DAVE: A Framework for Assisted Analysis of Document Collections in Knowledge-Intensive Domains

Ruben Agazzi[1,2] , **Renzo Alva Principe**[1] , **Riccardo Pozzi**[1] , **Marco Ripamonti**[1] and **Matteo Palmonari**[1]

[1]University of Milano-Bicocca, Milan, Italy
[2]National Interuniversity Consortium for Informatics (CINI), Rome, Italy
{ruben.agazzi,renzo.alvaprincipe,riccardo.pozzi,marco.ripamonti,matteo.palmonari}@unimib.it

Abstract

DAVE is a framework for assisting the analysis of documents in knowledge-intensive domains, based on an entity-centric approach supported by annotations of named entities in the documents. DAVE supports search & filtering, document exploration, question answering, and knowledge refinement. It is released as an open-source project that the community can further develop. DAVE's distinguishing features are: the integration of a chatbot interface based on recent RAG solutions into well-established entity-powered faceted search, the fusion of search and filtering features provided by entity-level annotations with the capability to ask questions on annotated documents; human-in-the-loop functions to consolidate knowledge while exploring information, allowing users to improve annotations from NLP algorithms.

1 Introduction

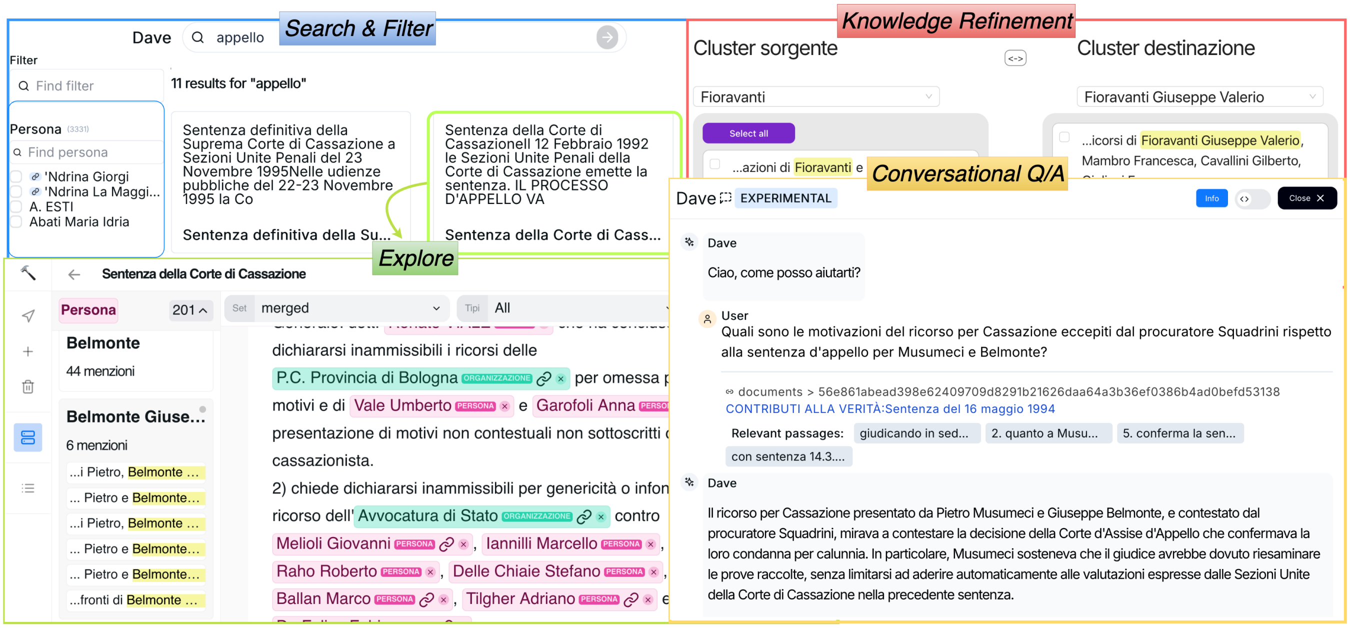
Digitalization has expanded access to large document collections, making vast amounts of previously inaccessible information available in digital form. To support users find useful information in document collections and analyze their content, different paradigms have been established to account for different information needs, from exploratory search interfaces [Liu *et al.*, 2024], which typically combine search and filtering functionalities (e.g., under the faceted search paradigm), to question answering systems. The latter ones have become more popular with the advent of Large Language Models (LLMs). Supported by Retrieval Augmented Generation (RAG) architectures, LLMs simplify the development of systems that answer questions in natural language on top of specific document collections [Lewis *et al.*, 2020].

