Towards Spatially Grounded and Multilingual Event Relation Extraction

Jan Bongard[1,*], Xuke Hu[1], Jens Kersten[1], Alexander Brenning[2] and Friederike Klan[1]

[1]Institute of Data Science, German Aerospace Center, Mälzerstraße 3-5, 07745 Jena, Germany
[2]Department of Geography, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany

Abstract

Identifying driving factors in cascading crisis events is essential for a coordinated response, yet the standard event-relation task lacks spatial and multilingual capabilities. This paper presents a graph-based formulation of document-level event relation extraction that jointly models event detection, causal relations, and spatial grounding in crisis-related texts. We introduce an annotation scheme and a small preliminary multilingual benchmark, annotated in English to enable language-independent evaluation. A cross-lingual node alignment strategy is proposed to compare predicted event graphs with English ground truth. Experiments with a large language model baseline show that node alignment is the main performance bottleneck that strongly affects relation extraction. The results highlight the need for robust cross-lingual alignment as a foundation for meaningful evaluation and further development of multilingual extraction pipelines.

Keywords

Event Relation Extraction, Spatial Grounding, Multilingual Evaluation

1. Introduction

Web and social media are important real-time sources of information for crisis management, providing insights into rapidly evolving situations such as natural disasters or humanitarian emergencies [1, 2]. One method to structure heterogeneous and multilingual text is to represent events, their attributes, and relationships in a knowledge graph, supporting reasoning, situational awareness, and decision making [3, 4]. Event Relation Extraction (ERE) aims to identify causal, temporal, and hierarchical relationships between events in unstructured text [5]. These relations form the foundation of event-centric KGs and enable decision-makers to track the development of events, recognize cascading effects, and respond effectively.

ERE typically models temporal and causal relationships, but rarely considers spatial context. Although event coreference resolution approaches usually consider spatial information, their focus is on linking multiple mentions of the same event across sentences and documents, rather than modeling spatiotemporal and causal relations among distinct events [5, 6]. In addition, existing multilingual ERE data sets are often limited in scope, language coverage, and data size [7, 8, 9], reducing their applicability to real-world crisis scenarios.

In practice the absence of training data leads to the use of predefined causal markers (e.g., "because", "led to", "due to") and hand-engineered syntactic rules, which limits ERE to sentence-level and explicit mentioned relationships [10]. However, causal and spatio-temporal relationships usually span across multiple sentences. Event Causality Identification and Spatial Grounding address this limitation by modeling global context and long-range dependencies, enabling the detection of implicit causal and spatial relations beyond lexical cue words. Large Language Models (LLMs) outperform traditional methods in detecting such implicit relations [11, 12], and their extracted event representations provide

GeoExT 2026: Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026, April 2, 2026, Delft, The Netherlands

*Corresponding author.
jan.bongard@dlr.de (J. Bongard); xuke.hu@dlr.de (X. Hu); jens.kersten@dlr.de (J. Kersten); alexander.brenning@uni-jena.de (A. Brenning); friederike.klan@dlr.de (F. Klan)
0000-0001-9453-7391 (J. Bongard); 0000-0002-5649-0243 (X. Hu); 0000-0002-4735-7360 (J. Kersten); 0000-0001-6640-679X (A. Brenning); 0000-0002-1856-7334 (F. Klan)

published 2026-04-29

a foundation for integrating information across sources to construct a more comprehensive picture of crisis events [13, 6].

These circumstances motivate approaches that jointly integrate spatial information and multilingual analysis. As a first step, this paper explores a graph-based formulation of ERE for crisis-related text, jointly addressing Event Detection (ED), document-level Event Causality Identification (DECI), and Spatial Grounding (SG). A tailored annotation scheme is proposed and a preliminary multilingual benchmark is annotated based on English translations including a mapping from original to English text parts to enable language-independent evaluations. Finally, an evaluation setup for multilingual results is proposed, using a LLM as a baseline.

2. Methodology

The proposed task defines multilingual ERE as a mapping problem. News articles are manually annotated on their English translations to produce ground truth graphs, which are then evaluated against predicted event graphs extracted from the source language. The design decouples annotations from the source language, enabling language-independent evaluation of multilingual predictions. This chapter introduces an annotation scheme and a preliminary benchmark for the creation of document-level, spatially grounded causal event graphs, while highlighting the core challenges of the proposed task and its evaluation.

