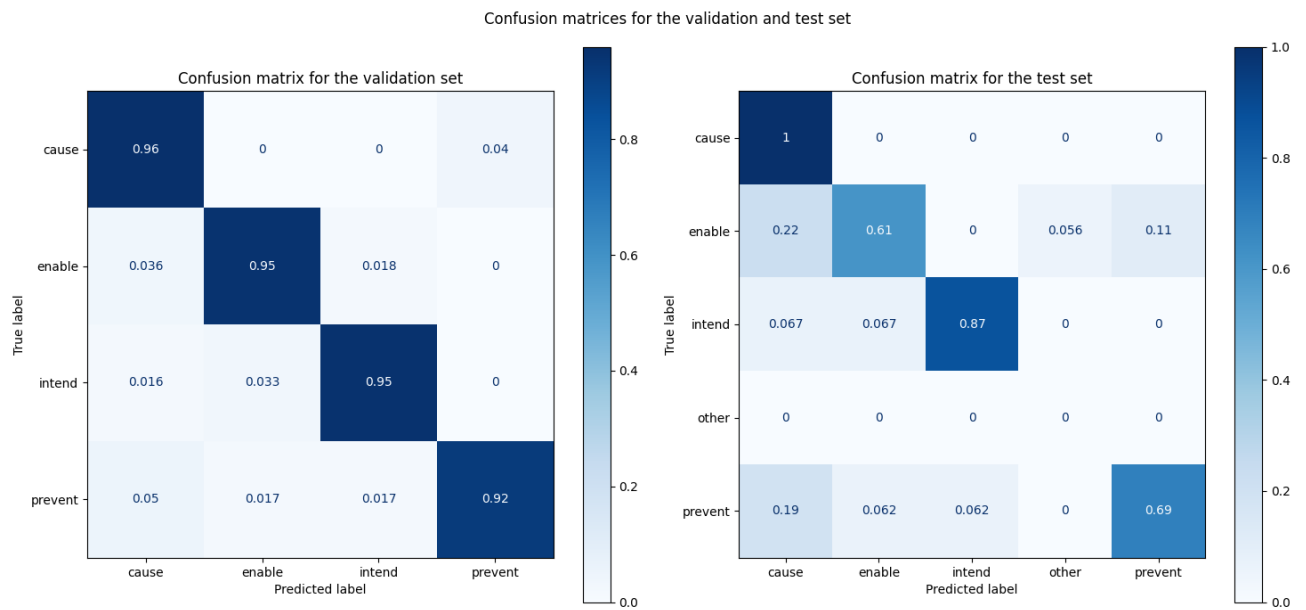
Figure 1: Confusion matrix of relationships in the validation- and testset

Relation	Frequency
has cause	1584
cause of destruction	55
has immediate cause	14
immediate cause of	9
end cause	3
may prevent	2

Table 11
Relationships in the REBEL train dataset

A.2. Event coreference resolution
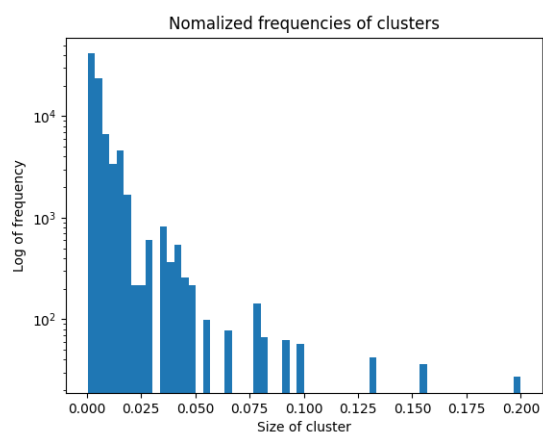
Figure 2: Normalized distribution of the cluster sizes