Notably, many domains are inherently entity-centric, where factual information is closely linked to entities that define the context and relevance of a document. For instance, in fields like law and healthcare, professionals seek information on case laws, regulations, diseases, and treatments to support precise search and compliance monitoring. As a result, their information needs are strongly entity-driven.

Consequently, a crucial step toward building entity-aware systems is the adoption of Entity Extraction (EE) approaches to identify and classify entity mentions (i.e., Named Entity Recognition [Li *et al.*, 2022]) and identify links across these mentions. These links can be derived indirectly using background Knowledge Bases (KBs) by applying Named Entity Linking (NEL) techniques [Sevgili *et al.*, 2022] (all mentions linked to the same identifier in the KB are deemed to refer to the same entity) or directly by applying co-reference resolution techniques [Logan IV *et al.*, 2021]. The two approaches can also be somehow combined in end-to-end pipelines combining different components [Pozzi *et al.*, 2023a]. The result of these EE extraction techniques can be used to attach entity-level annotations to the documents, supporting downstream applications for document search and filtering, for example, exploiting faceted search or other semantic search interfaces [Tunkelang, 2022]. However, it is worth noting that EE techniques or more sophisticated methods based on these techniques, are being increasingly used also in RAG applications to improve retrieval and answer formulation in chatbots. Flagship examples of these initiatives are Graph RAG approaches [Edge *et al.*, 2024]. Nonetheless, even more lightweight approaches that enrich content and questions with entities have been shown extremely effective in vertical domain [Xu *et al.*, 2024]. Yet, while both these two paradigms, faceted search and chatbots, can take advantage of entities, their integration is, to the best of our knowledge, limited. Even when used to improve RAG techniques behind chatbots, entities remain in the background.

In this paper, we introduce DAVE, a tool for assisted analysis of document collections in knowledge-intensive domains. The tool goal is to support search needs that span across different points of the extractive vs. abstractive spectrum, as discussed in [Worledge *et al.*, 2024]. It features a graphical user interface (GUI) that enables users to visualize, explore, and query documents. Our tool is specifically tailored for domains where entities are first-class citizens in document analysis and provide the main following features: (i) an entity-driven faceted search interface for entity-driven exploration; (ii) a conversational interface supporting complex natural language queries; (iii) seamless integration of faceted search and conversational interaction to refine document sources; (iv) a human-in-the-loop mechanism for refining entity annotations, ensuring corrections are propagated throughout the system.

DAVE is designed as an open system and is released as



11 results for "appello"

□ 'Ndrina Giorgi
□ 'Ndrina La Maggi...
□ A. ESTI
□ Abati Maria Idria

Sentenza definitiva della Suprema Corte di Cassazione a Sezioni Unite Penali del 23 Novembre 1995Nelle udienze pubbliche del 22-23 Novembre 1995 la Co

Sentenza definitiva della Su...

Sentenza della Corte di Cassazionell 12 Febbraio 1992 le Sezioni Unite Penali della Corte di Cassazione emette la sentenza. IL PROCESSO D'APPELLO VA

Sentenza della Corte di Cass...

Fioravanti

…azioni di Fioravanti e

Fioravanti Giuseppe Valerio

…icorsi di Fioravanti Giuseppe Valerio, Mambro Francesca, Cavallini Gilberto,

Sentenza della Corte di Cassazione

Belmonte
44 menzioni

Belmonte Giuse…
6 menzioni
...i Pietro, Belmonte …
... Pietro e Belmonte…
...i Pietro, Belmonte …
... Pietro e Belmonte…
... Pietro e Belmonte…
...fronti di Belmonte …

dichiararsi inammissibili i ricorsi delle P.C. Provincia di Bologna ORGANIZZAZIONE per omessa p motivi e di Vale Umberto PERSONA e Garofoli Anna PERSO presentazione di motivi non contestuali non sottoscritti cassazionista.