2.1. Benchmark Construction

News articles are collected from the GDELT news aggregator platform, which provides rich metadata including language, location, and Global Knowledge Graph (GKG) themes [14]. To cover different languages, Ukrainian, Chinese, and German are selected, and the Ukrainian city Kharkiv is chosen as the case study location. To ensure topical diversity, we group GKG themes into broader categories, using the World Bank Group Topical Taxonomy [15] as a rough reference, and focusing on the two topics of *Crime* and *Health Nutrition and Population*. For each topic and language, three articles are randomly sampled, resulting in a total of 18 news articles whose HTML content is retrieved, and boilerplate (i.e. non-content elements like advertisements or navigation menus) is removed to obtain the article text[1].

For the annotation, all texts are translated into English using Google Translate[2] and afterwards imported into an annotation tool with relation extraction capabilities[3]. So far, annotation was carried out by a single annotator to explore the feasibility of the process. Based on existing event-causality frameworks [16, 17], an initial annotation guideline is developed around the following five core components:

(1) Event detection identifies events that function as meaningful components within a cause-effect chain. The event relevance is defined at document level and depends on the main topic, argumentative focus, or narrative structure of the text. When an event is mentioned multiple times, only the most explicit formulation is annotated, and stylistic variations are not treated as distinct events.

(2) Event type classification assigns each detected event to one of the DPSIR categories (Driver, Pressure, State, Impact, Response). The DPSIR framework provides a generic structure for organizing causal processes [18, 19] and reflects basic causal principles of domain-specific models [20, 21]. It is applied as an ordering principle that supports the identification of events and their causal flow during the annotation. For a specific event, its type assignment may therefore vary across documents depending on the perspective and storyline.

(3) Causal relation annotation captures directed cause-effect relations between events, including both explicit and implicit causality. Relations are annotated only when both events are mentioned within the same paragraph or discourse unit, the causal direction is unambiguous, and the relation can be inferred from the document context alone, without introducing external assumptions. Complex causal structures, including many-to-many and many-to-one relations are allowed.

[1] https://pypi.org/project/trafilatura/
[2] https://pypi.org/project/deep-translator/
[3] https://labelstud.io/

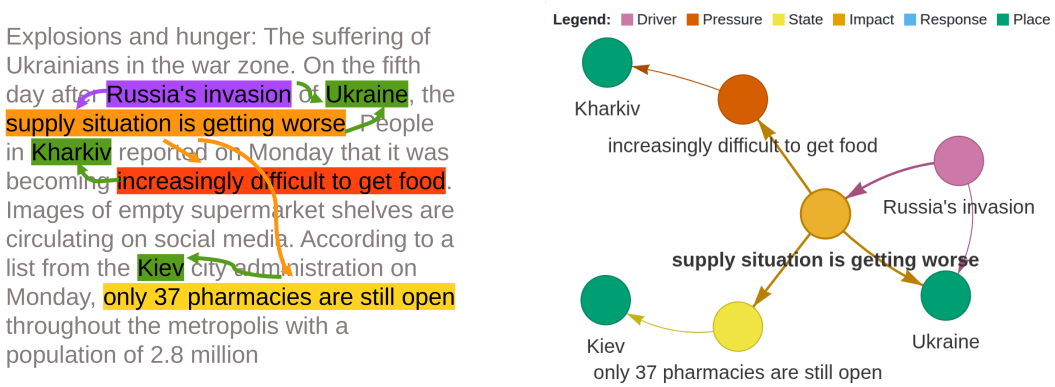

Figure 1: Annotation example. Translated and annotated text [22] (left), extracted ground truth graph (right).

(4) Spatial grounding links events to identifiable toponyms such as cities, countries, or other geographic features, while avoiding vague spatial references. Events may be associated with multiple locations, and implicit locations are inferred from contextual or linguistic cues.

(5) An annotator-based graph review finalizes the annotation process. As shown in Figure 1, the graph integrates events, directed causal relations, and spatial grounding, capturing the document's central event-related causal information while avoiding redundant structures.