A.3. JointGT: Qualitative analysis

Triples	Label	Generated
(109 Felicitas, apoapsis, 523329000.0 (kilometres), (109 Felicitas, temperature, 170.0 (kelvins))	(109 Felicitas, with a temperature of 170 kelvins, has an apoapsis of 523329000.0 (kilometres).), (The temperature of the asteroid called 109 Felicitas is 170.0 kelvins and it has an apoapsis of 523329000.0 km.), (The temperature of the asteroid called 109 Felicitas is 170.0 kelvins and it has an apoapsis of 523329000 kilometres.)	109 Felicitas has an apoapsis of 523329000.0 kilometres and a temperature of 170.0 kelvins.
(3Arena, owner, Live Nation Entertainment), (Dublin, is part of, Republic of Ireland), (3Arena, location, Dublin), (Dublin, is part of, Leinster)	(The owner of 3Arena, Dublin, Leinster, Republic of Ireland is Live Nation Entertainment.), (Dublin is part of Leinster and a city in the Republic of Ireland. Dublin is also home to the 3Arena which is currently owned by Live Nation Entertainment.)	3Arena is located in Dublin, Leinster, Republic of Ireland and is owned by Live Nation Entertainment.
(Barny Cakes, serving size, 30.0 g)	(Barny cakes can be served in 30 gram sizes.), (Serving size for the Barny cakes is 30.0g.), (The serving size of Barny cakes is 30.0g.)	Barny cakes have a serving size of 30.0 g.
(107 Camilla, discoverer, N.R. Pogson), (N.R. Pogson, death place, Chennai), (107 Camilla, periapsis, 479343000.0 (kilometres), (N. R. Pogson, birth place, Nottingham)	(Nottingham born, N.R. Pogson (who died in Chennai), discovered 107 Camilla, which has a periapsis of 479343000.0kilometres.), (N.R. Pogson discovered 107 Camilla which has a periapsis of 479,343,000 kilometres. Pogson was born in Nottingham and died in Chennai.), (107 Camilla was discovered by N.R. Pogson who was born in Nottingham. He died in Chennai. The periapsis of 107 Camilla is 479343000.0 (kilometres).)	N.R. Pogson was born in Nottingham and died in Chennai. He discovered 107 Camilla which has a periapsis of 479343000.0 kilometres.
('Fulton County, Georgia', country, United States)	(Fulton County is in Georgia, in the United States.), (Fulton County, in Georgia is part of the U.S.)	Fulton County, Georgia is in the United States.

Table 12: Sample of the WebNLG testset and the generated output of the base model.

Triples	Labels	Generated
(demand, cause, benefited)	(The company benefited from continued strong demand and higher selling prices for titanium dioxide, a white pigment used in paints, paper and plastics.)	benefited " is the cause of the demand .
(center, enable, innovation)	(We recently moved our headquarters and research and development center to Abu Dhabi to benefit from its rapidly growing innovation hub, and with the new AWS Region, we are excited to see it thrive at an accelerated rate.)	Innovation is the enabler of the center .
(made, cause, eaten)	("Cheerios and Honey Nut Cheerios have eaten away sales normally going to Kellogg's corn-based lines simply because they are made of oats," says Merrill Lynch food analyst William Maguire.)	The cause of eating is eaten .
(offer, cause, reimburse)	(The directors said if Messrs. Drabinsky and Gottlieb mail an offer to shareholders by Nov. 22, it will reimburse them a maximum of C\$8.5 million for expenses related to a bid.)	The cause of the offer is to reimburse .
(proclaimed, prevent, bloodhed)	(DR Congo President Felix Tshisekedi on May 6 proclaimed a 30-day "state of siege" – effectively martial law – in North Kivu and neighbouring Ituri province in a bid to curb bloodshed by the ADF.)	Bloodshed can be prevented .