2) chiede dichiararsi inammissibili per genericità o infon ricorso dell' Avvocatura di Stato ORGANIZZAZIONE contro Melioli Giovanni PERSONA, Iannilli Marcello PERSONA, Raho Roberto PERSONA, Delle Chiaie Stefano PERSONA, Ballan Marco PERSONA, Tilgher Adriano PERSONA e

Dave

Ciao, come posso aiutarti?

User

Quali sono le motivazioni del ricorso per Cassazione eccepiti dal procuratore Squadrini rispetto alla sentenza d'appello per Musumeci e Belmonte?

documents > 56e861abead398e62409709d8291b21626daa64a3b36ef0386b4ad0befd53138
CONTRIBUTI ALLA VERITÀ:Sentenza del 16 maggio 1994

Relevant passages: giudicando in sed... 2. quanto a Musu... 5. conferma la sen... con sentenza 14.3....

Dave

Il ricorso per Cassazione presentato da Pietro Musumeci e Giuseppe Belmonte, e contestato dal procuratore Squadrini, mirava a contestare la decisione della Corte d'Assise d'Appello che confermava la loro condanna per calunnia. In particolare, Musumeci sosteneva che il giudice avrebbe dovuto riesaminare le prove raccolte, senza limitarsi ad aderire automaticamente alle valutazioni espresse dalle Sezioni Unite della Corte di Cassazione nella precedente sentenza.

Figure 1: The DAVE framework applied to documents referring to the Bologna massacre of August 2, 1980 [Bologna Massacre, 2024].

open source under the Apache-2.0 license. Documentation, source code, and demonstration videos are available on the project GitHub page[1].

2 Framework Main Functionalities

Knowledge-intensive domains feature extensive collections of complex documents where information is highly factual and deeply connected to real-world entities. Moreover, details about a single entity are often distributed across multiple documents within the corpus. Recognizing that entities are the cornerstone of effective document analysis in knowledge-intensive domains, DAVE proposes an entity-centric exploration approach that goes beyond the exploitation of entity mentions, e.g., as an output of a NER algorithm, supporting mentions linked to entity identifiers and clusters of mentions referring to the same entity.

DAVE provides users with a number of functionalities to address the outlined requirements, which are discussed with examples in legal documents exploration (Figure 1):

- **Search & Filter**: Users can retrieve documents by entering keywords, which is ideal when they have a clear idea of what they are looking for. Additionally, they can refine their search results by applying multiple filters. For instance, users can filter the corpus based on specific individuals and locations. Figure illustrates the *Search & Filter* functionalities (the blue section), which allow users to retrieve documents by keyword and refine results through entity-based filters. For example, a user reviewing documents from a trial might search for the keyword "appeal". To narrow the focus, they could use the filtering panel on the left, which organizes filters by entity type (e.g., person, location). By selecting "Licio Gelli" from the list of identified entities, the user can dynamically refine the results, isolating only those documents where his name appears, with both the result set and available filters updating accordingly.
- **Explore**: Users can browse the full list of entities identified in each document and navigate directly to sections where they are referenced. In the figure, after selecting a document in the *Search & Filter* interface, it can be explored in detail via the *Explore* interface (green section). The entire document is shown with highlighted entity mentions based on type, and a list of entities grouped by mention appears on the left. For example, in the figure, the entity "Valerio Fioravanti" is mentioned 4 times. Clicking any mention takes the user directly to the corresponding text span.
- **Conversational Question Answering (QA)**: Natural language queries are supported, allowing users to ask questions about entities and factual information across multiple documents. In the example (yellow section), the user asks the chatbot about the appeal grounds raised by Prosecutor Squadrini in the Musumeci and Belmonte case, and the DAVE chatbot responds with relevant document passages.
- **Knowledge Refinement**: Users can refine entity clusters, ensuring that corrections are reflected across all system functionalities. In the figure (red section), the user has identified that the entities "Corte d'Assise di Bologna" and "Corte di Assise di Bologna" with their respective mentions are actually equivalent, therefore is collapsing the two entity clusters into one.