2.2. LLM Baseline and Task Evaluation

For a first evaluation, event-relation graphs are extracted from multilingual documents using an LLM and compared with English ground truth graphs, focusing on the ED, DECI, and SG tasks. GPT-OSS 120B [23] is used in a prompt-based setting (see Appendix A), where all extraction tasks are jointly requested based on the annotation guidelines, and the KOR package[4] is used to convert the outputs into structured, machine-readable event triples.

The cross-lingual comparison is enabled through node alignment. Node texts are embedded using the multilingual LaBSE model [24], and a global one-to-one semantic alignment is computed based on cosine similarity by selecting the best overall matching between predicted and ground truth nodes. Alignments with similarity above 0.75 are counted as true positives (TP), predicted nodes without alignment are false positives (FP), and unaligned ground truth nodes are false negatives (FN). This procedure is performed independently for event alignment and toponym alignment.

The node alignment is then used to evaluate spatial grounding and causal relations between events in terms of causal existence and causal direction. For each case, predicted triples (e.g., $event_1 - \texttt{causes} - event_2$ or $event_1 - \texttt{occurs\_in} - toponym$) are mapped to English via aligned nodes and counted as TP if they exactly match a ground truth triple, otherwise they are FP. False positives are further divided into FP node-aligned, where node alignment exists but the triple is incorrect, and FP node-unaligned, where no valid node alignment exists. Unmatched ground truth triples are FN.

Standard precision, denoted as $\text{Prec} = \frac{\text{TP}}{\text{number of predicted triples}}$, combines relation and alignment errors. Therefore, a conditional precision $\text{Prec}_{\text{cond}} = \frac{\text{TP}}{\text{number of node-aligned predicted triples}}$ is reported, which measures performance only over predictions with valid node alignment. Recall is defined as $\text{Rec} = \frac{\text{TP}}{\text{number of ground truth triples}}$. All metrics are computed globally by pooling all predicted and ground truth triples across the dataset.

2.3. Preliminary Results

As depicted in Figure 2, toponym alignment and event alignment achieve moderate performance, with F1 scores of 0.68 and 0.55, respectively. For event alignment, the error analysis shows a comparable number of FP and FN, indicating that many ground truth nodes lack a corresponding prediction while

[4] https://pypi.org/project/kor/

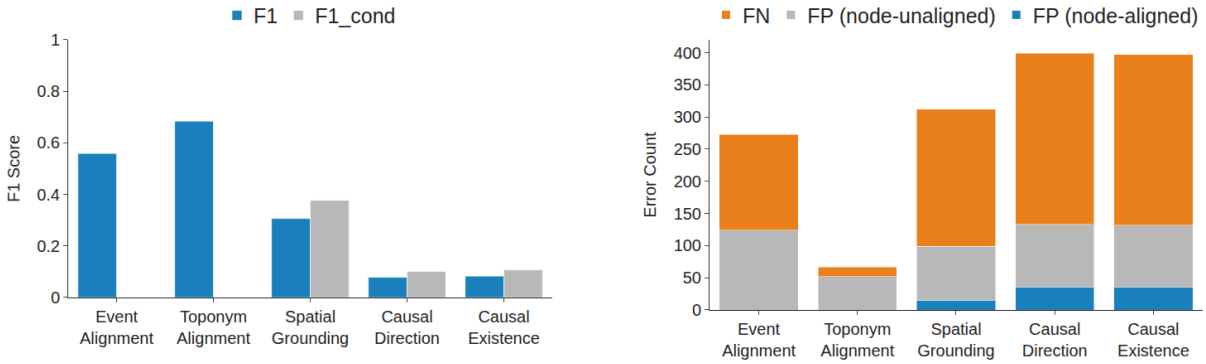

Figure 2: Performances of node matching and across relation types using F1 score (left) and the number of instances contributing to False Positives (FP) or False Negatives (FN) (right).

the model also generates additional nodes not present in the ground truth. This pattern suggests both node overgeneration and limitations in the semantic alignment procedure. In addition, it cannot be ruled out that some alignments are semantically similar but nevertheless incorrect.