Table 13: Sample of the FARO testset and the generated output of the base model

Triples	Labels	Generated
(demand, cause, benefited)	(The company benefited from continued strong demand and higher selling prices for titanium dioxide, a white pigment used in paints, paper and plastics.)	The company said it benefited from the strong demand for its products and services from a growing number of customers.
(center, enable, innovation)	(We recently moved our headquarters and research and development center to Abu Dhabi to benefit from its rapidly growing innovation hub, and with the new AWS Region, we are excited to see it thrive at an accelerated rate.)	The United States has created a new innovation center that will help stimulate economic growth and job creation.
(made, cause, eaten)	("Cheerios and Honey Nut Cheerios have eaten away sales normally going to Kellogg's corn-based lines simply because they are made of oats," says Merrill Lynch food analyst William Maguire.)	The company said it made a profit from the sale of the shares because it made enough money to make a profit from the sale.
(offer, cause, reimburse)	(The directors said if Messrs. Drabinsky and Gottlieb mail an offer to shareholders by Nov. 22, it will reimburse them a maximum of C$8.5 million for expenses related to a bid.)	The company has also announced that it will offer a new credit facility to small businesses, in an effort to reimburse them for the cost of capital expenditures.
(proclaimed, prevent, bloodhed)	(DR Congo President Felix Tshisekedi on May 6 proclaimed a 30-day "state of siege" – effectively martial law – in North Kivu and neighbouring Ituri province in a bid to curb bloodshed by the ADF.)	The United Nations has proclaimed a new resolution to prevent the spread of bloodshed.

Table 14: Sample of the FARO testset and the generated output of the combined model

A.4. Generating text from selected events

Operation Breaking Dawn

- begin date: 2022-08-05
- end date: 2021-08-07
- location: Israel
- participant: Israel Defense Forces
- triples: $< bombardment,\ cause,\ firing >$, $< bombardment,\ cause,\ retaliated >$, $< fired,\ cause,\ killed >$

Base model: *"Israel Defense Forces participated in Operation Breaking Dawn which began on August 5th, 2022 and ended on July 7th, 2022. The bombardment was retaliated by firing."*

Combined model: *"Israel Defense Forces participated in Operation Breaking Dawn which began on August 5th, 2022 and ended on July 7th, 2022. The Israelis were involved in Operation Breaking Dawn and were killed in the retaliation of the Israelis."*

2021 storming of the United States Capitol

- date: 2021-01-06
- location: United States of America
- participant: Proud Boys
- triples: $< measures,\ enable,\ removing >$, $< died,\ cause,\ shot >$, $< addressed,\ cause,\ shot >$

Base model: *"Save America March is located in the United States of America and was launched on January 6th, 2021. Its cause was stormed and it was removed."*

Combined model: *"Save America March is located in the United States of America and was launched on January 6th, 2021. The US government has implemented a series of measures to prevent the spread of disasters."*

2021 Fukushima earthquake

- : date: 2021-02-13
- location: Japan
- triples: $< earthquake,\ cause,\ collapsing >$, $< earthquake,\ cause,\ injured >$, $< broken,\ cause,\ damage >$

Base model: *"2021 Fukushima earthquake, which was caused by collapsing, is located in Japan and was broken."*

Combined model: *"The 2021 Fukushima earthquake, which hit Japan on February 13th,*

2021, injured many people and caused extensive damage and collapsing."

2021 Sundance Film Festival

- begin date: 2021-01-28
- end date: 2021-02-03
- location United States of America
- triples: $< whittled, cause, unveiling >, < awards, enable, stunt >, < coronaviruspandemic, cause, delayed >$

Base model: *"2021 Sundance Film Festival, located in the United States of America, began on January 28th, 2021 and ended on March 3rd, 2021-02-03. The event was delayed by the coronavirus pandemic, and awards were whittled."*

Combined model: *"2021 Sundance Film Festival, which is located in the United States, began on January 28th, 2021 and ended on March 3rd, 2021. It was delayed because of the coronavirus pandemic."*

