Leveraging LLMs for Event Extraction in Italian Documents: a Roadmap for Future Research

Federica Rollo*, Giovanni Bonisoli and Laura Po

"Enzo Ferrari" Engineering Department, University of Modena and Reggio Emilia, MO 41121 Italy

Abstract

Event extraction is a task of significant interest in the field of Natural Language Processing (NLP) and plays a vital role in various applications, such as information retrieval and document summarization. Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. In this paper, we present a roadmap for the application of LLMs for event extraction from Italian documents, aiming to address the gap in research and resources for event extraction in non-English languages. We first discuss the challenges of event extraction and the current state-of-the-art approaches based on LLMs. Next, we present potential Italian datasets suitable for adapting linguistic models to the domain of event extraction. Furthermore, we outline future research directions and potential areas for improvement in this evolving field.

Keywords

event extraction, Large Language Model, Italian language

1. Introduction

The recent development of Large Language Models (LLMs) poses significant promise for advancing several natural language-based tasks, including event extraction from lengthy text. LLMs such as GPT models [1] have demonstrated remarkable capabilities in understanding and generating natural language text. The application of LLMs for event extraction offers several advantages. Firstly, these models can process vast amounts of text data, enabling comprehensive analysis of events described in natural language. Secondly, LLMs can capture complex linguistic structures and contextual nuances typical of different kinds of documents, enhancing the accuracy of extracted event details. The continuous learning ability of LLMs allows them to adapt to different writing styles and language conventions.

However, challenges persist in leveraging LLMs for event extraction in languages other than English, particularly in languages with limited available resources such as Italian. Fine-tuning requires curated datasets that accurately represent the diversity of language and scenarios, and the annotation of different event-related data.

Despite these challenges, the potential of LLMs to revolutionize event extraction is substantial. For instance, Question Answering (QA) models can facilitate rapid and efficient access to relevant information by automatically identifying text spans containing the desired answers to specific questions. While other models can be provided with detailed instructions to extract specific data from the text. Integrating these models into NLP pipelines can streamline the process of real-time event analysis, allowing for timely and efficient extraction of event-related information from textual data. This paper explores the role of LLMs in advancing event extraction from lengthy text. In particular, we focus on the Italian language and we explore the resources available for adapting and evaluating LLMs to event extraction on Italian documents. In the end, we define possible future directions for research in this dynamic field.

2. Event Extraction

2.1. Task formulation

Event extraction aims at identifying and categorizing events described within a text, including the recognition of the entities involved in the event (such as individuals, organizations, or locations), and the extraction of temporal references and any elements that are relevant for the event. This task has gained significant popularity in recent years due to its broad applicability and practical utility in various real-world scenarios. Figure 1 shows an example of the results of event extraction from a document describing an air crash. In addition to the identification of the event type, different event roles have been annotated, e.g., the date of the event occurrence, the aircraft agency.

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding author.

federica.rollo@unimore.it (F. Rollo); giovanni.bonisoli@unimore.it (G. Bonisoli); laura.po@unimore.it (L. Po)

0000-0002-3834-3629 (F. Rollo); 0000-0001-8538-8347 (G. Bonisoli); 0000-0002-3345-176X (L. Po)

Flight 345 was a SriLankan Airlines flight from London Heathrow Airport via Dubai to Colombo (Bandaranaike International Airport).

On 3 June 2015, the Airbus A330-300 operating the flight was on the ground in Colombo, about to fly on to Malé, when an explosion ripped the aircraft in two, destroying it.

Flight 345 carried mainly French, German, British and Japanese tourists; 23 people were killed on the aircraft, including 3 British, 4 French, 2 Japanese; 44 people were injured.

Boarding of the flight had been delayed due to the aircraft being damaged during cargo loading.

During boarding, a bomb, hidden in the aircraft's 'Fly Away Kit' (a collection of small spare parts), exploded…

Event Type:	Air Crash
Event Roles	
Date	"3 June 2015"
Aircraft Agency	"SriLankan Airlines"
Taking-off Place	"London Heathrow Airport"
Flight Number	"Flight 345"
Casualties and Losses	"23 people were killed", "44 people were injured"

Figure 1: Example of event extraction.