DAVE also supports *search composability*, enabling users to combine Search, Filter, and Conversational Question Answering. By filtering results before querying the RAG-LLM

[1] https://github.com/unimib-datAI/DAVE

system, the chatbot works with a more focused document set, improving answer precision and relevance. Additionally, DAVE ensures data privacy through on-premise servers and access controls, crucial for sensitive domains.

By offering this suite of functionalities, DAVE enables precise control and in-depth understanding of large corpora, providing diverse exploration and search capabilities tailored to knowledge-intensive domains and specialized users. To this aim we implemented these functionalities by mixing the following techniques and technologies:

- **Entity-centric management**. Entity-level annotations are represented in the GATE format [Cunningham, 2002] and stored in a MongoDB database; the DB stores information about the annotations for every entity mentions, entity identifiers, and links between annotations and entity identifiers. The current prototype considers annotations resulting from pipelines that apply algorithms for NER, Named Entity Linking (with links to Wikipedia), NIL Prediction (to identify entities not represented in Wikipedia), and NIL Clustering (to cluster NIL entities and create identifiers for each cluster)[Pozzi *et al.*, 2023a; Pozzi *et al.*, 2023b; Bellandi *et al.*, 2024]. As a result, each entity mention can be linked to an entity identifier, either external (e.g. a Wikipedia URI), or internal (identifying a local cluster). Each cluster is associated with a default surface form used to display the entities in the interface.
- **Keyword and Faceted Search**: The *Search* engine uses keyword matching to efficiently retrieve relevant documents and serves as the foundation for several other functionalities. To support the *Search & Filter* and *Explore* functionalities, DAVE employs the well-established faceted search paradigm. This technique provides filtering facets based on entities, allowing users to refine their search results. By providing a structured way to narrow down results, faceted search enhances the user's ability to explore large corpora. Users can apply filters either across the entire corpus or within a subset of documents for a more granular exploration.
- **Human-in-the-Loop (HITL)**: *The Knowledge Consolidation* feature follows the HITL paradigm, ensuring continuous user involvement in refining the system. Users can correct and refine annotations and entity clusters, and these corrections are reflected across all system functionalities. This active participation helps the system improve over time, ensuring a pay-as-you-go consolidation of the background data as proposed for similar tasks [De Castilho *et al.*, 2024; De Paoli *et al.*, 2019; Cutrona *et al.*, 2019; Cruz *et al.*, 2016].
- **Retrieval Augmented generation (RAG)**: The Conversational QA functionality is powered by an LLM-based chatbot, implemented through the RAG paradigm. This enables users to ask fact-based queries about entities across multiple documents. While LLMs excel in natural language understanding and zero-shot learning, RAG ensures responses are grounded in retrieved documents, addressing concerns about hallucinations and limited knowledge, which is crucial in certain domains.

3 Applications, Main Contributions and Demonstration

DAVE has been used in prototypes for Italian projects in the legal domain [Batini *et al.*, 2024] with the goal of showing stakeholders AI-powered search functionalities, we have experimented DAVE in i) search on court decision in civil trials [Bellandi *et al.*, 2024], ii) criminal investigations and chat analysis [Pozzi *et al.*, 2025], and iii) analysis of the documentation about the Bologna massacre of August 2, 1980.

In relation with related work, we discuss below the three main novelties that we believe DAVE presents as a system:

- *Mixing entity-driven faceted search and conversational assistant*. Entity-driven faceted search is a mainstream technology in many knowledge-driven scenarios where documents are analyzed using entity extraction (EE) methods [Guo *et al.*, 2023; Hirsch *et al.*, 2021]. On the other hand, RAG systems on pre-filtered data have been studied, with approaches like agent-based filtering [Poliakov and Shvai, 2024], metadata-based filtering [Chang *et al.*, 2024], and natural language inference [Yoran *et al.*, 2023]. Our work combines faceted search with a RAG system, enabling dynamic, entity-driven document filtering.
- *Integrating interactive entity-driven knowledge consolidation in an information exploration interface*. Several platforms support text annotations, with Doccano [Nakayama *et al.*, 2018] being widely used; Very recent work has surveyed interactive approaches to improve annotations, minimize user effort, manage annotation teams, support pre-annotated data, and enable customizable task design [De Castilho *et al.*, 2024]. While we haven't fully integrated the advanced annotation quality methods from [Klie *et al.*, 2020], our application is the first to integrate incorporate interactive methods editing into an *exploratory search* interface, focusing on improving entity clustering, a key challenge in end-to-end EE pipelines.
- *Entity-centric RAG prototyping and grounding*. Frameworks like LangChain [Chase, 2022] facilitate rapid prototyping and configuration of RAG systems, while tools such as RAGAS [Es *et al.*, 2024] and RAGChecker [Ru *et al.*, 2024] allow for detailed evaluation through a wide range of metrics. However, existing applications do not support the prototyping of highly entity-centric LLM-RAG systems that enable direct analysis of the corpus and its entities to verify factual accuracy, making effective debugging more challenging.