These node-level errors propagate to relation predictions, where FN dominate, indicating that the model often fails to predict relations present in the ground truth graph. This highlights recall as the primary limitation. In contrast, precision is mainly affected by predictions that cannot be aligned to any ground truth node, resulting in a higher number of node-unaligned FP.

Across relation types, spatial grounding achieves higher F1 scores than causal relations. Comparing standard F1 with conditional F1 further shows that relation predictions become substantially more reliable once correct node alignment is given, indicating again that node alignment is the dominant source of error. Moreover, no meaningful difference is observed between causal existence and causal direction, suggesting that the primary challenge lies in identifying causal links at all, rather than in predicting their directionality.

3. Conclusion & Outlook

This work proposes a graph-based formulation of document-level ERE that jointly models event detection, causal relation extraction, and spatial grounding, investigated in crisis contexts. By introducing an annotation scheme and a preliminary language-independent evaluation framework, this work addresses key limitations of prior ERE research, including the lack of spatial modeling and multilingual document-level benchmarks.

Evaluating predicted event graphs in the source language against annotated English ground truth graphs at document level represents a promising direction for multilingual ERE assessment. Nevertheless, experimental results identify the cross-lingual node alignment as the primary source of error, affecting the evaluation of ERE. Since surface-level string matching via text spans are not suited to multilingual settings, the investigated semantic similarity approach offers a viable starting point, but more robust alignment methods are needed to establish event-agnostic mapping between English ground truth and predicted extractions from the source language.

In addition, the annotation scheme requires further development. Determining what constitutes a relevant event remains a core challenge to be able to reduce future inter-annotator inconsistencies. Particularly in open-domain extraction, annotators should be more guided in extracting salient events, carrying relational significance or are central to the document's context. To reduce cognitive load, future annotation will adopt a staged pipeline, addressing event detection and causal relations prior to spatial relations and geocoding.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-OSS 120B in order to: Improve writing style, Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Bhoi, H. B. Bhuyan, R. P. Nayak, R. C. Balabantaray, A. Pattanaik, A. Chinmay, Smart crisis response leveraging social media content for effective disaster management, Discover Computing 28 (2025) 1–25.

[2] N. Vračević, S. Schmidt, M. Keskin, D. Hanny, B. Resch, More than just tweets: the potential of alternative geo-social media data for disaster management, Social Network Analysis and Mining 15 (2025) 74.

[3] C. Li, Y. Wang, X. Cui, C. Zhu, et al., Review on the construction and application of knowledge graphs for natural disaster emergency response, Academic Journal of Environment & Earth Science 7 (2025).

[4] R. Zhu, C. Shimizu, S. Stephen, C. K. Fisher, T. Thelen, K. Currier, K. Janowicz, P. Hitzler, M. Schildhauer, W. Li, et al., The knowwheregraph: A large-scale geo-knowledge graph for interdisciplinary knowledge discovery and geo-enrichment, arXiv preprint arXiv:2502.13874 (2025).

[5] K. Liu, Y. Chen, J. Liu, X. Zuo, J. Zhao, Extracting events and their relations from texts: A survey on recent research progress and challenges, AI Open 1 (2020) 22–39.

[6] Q. Gao, Z. Meng, B. Li, J. Zhou, F. Li, C. Teng, D. Ji, Harvesting events from multiple sources: Towards a cross-document event extraction paradigm, arXiv preprint arXiv:2406.16021 (2024).

[7] V. D. Lai, A. P. B. Veyseh, M. Van Nguyen, F. Dernoncourt, T. H. Nguyen, Meci: A multilingual dataset for event causality identification, in: Proceedings of the 29th international conference on computational linguistics, 2022, pp. 2346–2356.

[8] A.-L. Minard, M. Speranza, R. Urizar, B. Altuna, M. Van Erp, A. Schoen, C. Van Son, Meantime, the newsreader multilingual event and time corpus, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 4417–4422.

[9] C. Walker, S. Strassel, J. Medero, K. Maeda, Ace 2005 multilingual training corpus, (No Title) (2006).

[10] X. Dong, E. Mas, B. Adriano, S. Koshimura, Towards real-time extraction of cascading effect and spatiotemporal analysis using social media data, International Journal of Disaster Risk Reduction (2025) 105512.

