Luca Ragazzi

"Things don't happen by closing your eyes. Help yourself and you will be helped."

Personal

17/01/1996 Forlì luca.ragazzi9696@gmail.com

Research & Academy

Scholar Website LinkedIn GitHub

UniboNLP ORCID l.ragazzi@unibo.it

Bio

Luca Ragazzi, Ph.D. (2024), is a postdoctoral researcher at the Department of Computer Science and Engineering, University of Bologna. He holds both a bachelor's degree (2018) and a master's degree with honors (2020) from the same institution. His research focuses on Natural Language Processing, with particular emphasis on Large Language Models and Text Summarization/Generation in low-resource settings, especially in high social-impact domains such as law and biomedicine. Luca has co-authored around 20 peer-reviewed papers in leading international venues, including AAAI, ACL, EMNLP, ICLR, and Neurocomputing. His extensive reviewing experience spans over 65 submissions to high-impact AI and NLP conferences and journals, where he currently holds the Reviewing Editor role with Springer Nature and Area Chair for ACL ARR (Association for Computational Linguistics Rolling Review). He served as session chair at AAAI 2023 in Washington, D.C., and gained valuable research experience through an internship at the renowned French research institute EURECOM. In academia, he has delivered numerous invited lectures at both bachelor's and master's levels and co-supervised nearly 20 theses.

Education

2020 – 2024 **Ph.D.** in Computer Science and Engineering, University of Bologna, Cesena, Italy.
5/5 - Excellent
Supervisor: Gianluca Moro
Tutors: Marco Antonio Boschetti, Ilaria Bartolini
Topics: *Natural Language Processing – Text Generation – Text Summarization – Question Answering – Low Resource Regimes – Large Language Models*.
Thesis title: *Abstractive Long-Input Summarization in Low-Resource Regimes: Methods, Datasets and Benchmarks*.

2018 – 2020 **M.Sc.** in Computer Science and Engineering, University of Bologna, Cesena, Italy.
110L/110 - Magna cum Laude
Thesis title: *Abstractive Summarization on Legal Case Reports: New State-of-the-art Solutions with Transformer-based Language Models*.

2015 – 2018 **B.Sc.** in Computer Science and Engineering, University of Bologna, Cesena, Italy.
Thesis title: *Design and Development of an Offline Web Application with Angular Service Worker and IndexedDB*.

2010 – 2015 **Secondary High School Diploma** in Liceo Scientifico Statale Fulcieri Paulucci di Calboli, Forlì, Italy.

Work Experience

2024 – *now* **Post-Doctoral Researcher** in Computer Science and Engineering, University of Bologna, Cesena, Italy.
Topics: *Natural Language Processing – Text Generation – Text Summarization – Question Answering – Low Resource Regimes – Large Language Models – Benchmarking.*
Supervisor: Antonella Carbonaro
Research grant by DARE (Digital Lifelong Prevention).

2023 **Postgraduate Visiting Researcher** at EURECOM, Sophia Antipolis, France.
Supervisor: Paolo Papotti
Topic: *Automatic Data Generation for Computational Fact-Checking.*

2022 **Scientific and Technological Consultant** for L&G Solution, Foggia, Italy.
Topic: *Retrieval-based Italian Chatbot.*

2018 **Web Designer & Developer.** Curricular Internship at Librasoft, Forlì, Italy.
Topic: *Web Application Design and Development with React and Angular Frameworks.*

Languages

Italian Mother language.

English B2 level.

Skills

Coding	Python, Bash, Java, Scala, LaTeX, JSON
ML Libraries	PyTorch, HuggingFace, TensorFlow, NumPy, Pandas.
Web Dev	JavaScript, TypeScript, HTML, CSS, Angular, Vue.
Sw. & Tools	Docker, Git, Slurm.
Operating Systems	Mac OS X, Microsoft Windows, Linux.
Misc.	Leadership, Teamwork, Motivation, Problem Solving, Perseverance, Calmness.