Giza church fire

- date: 2022-08-14
- location: Egypt
- triples: $< mobilized, cause, taken >, < went, cause, deaths >, < shortcircuit, cause, fire >$

Base model: *"The cause of the Giza church fire, which was located in Egypt, was a short circuit. The church was mobilised on the 14th of August, 2022."*

Combined model: *"The Giza church fire, which occurred on 2022-08-14, was located in Egypt. It was a short circuit that killed several people."*

Nationwide COVID-19 memorial

- date: 2021-01-19
- location: United States of America
- triples: $< pandemic, prevent, swearing - in >, < confirmation, cause, run >, < sworn, cause, saluted >$

Base model: *"The nationwide COVID-19 Memorial, located in the United States of America, was sworn in on January 19th, 2021. It was saluted and run to prevent a pandemic."*

Combined model: *"The National COVID-19 Memorial is located in the United States of*

America and was inaugurated on January 19th, 2021. It was saluted and sworn in to prevent a pandemic."

Save America March

- date: 2021-01-06
- location: United States of America
- triples: $< measures,\ enable,\ removing >,\ < emergencies,\ cause,\ died >,\ < stormed,\ cause,\ drew >$

Base model: *"Save America March is located in the United States of America and was launched on January 6th, 2021. Its cause was stormed and it was removed."*

Combined model: *"Save America March is located in the United States of America and was launched on January 6th, 2021. The US government has implemented a series of measures to prevent the spread of disasters."*

A.5. Annotation of article: Russia launches Iranian satellite amid Ukraine war concerns

An Iranian satellite launched by Russia blasted off from Kazakhstan Tuesday and reached orbit amid controversy that Moscow might use it to boost its surveillance of military targets in Ukraine

Triples: ("Russia", "launch", "satellite"), ("launch", "location", "Kazakhstan"), ("launch", "time", "Tuesday"), ("satellite", "enable", "military surveillance")

Base model: *"The launch of a satellite in Kazakhstan , which has a military surveillance capability , took place on Tuesday ."*

Combined model: *"The launch of a satellite from Russia was on Tuesday , in Kazakhstan . The satellite will enable military surveillance of the country ."*

As Russia's international isolation grows following Western sanctions over its invasion of Ukraine, the Kremlin is seeking to pivot towards the Middle East, Asia and Africa and find new clients for its embattled space programme.

Triples: ("Russia", "cause", "invasion of Ukraine"), ("invasion of Ukraine", "cause", "western sanctions"), ("western sanctions", "cause", "Russia's international isolation"), ("Russia", "intend", "find new clients")

Base model: *"Russia 's invasion of Ukraine is caused by western sanctions and Russia 's*

international isolation . Russia 's intention is to find new clients ."

Combined model: *"The invasion of Ukraine is a cause of Russia 's isolation and western sanctions are a reason for Russia's invasion of Ukraine . Russia also seeks to find new clients in the region ."*

Speaking at the Moscow-controlled Baikonur Cosmodrome in the Kazakh steppe, Russian space chief Yury Borisov hailed "an important milestone in Russian-Iranian bilateral cooperation, opening the way to the implementation of new and even larger projects".

Triples: ("Russian-Iranian bilateral cooperation", "enable", "implementation of new and even larger projects")

Base model: *"The Russian-Iranian bilateral cooperation is enabled by the implementation of new and even larger projects ."*

Combined model: *"The Russian-Iranian bilateral cooperation will enable the implementation of new and even larger projects ."*

Iran's Telecommunications Minister Issa Zarepour, who also attended the launch of the Khayyam satellite, called the event "historic" and "a turning point for the start of a new interaction in the field of space between our two countries".