2.2. Challenges

Due to the complexity of the natural language, event extraction poses several challenges that require sophisticated techniques to address effectively.

The first challenge consists of detecting multiple events described in the same document and understanding which are the references to each event. Natural language often contains ambiguous expressions that can refer to multiple events or entities. This ambiguity, along with the use of coreference, further complicates the task of accurately extracting event data from text since resolving ambiguity requires contextual understanding and disambiguation techniques.

Identifying relevant elements for each event requires distinguishing between event triggers (words or phrases that indicate the occurrence of an event) and background information and noise. Another complexity is given by the variability in language usage, writing styles, syntactic structures, and document length. Indeed, event extraction can be performed on short text like tweets, longer documents such as news articles, and lengthy documents such as investigative reports or government documents. All these factors require the use of techniques able of accommodating these variations to achieve accurate and reliable results across diverse text types and genres.

Two of the key aspects of events are the time and the space, i.e., when the event took place and where. The recognition and standardization of temporal and spatial expressions could be complex since temporal reference can be expressed in various formats (such as dates, times, part of the day). In addition, a document describing an event can refer to the location of the event providing information at different granularity, for example indicating the name of the city, specifying the address, and/or describing the type of the place like an apartment, a shop, or a park. During event extraction, the references to all these locations should be identified.

2.3. Large Language Models based approaches

Several approaches have been proposed for event extraction in recent surveys, from traditional methods which rely on the use of linguistic rules for pattern identification within the text to more advanced solutions such as machine learning and deep learning algorithms able to learn patterns after training on annotated data, and the use of pre-trained language models [2, 3]. LLMs based approaches have emerged as a promising avenue for event extraction in recent years. These models leverage the power of machine learning and deep learning algorithms as they are pre-trained on vast amounts of text data and then fine-tuned for specific tasks. By encoding contextual information and capturing semantic relationships within the text, LLMs seem to be promising in identifying and extracting events from various sources.

We identified three main approaches based on the use of LLMs that could reach good performance in event extraction: sequence labeling models, extractive Question Answering (QA) models and instruction-tuned models.

Sequence Labeling models In Sequence labeling each token in a sequence is assigned a label based on its role or category within the context of the sequence. Sequence labeling models can be used to identify those text spans reporting relevant information within a text. Therefore, it is widely employed for several classical NLP tasks like part-of-speech (POS) tagging, named entity recognition (NER), text chunking.

Sequence labeling models are suitable for the scenario of event extraction, where they can identify and classify those parts of text reporting information about events. Indeed, some works in literature have already treated event extraction as a sequence labeling or NER problem, [4, 5], also for Italian Language [6].

Extractive Question Answering The goal of extractive QA models is to understand an input question in natural language and extract the answer as a span from an input text. QA models can facilitate rapid and efficient access to event-related information by automatically identifying text spans containing the desired answers to specific questions. For instance, the question "When did the event take place?" (Q1) can be formulated to retrieve the date of the event.

The results of these models depend significantly on the quality of the input documents, as well as the structure of the questions provided to the models. Prior knowledge about the kind of event described in the document allows to formulate ad hoc questions. For instance, considering the document in Figure 1, the question "When did the air crash take place?" (Q2) should provide more accurate answers than Q1. In addition, questions should be enriched by other details about the event after a partial process of event extraction. For example, the question "When did the Flight 345 crash?" (Q3) contain the reference to the flight number and should help the QA models to select the correct context for the extraction of the date.

Within the QA models, distinctions arise between Single-Span QA (SQA) and Multi-Span QA (MQA). While the former identifies a single text segment for each question, the latter locates answers even when distributed across non-consecutive text segments, potentially located far apart within a document. Given the prevalence of such scenarios, especially in complex inquiries and detailed documents, the limitations of SQA models are evident. An example is the annotation of "causalities and losses" in Figure 1. The recent surge in MQA model development [7, 8, 9] underscores a notable interest.