Research Publications

Journal Articles

1. G. Frisoni, L. Ragazzi, D. Cohen, G. Moro, A. Carbonaro, and C. Sartori, "Abstractive Summarization through the Prism of Decoding Strategies," *Neural Networks*, 2025.
In natural language generation, abstractive summarization (AS) is advancing rapidly due to transformer-based language models (LMs). Although decoding strategies significantly influence generated summaries, their significance is often overlooked. Given the abundance of token selection heuristics and associated hyperparameters, the community needs guidance to make well-informed decisions based on the specific task and target metrics. To address this gap, we conduct a comparative assessment of the effectiveness and efficiency of decoding-time techniques for short, long, and multi-document AS. We explore over 3,500 combinations involving three widely used million-scale autoregressive encoder-decoder LMs, two billion-scale decoder-only LMs, six datasets, and nine decoding settings. Our findings highlight that optimized decoding choices can lead to substantial performance improvements. Alongside human evaluation, we quantitatively measure effects using ten automatic metrics, covering dimensions such as semantic similarity, factuality, compression, redundancy, and carbon footprint. To set the stage for differentiable selection and optimization of decoding options, we introduce Prism, a first-of-its-kind dataset that pairs AS gold input-output examples with our LM predictions across a diverse range of decoding options.

2 P. Italiani, G. Moro, and L. Ragazzi, "Clash-of-Leges: A Bilingual Dataset for Conflict Detection and Explanation in Statutory Law," *Expert Systems with Applications*, 2025.
Legal conflicts between statutes or constitutional articles present a significant challenge in maintaining consistency and coherence within legal systems. Addressing these conflicts requires extensive human expertise, making the process labor-intensive and time-consuming. In this paper, we introduce Clash-of-Leges, a novel multilingual dataset derived from rulings by the Constitutional Court of the Italian Republic, designed to aid the automation of conflict detection and explanation between legal articles. We identify three key tasks: Conflict Classification, which determines whether two legal articles are in conflict; Conflict Explanation Generation, which provides detailed explanations for identified conflicts; and Reference Retrieval, which sources relevant legal bases and precedents to substantiate interpretations. These tasks are intended to facilitate the development of AI models that can automatically identify and explain contradictions between legal provisions.

3 P. Italiani, G. Moro, and L. Ragazzi, "Enhancing Legal Question Answering with Data Generation and Knowledge Distillation from Large Language Models," *Artificial Intelligence and Law*, 2025.
Legal question answering (LQA) relies on supervised methods to automatically handle law-related queries. These solutions require a substantial amount of carefully annotated data for training, which makes the process very costly. Although large language models (LLMs) show promise in zero-shot QA, their computational demands limit their practical use, making specialized small language models (SLMs) more favorable. Furthermore, the growing interest in synthetic data generation has recently surged, spurred by the impressive generation capabilities of LLMs. This paper presents Ace-Attorney, an LLM distillation approach devised to develop LQA data and supervised models without human annotation. Given a textual prompt, a frozen LLM generates artificial examples that are used as knowledge to train a student SLM with an order of magnitude fewer parameters. Taking into account a realistic retrieval-based scenario to fetch the correct document for answer generation, we propose Selective Generative Paradigm, a novel approach designed to improve retrieval efficacy. Extensive experiments demonstrate the effectiveness and efficiency of distilled models on Syn-LeQA, our human-free synthetic dataset, and a public expert-annotated corpus. Notably, by using only a few dozen training samples, our best SLM achieves LLM-comparable performance with ≈1200% less CO_2 emissions.

4 P. Italiani, L. Ragazzi, and G. Moro, "Read Between the Tokens: Differentiable Text Pruning via Perturbed Top-k Selection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2025.
Transformer-based pretrained language models (PLMs) face scalability issues due to their computational expense, which increases with the length of the input sequence, and often struggle to maintain focus on relevant information. To mitigate this, we introduce PrunePert, a novel model featuring a learnable mechanism that identifies and removes uninformative tokens from the context. By doing so, our method not only addresses performance concerns but also enhances interpretability, offering valuable insights into the tokens utilized in the model's decision-making process. Specifically, our approach employs a differentiable perturbed top-k token selection module within the transformer layers to prune a user-defined percentage of tokens. It can be integrated with any downstream PLMs, allowing them to be trained end-to-end using backpropagation. We demonstrate the application of PrunePert in text summarization and classification tasks, utilizing both encoder-decoder PLMs and contemporary decoder-only large language models. Notably, our findings reveal that models equipped with PrunePert achieve up to 2x higher throughput and exhibit comparable performance in text summarization, while demonstrating superior performance in text classification tasks. Code is available at https://anonymous.4open.science/r/llm$_{p}runing - 6B58/$..