Triples: ("Iran Telecommunication", "minister", "Issa Zarepour"), ("Issa Zarepour", "attended", "launch of the Khayyam satellite"), ("satellite", "enable", "a turning point for the start of a new interaction")

Base model: *"Iran Telecommunications Minister Issa Zarepour attended the launch of the Khayyam satellite , which was a turning point for the start of a new interaction ."*

Combined model: *"Iran Telecommunication Minister Issa Zarepour attended the launch of the Khayyam satellite , which was a turning point for the start of a new interaction ."*

Nasser Kanani, the Iranian foreign ministry spokesman, said on Twitter that "the brilliant path of scientific and technological progress of the Islamic republic of Iran continues despite sanctions and the enemies' maximum pressure".

Triples: ("western sanctions", "not prevent", "scientific and technological progress of Iran"), ("enemies' maximum pressure", "not prevent", "scientific and technological progress of Iran")

Base model: *"The enemies ' maximum pressure is the scientific and technological progress of Iran ."*

Combined model: *"The United Nations has imposed a maximum pressure on Iran in an effort to prevent the country from achieving scientific and technological progress."*

Iran, which has maintained ties with Moscow and refrained from criticism of the Ukraine invasion, has sought to deflect suspicions that Moscow could use Khayyam to spy on Ukraine.

Triples: ("Iran", "maintained ties", "Moscow", 5), ("Iran", "not criticized", "Ukraine invasion", 5), ("Khayyam", "enable", "spy", 5)

Base model: *"Khayyam is a spy in Iran , which has maintained ties with Moscow . The Ukraine invasion is not criticized in Iran ."*

Combined model: *"Khayyam was allowed to spy on Moscow and Iran , which has not been criticized for its invasion of Ukraine ."*

Responding to the launch, Washington said Russia's growing cooperation with Iran should be viewed as a "profound threat".

Triples: ("Washington", "respond", "launch"), ("Russia's growing cooperation with Iran", "cause", "profound threat")

Base model: *"Russia 's growing cooperation with Iran is the cause of Washington 's response , which is a profound threat ."*

Combined model: *"Russia's growing cooperation with Iran has created a profound threat to the region, and the United States has responded with a missile launch."*

"No third country is able to access the information" sent by the satellite due to its "encrypted algorithm", it said.

Triples: ("satellite", "sends", "information"), ("encrypted algorithm", "prevent", "third countries accessing information"), ("satellite", "has", "encrypted algorithm")

Base model: *"The satellite has an encrypted algorithm which prevents third countries accessing information ."*

Combined model: *"The satellite sends information via an encrypted algorithm , which*

prevents third countries from accessing information ."

Iran is currently negotiating with world powers, including Moscow, to salvage a 2015 deal aimed at reining in Tehran's nuclear ambitions.

Triples: ("Iran", "intend", "salvage 2015 deal"), ("2015 deal", "prevent", "Iran's nuclear ambitions")

Base model: *"Iran 's nuclear ambitions are intended to prevent the 2015 deal being salvaged ."*

Combined model: *"The United Nations has imposed a new nuclear deal on Iran in an effort to prevent the country from developing nuclear weapons."*

Western governments worry that satellite launch systems incorporate technologies interchangeable with those used in ballistic missiles capable of delivering a nuclear warhead, something Iran has always denied wanting to build.

Triples: ("satellite", "contains", "ballistic missle technologies", 9), ("ballistic missile technologies", "enable", "delivery of nuclear warhead", 9)

Base model: *"The satellite contains ballistic missile technologies which enable the delivery of nuclear warheads ."*

Combined model: *"The satellite contains ballistic missile technologies which enable the delivery of nuclear warheads ."*

Iran successfully put its first military satellite into orbit in April 2020, drawing a sharp rebuke from the United States.