In the current state-of-the-art, the only Italian dataset properly designed for training QA models is SQuAD-it [10], derived from the automatic translation of the English SQuAD dataset, consisting of a list of pairs question-answer. However, this dataset can be used only for SQA, therefore it is unsuitable for complex tasks like event extraction which requires the ability to retrieve multiple spans for one question.

Instruction-Tuned models Among LLMs, Auto-Regressive models such as GPT [1] or Llama [11] series stand out. These models leverage advanced deep learning techniques to predict the subsequent word based on an input text. This prediction process is repeated multiple times, with each predicted word being added to the original text. By training on vast amounts of text data, Auto-Regressive LLMs effectively capture complex patterns and structures in language, leading them to generate full and coherent text which is contextually relevant to input text.

The research in recent years has led to the development of instruction tuning [12] to bridge the gap between the next-word prediction objective of LLMs and the users' objective of following their instructions helpfully and safely. Instruction-tuning involves a fine-tuning of Auto-Regressive LLMs with input-output pairs, where input denotes the human instructions, and output denotes the desired output that follows the instruction. The results of this process are the Instruction-Tuned LLMs, designed specifically to provide appropriate results based on instruction inputs. This ability is also enhanced as a cross-task generalization, leading Instruction-Tuned LLMs to better performances on novel tasks.

Instruction-Tuned LLMs can be employed to solve a wide range of NLP tasks through various techniques of prompt engineering [13], i.e., the process of designing task-specific instructions to guide model output. Therefore, the utilization of these models can also yield benefits for event extraction.

Currently, there are several Instruction-Tuned LLMs capable of understanding and generating text. For those, Italian represents a minority percentage in the training data compared to more widely used languages on the web such as English. Among these, there are proprietary models like GPT-3.5 and GPT-4 from OpenAI, Gemini from Google, and open-source families of LLMs like Mistral [14] and Mixtral [15] from Mistral AI and Llama [11] and Llama 2 [16] from Meta. From this last family, Llamantino [17] has been derived through a language adaptation process to the Italian Language.

3. Italian datasets

Currently, there are few Italian datasets suitable for event extraction. Some of them provide a comprehensive annotation of event-related data, while in other cases, only one type of information (e.g., the temporal references) is annotated.

3.1. EVENTI

The EVENTI[1] corpus was built in 2014 for the evaluation of Temporal Information Processing systems of the EVENTI evaluation exercise [18] in the EVALITA workshop. The corpus consists of three datasets: the Main task training data (274 documents) and test data (92 documents) of contemporary news articles and the Pilot task (10 documents) test data of historical news articles. The annotation guidelines involve the use of four tags to annotate different elements within news texts: the EVENT tag is used to annotate all the mentions of events including verbs, nouns, prepositional phrases and adjectives; the TIMEX3 tag is used for temporal expressions; the SIGNAL tag identifies textual items which encode a relation either between EVENTs, or TIMEX3s or both; the

[1] https://sites.google.com/site/eventievalita2014/data-tools

TLINK tag is used for temporal dependencies between EVENTs and/or Temporal Expressions.

3.2. NewsReader MEANTIME

The NewsReader MEANTIME (Multilingual Event ANd TIME) is a multilingual semantically annotated corpus of 480 Wikinews articles in four languages: English, Italian, Spanish, and Dutch [19]. The corpus was released in 2016 and derives from the NewsReader Project[2] [20] which aims at extracting information about what happened to whom, when, and where, processing a large volume of financial and economic data. The corpus is enriched with annotations that span multiple levels, including entities, entity mentions, events, temporal information, semantic roles, and intra-document and cross-document event and entity coreference.

3.3. De Gasperi

The De Gasperi corpus [21] is a collection of historical documents by Alcide De Gasperi, the first Prime Minister of the Italian Republic. The corpus was released in 2019 and includes 2,762 documents published between 1901 and 1954, originally released in an oral or written form. In addition to the raw text, a set of meta-data and additional semiautomatically annotated information are available. The corpus contains different kinds of documents, like daily press written by De Gasperi when he worked as a journalist for newspapers in Trentino, and speeches in institutional venues when he was a Member of the Italian Parliament. In each document, references to persons and places are annotated.