5 G. Moro, L. D. M. Magnani, and L. Ragazzi, "Legal Lay Summarization: Exploring Methods and Data Generation with Large Language Models," *Artificial Intelligence Review*, 2025.
This paper explores advancements in Natural Language Processing (NLP) for legal lay summarization by systematically analyzing existing methodologies, datasets, and research findings. We review current literature, highlighting key challenges such as data scarcity and the complexity of legal language. A primary contribution of this study is the development of LegalEase, a specialized dataset designed to improve model training for summarizing legal documents in layman's terms. Our findings demonstrate

that subdomain-specific datasets within the legal domain outperform general legal datasets in enhancing NLP model performance for generating accurate and comprehensible legal summaries. The insights and methodologies presented provide a foundation for future research in legal lay summarization.

6 L. Ragazzi, G. Moro, L. Valgimigli, and R. Fiorani, "Cross-Document Distillation via Graph-based Summarization of Extracted Essential Knowledge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2025. URL: https://ieeexplore.ieee.org/abstract/document/10740791.
Abstractive multi-document summarization aims to generate a comprehensive summary that encapsulates crucial content derived from multiple input documents. Despite the proficiency exhibited by language models in text summarization, challenges persist in capturing and aggregating salient information dispersed across a cluster of lengthy sources. To accommodate more input, existing solutions prioritize sparse attention mechanisms, relying on sequence truncation without incorporating graph-based modeling of multiple semantic units to locate essential facets. Furthermore, the limited availability of training examples adversely impacts performance, thereby compromising summarization quality in real-world few-shot scenarios. In this paper, we present G-Seek-2, a graph-enhanced approach designed to distill multiple topic-related documents by pinpointing and processing solely the pertinent information. We use a heterogeneous graph to model the input cluster, interconnecting various encoded entities via informative semantic edges. Then, a graph neural network locates the most salient sentences that are provided to a language model to generate the summary. We extensively evaluate G-Seek-2 across seven datasets spanning various domains—including news articles, lawsuits, government reports, and scientific texts—under few-shot settings with a limited training sample size of only 100 examples. The experimental findings demonstrate that our model consistently outperforms advanced summarization baselines, achieving improvements as measured by syntactic and semantic metrics.

7 L. Ragazzi, G. Moro, S. Guidi, and G. Frisoni, "LAWSUIT: a LArge expert-Written SUmmarization dataset of ITalian constitutional court verdicts," *Artificial Intelligence and Law*, 2024. URL: https://link.springer.com/article/10.1007/s10506-024-09414-w.
Large-scale public datasets are vital for driving the progress of abstractive summarization, especially in law, where documents have highly specialized jargon. However, the available resources are English-centered, limiting research advancements in other languages. This paper introduces LAWSUIT, a collection of 14K Italian legal verdicts with expert-authored abstractive maxims drawn from the Constitutional Court of the Italian Republic. LAWSUIT presents an arduous task with lengthy source texts and evenly distributed salient content. We offer extensive experiments with sequence-to-sequence and segmentation-based approaches, revealing that the latter achieve better results in full and few-shot settings. We openly release LAWSUIT to foster the development and automation of real-world legal applications.

8 G. Moro, N. Piscaglia, L. Ragazzi, and P. Italiani, "Multi-Language Transfer Learning for Low-Resource Legal Case Summarization," *Artificial Intelligence and Law*, pp. 1–29, 2023. DOI: 10.1007/s10506-023-09373-8.
Analyzing and evaluating legal case reports are labor-intensive tasks for judges and lawyers, who usually base their decisions on report abstracts, legal principles, and commonsense reasoning. Thus, summarizing legal documents is time-consuming and requires excellent human expertise. Moreover, public legal corpora of specific languages are almost unavailable. This paper proposes a transfer learning approach with extractive and abstractive techniques to cope with the lack of labeled legal summarization datasets, namely a low-resource scenario. In particular, we conducted extensive multi- and cross-language experiments. The proposed work outperforms the state-of-the-art results of extractive summarization on the Australian Legal Case Reports dataset and sets a new baseline for abstractive summarization. Finally, syntactic and semantic metrics assessments have been carried out to evaluate the accuracy and the factual consistency of the machine-generated legal summaries.