Triples: ("Iran", "launch", "first military satellite", 10), ("first military satellite", "orbit", "April 2020", 10), ("first military satellite", "cause", "sharp rebuke from the United States", 10)

Base model: *"The first military satellite , launched in April 2020 , was launched in Iran . A sharp rebuke from the United States was the cause of the first military satellite ."*

Combined model: *"Iran launched its first military satellite in April 2020 , causing a sharp rebuke from the United States ."*

Borisov, who last month replaced bombastic nationalist Dmitry Rogozin as

head of the Russian space agency, had acknowledged that the national space industry is in a "difficult situation" amid tensions with the West. Triples: ("Borisov", "replaced", "Dmitry Rogozin", 11), ("Borisov", "acknowledged", "difficult situation", 11), ("tensions with the West", "cause", "difficult situation ", 11)

Base model: *"Borisov , who was replaced by Dmitry Rogozin , is in a difficult situation which has caused tensions with the West ."*

Combined model: *"Borisov , who was replaced by Dmitry Rogozin , acknowledged a difficult situation with the West because of tensions with the West ."*

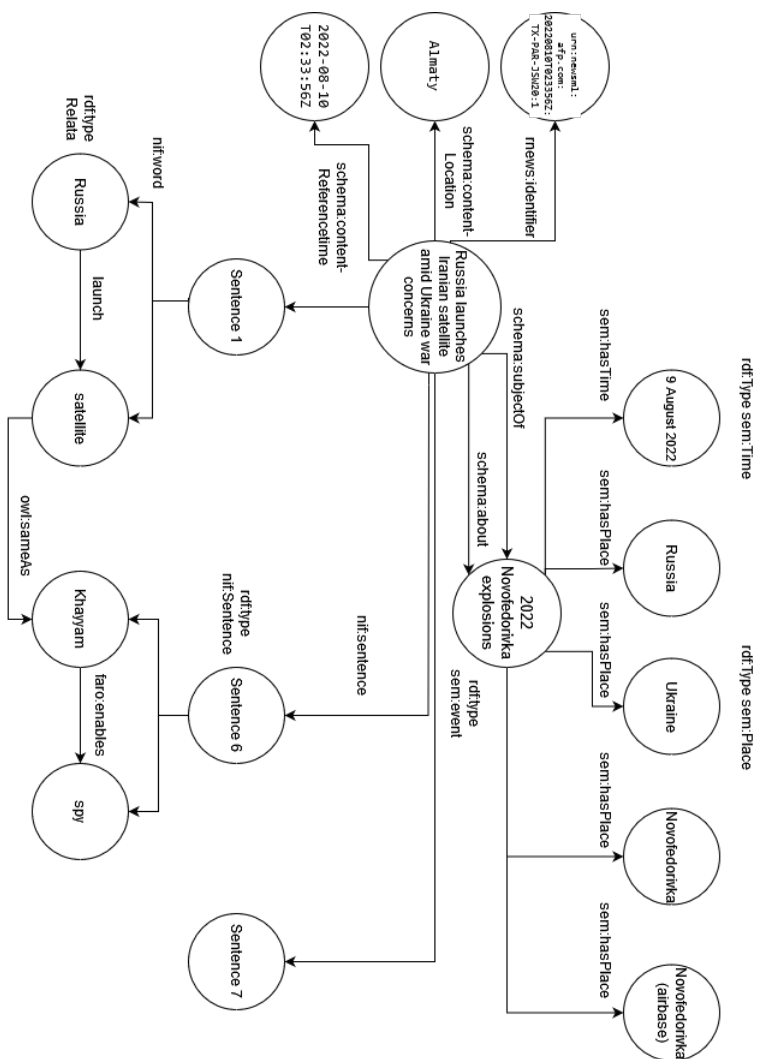### A.6. Example Narrative Graph of article: Russia launches Iranian satellite amid Ukraine war concerns

Figure 3: Part of the constructed narrative graph of the article "Russia launches Iranian satellite amid Ukraine war concerns"