3.4. DICE

DICE[3] [22] is a collection of 10,395 Italian news articles describing crime events that happened in the Modena province between 2011 and 2021. The news articles are extracted from one of the most popular local newspapers, "Gazzetta di Modena", following the approach described in [23]. Thanks to an agreement between the University of Modena and Reggio Emilia and the Gazzetta di Modena, DICE was released online in 2023, free to redistribute and transform without encountering legal copyright issues under an Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Along with the data related to the title, the text, and the publication date of each news article that are crawled from the newspaper's webpage, several annotations are available on the data. The crime event category (e.g., theft, robbery) is assigned to each news article using text categorization approaches based on word embeddings [24, 25]. The news articles underwent automated NLP processes to extract temporal references, entities, and corresponding DBpedia resources. Duplicates are annotated to identify news articles referring to the same crime event. The theft-related news articles are annotated manually following a sophisticated annotation schema to identify stolen items (`What`), crime locations (`Where`), references to authors and victims, and their sociodemographic characteristics (`Who`). The annotation provided in the dataset is multi-span since it involves identifying and linking multiple text spans within the document.

3.5. EventNet-ITA

EventNet-ITA[4] [26] is an Italian corpus for Frame Parsing applied to events released in 2024. Semantic Frame Parsing is a task which aims at identifying semantic frames within textual data. A semantic frame [27] is a cognitive structure that organizes and represents knowledge about a concept or situation. It consists of a set of interconnected elements such as roles, attributes, and relations, which collectively define the meaning and typical features of that concept or situation. Frames help humans understand and interpret language by providing a mental framework for comprehending and categorizing information.

EventNet-ITA is built upon the idea of enabling frame parsing for event extraction. It is composed of 53,854 sentences manually annotated with 205 semantic frames of events and covers different domains, like conflictual, social, communication, legal, geopolitical, economic and biographical events.

4. Future directions

Automated information extraction from documents continues to captivate the scientific community due to its manifold advantages, facilitating improved information accessibility across various domains. By leveraging LLMs and exploiting annotated datasets, researchers can develop robust event extraction systems capable of achieving high accuracy and efficiency across a wide range of text sources. As the field continues to advance, further research into LLMs and their applications in event extraction is expected to drive continued innovation and progress in this area.

Future directions will focus on three key aspects:

- **Definition of an Italian benchmark**: while we have identified five Italian datasets suitable for event extraction, further efforts are needed to expand their annotation and support comprehensive event extraction tasks. This entails defining

[2] http://www.newsreader-project.eu/
[3] https://github.com/federicarollo/Italian-Crime-News
[4] https://huggingface.co/datasets/mrovera/eventnet-ita

a standardized benchmark for evaluating event extraction systems. Such a benchmark would serve as a common evaluation dataset, enabling comparisons between different approaches and fostering the development of more accurate and reliable event extraction models.

- **Evaluation of LLMs on the benchmark**: despite the limited literature on Italian event extraction, our preliminary evaluation of three BERT-based QA models on the DICE dataset revealed promising results [22]. However, challenges persist, particularly related to the size and quality of the annotated data. Once the benchmark is defined, future efforts will focus on evaluating and comparing various approaches outlined in Section 2.3. The evaluation will include the recent Minerva models that represent the first family of LLMs trained from scratch on Italian documents developed by Sapienza NLP.
- **Creation of a synthetic annotated dataset**: since manual annotation is a time and resource-consuming process, new strategies will be studied to automate the process of annotation. Employing LLMs for data augmentation (i.e., to expand the annotated dataset) is now the most promising approach, especially focusing on text generation models. Given a list of desired annotations, i.e., the spans to extract from the text (like "May 14th" as the date of the event), the LLM is asked to create a document with that span with the expected role in the event described (like "create a document describing an event that occurred on May 14th"). This methodology allows for obtaining a synthetic dataset that is also already annotated. Furthermore, this approach offers control over text generation and ensures fairness in dataset composition, ultimately contributing to the development of balanced and unbiased datasets essential for training accurate and equitable AI models.

References

[1] 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: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[2] G. Frisoni, G. Moro, A. Carbonaro, A survey on event extraction for natural language understanding: Riding the biomedical literature wave, IEEE Access 9 (2021) 160721 – 160757. doi:10.1109/ACCESS.2021.3130956.