These innovations arise from the need for domain experts to thoroughly analyze and explore documents annotated by entity extraction pipelines, using established, user-friendly search paradigms and allowing experts to improve the system by refining annotations. A first quantitative evaluation where DAVE outputs are compared to outputs of top-tier models is ongoing in the context of a Civil Appeal Proceedings use case [Agazzi *et al.*, 2024].

During the demonstration session, users are guided in exploring a document collection using DAVE's features.

Acknowledgments

The work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreements No. 101189771 (DataPACT) and No. 101070284 (enRichMyData), and the European Community – Next Generation EU under the Italian Ministry of Justice, and the Italian PRIN project Discount Quality for Responsible Data Science (202248FWFS). We also thank "Associazione 2 Agosto 1980" for the useful discussions and their support for using their data in our demo.

References

[Agazzi *et al.*, 2024] Ruben Agazzi, Carlo Batini, Matteo Palmonari, and Valentina Monica. Legal document query language: Conceptualizing linguistic commands for ai assistants in civi apeal proceedings. In *Proceedings of SEBD 2025)*, pages 35–50. CEUR-Ws, 2024.

[Batini *et al.*, 2024] Carlo Batini, Gaetano Santucci, Matteo Palmonari, Valerio Bellandi, Elisabetta Fersini, Fabio Zanzotto, Barbara Pernici, Giancarlo Vecchi, and Stefano Ronchi. Towards a semantic document management system for public administration. In *Proceedings of the Ital-IA Intelligenza Artificiale - Thematic Workshops co-located with the 4th CINI National Lab AIIS Conference on Artificial Intelligence (Ital-IA 2024), Naples, Italy, May 29-30, 2024*, volume 3762 of *CEUR Workshop Proceedings*, pages 396–401. CEUR-WS.org, 2024.

[Bellandi *et al.*, 2024] Valerio Bellandi, Christian Bernasconi, Fausto Lodi, Matteo Palmonari, Riccardo Pozzi, Marco Ripamonti, and Stefano Siccardi. An entity-centric approach to manage court judgments based on natural language processing. *Computer Law Security Review*, 52:105904, 2024.

[Bologna Massacre, 2024] Families Of The Victims Of The Bologna Massacre. 2 agosto 1980. Comune di Bologna, Italy, 2024.

[Chang *et al.*, 2024] Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, et al. Main-rag: Multi-agent filtering retrieval-augmented generation. *arXiv preprint arXiv:2501.00332*, 2024.

[Chase, 2022] Harrison Chase. Langchain, 2022.

[Cruz *et al.*, 2016] Isabel F Cruz, Matteo Palmonari, Francesco Loprete, Cosmin Stroe, and Aynaz Taheri. Quality-based model for effective and robust multi-user pay-as-you-go ontology matching. *Semantic Web*, 7(4):463–479, 2016.

[Cunningham, 2002] Hamish Cunningham. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of ACL 2002*, pages 168–175, 2002.

[Cutrona *et al.*, 2019] Vincenzo Cutrona, Michele Ciavotta, Flavio De Paoli, and Matteo Palmonari. Asia: A tool for assisted semantic interpretation and annotation of tabular data. In *Proceedings of the ISWC 2019 Satellite Tracks (Posters Demonstrations, Industry, and Outrageous Ideas)*, volume 2456, pages 209–212. CEUR-WS, 2019.

[De Castilho *et al.*, 2024] Richard Eckart De Castilho, Jan-Christoph Klie, and Iryna Gurevych. Integrating inception into larger annotation processes. In *Proceedings of EMNLP 2024: System Demonstrations*, pages 110–121, 2024.