[11] Q. Cheng, Z. Zeng, X. Hu, Y. Si, Z. Liu, A survey of event causality identification: Taxonomy, challenges, assessment, and prospects, ACM Computing Surveys 58 (2025) 1–37.

[12] G. Katz, H. Sitton, G. Gonen, Y. Kaplan, Beyond the surface: Uncovering implicit locations with llms for personalized local news, arXiv preprint arXiv:2502.14660 (2025).

[13] Y. Tian, W. Li, Geoai for knowledge graph construction: Identifying causality between cascading events to support environmental resilience research, arXiv preprint arXiv:2211.06011 (2022).

[14] GDELT Project, GDELT: Global database of events, language, and tone, https://www.gdeltproject.org/, 2013. Accessed: 2026-01-28.

[15] World Bank Group, Wbg topical taxonomy, n.d. URL: https://vocabularyserver.com/worldbank/taxonomy/, accessed: 2026-01-28.

[16] Linguistic Data Consortium, ACE (automatic content extraction) english annotation guidelines for events: Version 5.4.3, https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf, 2005. Accessed: 2026-01-28.

[17] T. Caselli, P. Vossen, The event storyline corpus: A new benchmark for causal and temporal relation extraction, in: Proceedings of the Events and Stories in the News Workshop, 2017, pp. 77–86.

[18] E. Eea, Environmental indicators: Typology and overview, European Environmental (1999).
[19] S. R. Gari, A. Newton, J. D. Icely, A review of the application and evolution of the dpsir framework with an emphasis on coastal social-ecological systems, Ocean & coastal management 103 (2015) 63–77.
[20] P. Blaikie, T. Cannon, I. Davis, B. Wisner, At risk: natural hazards, people's vulnerability and disasters, Routledge, 2014.
[21] World Bank Group, Wbg topical taxonomyshardesai, shonali and wam, per, n.d.2002. URL: https://vocabularyserver.com/worldbank/taxonomy/hdl.handle.net/10986/11335, accessed: 2026-01-28.
[22] Vienna Online, Explosionen und hunger: Das leiden der ukrainer im kriegsgebiet, Vienna.at (2022). URL: https://www.vienna.at/explosionen-und-hunger-das-leiden-der-ukrainer-im-kriegsgebiet/7309724, accessed: 2026-01-29.
[23] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, et al., gpt-oss-120b & gpt-oss-20b model card, arXiv preprint arXiv:2508.10925 (2025).
[24] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 878–891.

A. LLM Extraction Prompt

The following instruction was passed to the KOR extraction chain alongside the document text and forwarded to the OpenAI GPT OSS 120B model. Nodes and edges were defined as KOR schema objects, with properties such as description, type or relation specified as object attributes. Returned nodes and edges were subsequently used to construct the document-level event graph.

KOR Instruction Template for Causal Graph Extraction

Your task is to extract a directed causal graph from the provided text to support cause-effect analysis for disaster response. Focus on mechanisms, not narrative background or rhetorical framing.

NODES

Description: Extract only nodes explicitly stated in the text. Each node represents a distinct causal event or geographic location. Select the most explicit instance; do NOT invent, infer, or duplicate nodes.

Type: Classify each node as exactly one of:

- `Driver` – root cause or external forcing factor
- `Pressure` – intermediate stress on a system
- `State` – condition or status of the affected system
- `Impact` – observed negative consequence
- `Response` – mitigation or reaction measure
- `Place` – only real-world geographic toponyms; events must never be classified as Place

Id: Short label in the source language:

- *Place nodes:* copy toponym verbatim (e.g. `Kharkiv, Luhansk, Ukraine`)
- *Event nodes:* concise description of the event extracted from the text verbatim

EDGES

Description: Extract edges explicitly stated or reasonably inferable. Every source and target must exactly match an existing node `id`; do NOT introduce new `ids`.

Relation: Exactly one of:

- `causes` – directed causal link; source and target must both be non-Place nodes
- `occurs_in` – links event node (source) to its Place node (target)

Source: The id of the source node. Must exactly match an existing node id
Target: The id of the target node. Must exactly match an existing node id