[3] W. Xiang, B. Wang, A survey of event extraction from text, IEEE Access 7 (2019) 173111–173137. doi:10.1109/ACCESS.2019.2956831.

[4] A. Ramponi, R. van der Goot, R. Lombardo, B. Plank, Biomedical event extraction as sequence labeling, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 5357–5367. doi:10.18653/v1/2020.emnlp-main.431.

[5] S. Pongpaichet, B. Sukosit, C. Duangtanawat, J. Jamjongdamrongkit, C. Mahacharoensuk, K. Matangkarat, P. Singhajan, T. Noraset, S. Tuarob, Camelon: A system for crime metadata extraction and spatiotemporal visualization from online news articles, IEEE Access 12 (2024) 22778–22802. doi:10.1109/ACCESS.2024.3363879.

[6] N. Viani, T. A. Miller, D. Dligach, S. Bethard, C. Napolitano, S. G. Priori, R. Bellazzi, L. Sacchi, G. K. Savova, Recurrent neural network architectures for event extraction from italian medical reports, in: A. ten Teije, C. Popow, J. H. Holmes, L. Sacchi (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, Cham, 2017, pp. 198–202.

[7] H. Li, M. Tomko, M. Vasardani, T. Baldwin, Multispanqa: A dataset for multi-span question answering, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 1250–1260. doi:10.18653/V1/2022.NAACL-MAIN.90.

[8] E. Segal, A. Efrat, M. Shoham, A. Globerson, J. Berant, A simple and effective model for answering multi-span questions, 2020, p. 3074 – 3080.

[9] M. Zhu, A. Ahuja, D. Juan, W. Wei, C. K. Reddy, Question answering with long multiple-span answers, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, Association for Computational Linguistics, 2020, pp. 3840–3849. doi:10.18653/V1/2020.FINDINGS-EMNLP.342.

[10] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, Lecture Notes

in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11298 LNAI (2018) 389 – 402. doi:10.1007/978-3-030-03840-3_29.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[12] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, G. Wang, Instruction tuning for large language models: A survey, 2024. arXiv:2308.10792.

[13] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, 2024. arXiv:2402.07927.

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[15] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. arXiv:2401.04088.

[16] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[17] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.

[18] T. Caselli, R. Sprugnoli, M. Speranza, M. Monachini, Eventi evaluation of events and temporal information at evalita 2014, 2014. doi:10.12871/clicit201425.

[19] 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), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 4417–4422.

[20] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. P. Aprosio, G. Rigau, et al., Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news, Knowledge-Based Systems 110 (2016) 60–85.

[21] S. Tonelli, R. Sprugnoli, G. Moretti, Prendo la parola in questo consesso mondiale: A multi-genre 20th century corpus in the political domain, volume 2481, 2019.

[22] G. Bonisoli, M. P. di Buono, L. Po, F. Rollo, DICE: a dataset of italian crime event news, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 2985–2995. doi:10.1145/3539618.3591904.

[23] F. Rollo, L. Po, Crime event localization and deduplication, in: J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), The Semantic Web – ISWC 2020, Springer International Publishing, Cham, 2020, pp. 361–377.

[24] F. Rollo, G. Bonisoli, L. Po, A comparative analysis of word embeddings techniques for italian news categorization, IEEE Access 12 (2024) 25536 – 25552. doi:10.1109/ACCESS.2024.3367246.

[25] F. Rollo, G. Bonisoli, L. Po, Supervised and unsupervised categorization of an imbalanced italian crime news dataset, Lecture Notes in Business Information Processing 442 LNBIP (2022) 117 – 139. doi:10.1007/978-3-030-98997-2_6.

[26] M. Rovera, EventNet-ITA: Italian frame parsing for events, in: Y. Bizzoni, S. Degaetano-Ortlieb, A. Kazantseva, S. Szpakowicz (Eds.), Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 77–90.

[27] C. J. Fillmore, C. F. Baker, Frame semantics for text understanding, in: Proceedings of WordNet and Other Lexical Resources Workshop, NAACL, volume 6, 2001.