A.7. Overview Narrative Graph

```
@prefix sem: <http://semanticweb.cs.vu.nl/2009/11/sem/> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix faro: <https://purl.org/faro/> .
@prefix schema: <http://schema.org/> .
@prefix rnews: <http://iptc.org/std/rNews/2011-10-07#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```
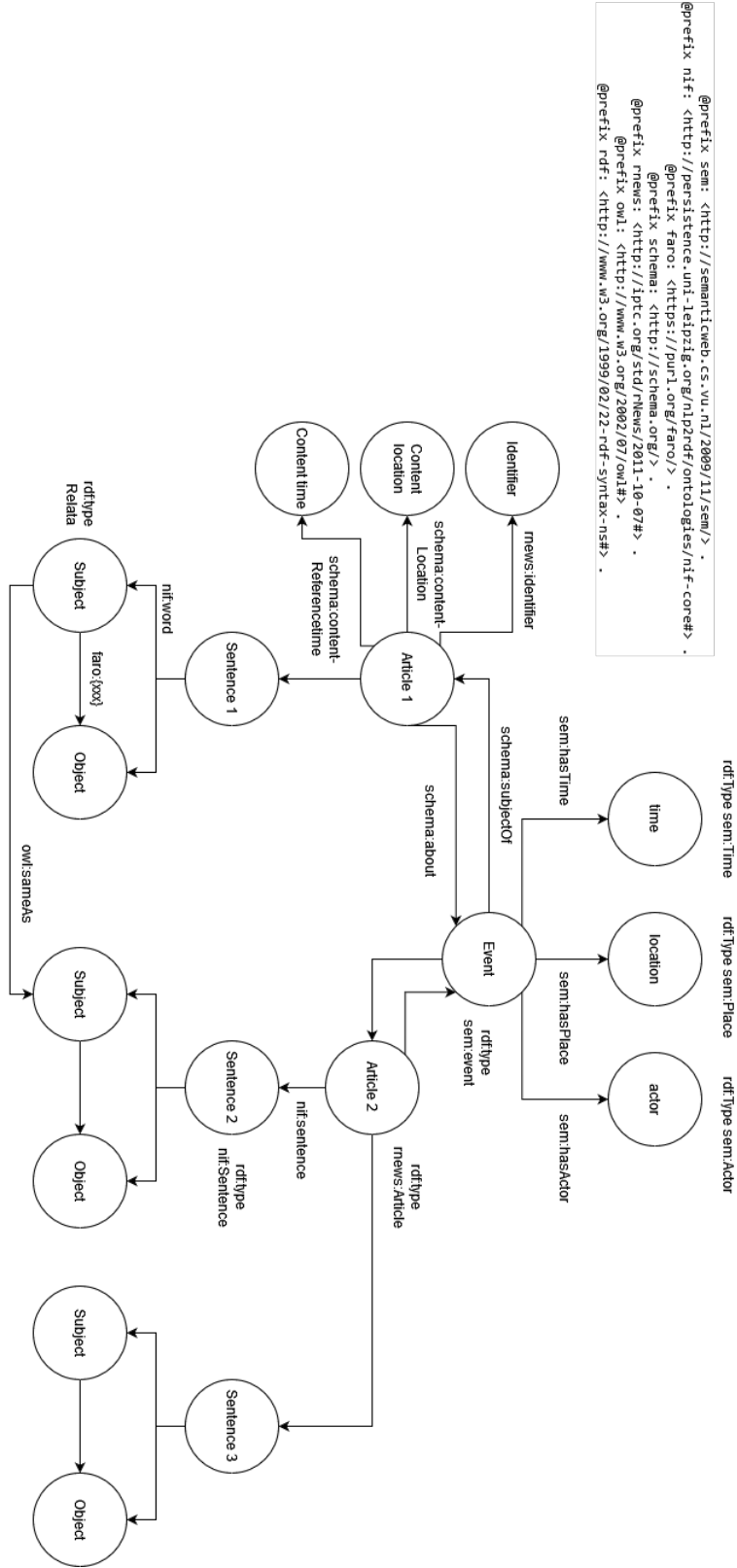
Figure 4: Overview of the Narrative Graph