[De Paoli *et al.*, 2019] Flavio De Paoli, Roberto Avogadro, Marco Ripamonti, and Matteo Palmonari. Interactive enrichment of tabular data with semtui. In *Proceedings of the ISWC 2024 Posters, Demos and Industry Tracks*, volume 2456, pages 209–212. CEUR-WS, 2019.

[Edge *et al.*, 2024] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. arxiv 2024. *arXiv preprint arXiv:2404.16130*, 2024.

[Es *et al.*, 2024] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of ACL 2024: System Demonstrations*, pages 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics.

[Guo *et al.*, 2023] Mengtian Guo, Zhilan Zhou, David Gotz, and Yue Wang. Grafs: Graphical faceted search system to support conceptual understanding in exploratory search. *ACM Trans. Interact. Intell. Syst.*, 13(2), May 2023.

[Hirsch *et al.*, 2021] Eran Hirsch, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan. iFacetSum: Coreference-based interactive faceted summarization for multi-document exploration. In *Proceedings of EMNLP 2021: System Demonstrations*, pages 283–297. Association for Computational Linguistics, November 2021.

[Klie *et al.*, 2020] Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In *Proceedings of ACL 2020*. Association for Computational Linguistics, July 2020.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of NeurIPS 2020*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[Li *et al.*, 2022] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2022.

[Liu *et al.*, 2024] Yaxi Liu, Chunxiu Qin, Yulong Wang, and XuBu Ma. Exploratory search in information systems: a systematic review. *The Electronic Library*, 42(2):308–339, 2024.

[Logan IV *et al.*, 2021] Robert L Logan IV, Andrew McCallum, Sameer Singh, and Dan Bikel. Benchmarking scalable methods for streaming cross document entity coreference. In *ACL-IJCNLP 2021*, volume 1, pages 4717–4731. ACL, 2021.

[Nakayama *et al.*, 2018] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text annotation tool for human, 2018. Software available from https://github.com/doccano/doccano.

[Poliakov and Shvai, 2024] Mykhailo Poliakov and Nadiya Shvai. Multi-meta-rag: Improving rag for multi-hop queries using database filtering with llm-extracted metadata. *arXiv preprint arXiv:2406.13213*, 2024.

[Pozzi *et al.*, 2023a] Riccardo Pozzi, Federico Moiraghi, Fausto Lodi, and Matteo Palmonari. Evaluation of incremental entity extraction with background knowledge and entity linking. In *11th International Joint Conference on Knowledge Graphs*, IJCKG 2022. ACM, 2023.

[Pozzi *et al.*, 2023b] Riccardo Pozzi, Riccardo Rubini, Christian Bernasconi, and Matteo Palmonari. Named entity recognition and linking for entity extraction from italian civil judgements. In *AIxIA 2023 - Advances in Artificial Intelligence - XXIInd International Conference of the Italian Association for Artificial Intelligence, AIxIA 2023, Rome, Italy, November 6-9, 2023, Proceedings*, volume 14318 of *Lecture Notes in Computer Science*, pages 187–201. Springer, 2023.

[Pozzi *et al.*, 2025] Riccardo Pozzi, Valentina Barbera, Renzo Alva Principe, Davide Giardini, Riccardo Rubini, and Matteo Palmonari. Combining knowledge graphs and nlp to analyze instant messaging data in criminal investigations. In *Web Information Systems Engineering – WISE 2024*, pages 427–442. Springer Nature Singapore, 2025.

[Ru *et al.*, 2024] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*, 2024.

[Sevgili *et al.*, 2022] Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570, 2022.

[Tunkelang, 2022] Daniel Tunkelang. *Faceted search*. Springer Nature, 2022.

[Worledge *et al.*, 2024] Theodora Worledge, Tatsunori Hashimoto, and Carlos Guestrin. The extractive-abstractive spectrum: Uncovering verifiability trade-offs in llm generations. *arXiv preprint arXiv:2411.17375*, 2024.

[Xu *et al.*, 2024] Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of SIGIR '24*, pages 2905–2909, 2024.

[Yoran *et al.*, 2023] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.