9 G. Moro and L. Ragazzi, "Align-Then-Abstract Representation Learning for Low-Resource

Summarization," *Neurocomputing*, vol. 548, p. 126 356, 2023. DOI: 10.1016/J.NEUCOM.2023.126356.
Generative transformer-based models have achieved state-of-the-art performance in text summarization. Nevertheless, they still struggle in real-world scenarios with long documents when trained in low-resource settings of a few dozen labeled training instances, namely in low-resource summarization (LRS). This paper bridges the gap by addressing two key research challenges when summarizing long documents, i.e., long-input processing and document representation, in one coherent model trained for LRS. Specifically, our novel align-then-abstract representation learning model (ATHENA) jointly trains a segmenter and a summarizer by maximizing the alignment between the chunk-target pairs in output from the text segmentation. Extensive experiments reveal that ATHENA outperforms the current state-of-the-art approaches in LRS on multiple long document summarization datasets from different domains.

10 G. Moro, L. Ragazzi, L. Valgimigli, G. Frisoni, C. Sartori, and G. Marfia, "Efficient Memory-Enhanced Transformer for Long-Document Summarization in Low-Resource Regimes," *Sensors*, vol. 23, no. 7, p. 3542, 2023. DOI: 10.3390/S23073542.
Long document summarization poses obstacles to current generative transformer-based models because of the broad context to process and understand. Indeed, detecting long-range dependencies is still challenging for today's state-of-the-art solutions, usually requiring model expansion at the cost of an unsustainable demand for computing and memory capacities. This paper introduces EMMA, a novel efficient memory-enhanced transformer-based architecture. By segmenting a lengthy input into multiple text fragments, our model stores and compares the current chunk with previous ones, gaining the capability to read and comprehend the entire context over the whole document with a fixed amount of GPU memory. This method enables the model to deal with theoretically infinitely long documents, using less than 18 and 13 GB of memory for training and inference, respectively. We conducted extensive performance analyses and demonstrate that EMMA achieved competitive results on two datasets of different domains while consuming significantly less GPU memory than competitors do, even in low-resource settings.

Conference Proceedings

1 A. Cocchieri, L. Ragazzi, G. Tagliavini, and G. Moro, "ReMedQA: Are We Done With Medical Multiple-Choice Benchmarks?" In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Rabat, Morocco: Association for Computational Linguistics, Mar. 2026.
Medical multiple-choice question answering (MCQA) benchmarks show that models achieve near-human accuracy, with some benchmarks approaching saturation—leading to claims of clinical readiness. Yet a single accuracy score is a poor proxy for competence: models that change answers under minor input perturbations cannot be considered reliable. We argue that reliability underpins accuracy—only consistent predictions make correctness meaningful. We release ReMedQA, a new benchmark that augments three standard medical MCQA datasets with open-ended answers and systematically perturbed options. Building on this, we introduce ReAcc and ReCon, two reliability metrics: ReAcc measures the proportion of questions answered correctly across all variations, while ReCon measures the proportion answered consistently regardless of correctness. Our evaluation shows that high MCQA accuracy masks low reliability: models remain sensitive to format and perturbation changes, and domain specialization offers no robustness gain. MCQA underestimates smaller models while inflating large ones that exploit structural cues—with some exceeding 50% accuracy even when the original questions are hidden. This shows that, despite near-saturated accuracy, we are not yet done with medical MCQA benchmarks.

2 I. Paolo, G.-G. David, R. Luca, M. Gianluca, and R. Paolo, "MemeWeaver: Inter-Meme Graph Reasoning for Sexism and Misogyny Detection," in *Findings of the Association for Computational Linguistics: EACL 2026*, Rabat, Morocco: Association for Computational Linguistics, Mar. 2026.

Women are twice as likely as men to face online harassment due to their gender. Despite recent advances in multimodal content moderation, most approaches still overlook the social dynamics behind this phenomenon, where perpetrators reinforce prejudices and group identity within like-minded communities. Graph-based methods offer a promising way to capture such interactions, yet existing solutions remain limited by heuristic graph construction, shallow modality fusion, and instance-level reasoning. In this work, we present MemeWeaver, an end-to-end trainable multimodal framework for detecting sexism and misogyny through a novel inter-meme graph reasoning mechanism. We systematically evaluate multiple visual-textual fusion strategies and show that our approach consistently outperforms state-of-the-art baselines on the MAMI and EXIST benchmarks, while achieving faster training convergence. Further analyses reveal that the learned graph structure captures semantically meaningful patterns, offering valuable insights into the relational nature of online hate.

3 A. Cocchieri, L. Ragazzi, P. Italiani, G. Tagliavini, and G. Moro, "What do you call a dog that is incontrovertibly true? Dogma: Testing LLM Generalization through Humor," in *Proceedings of the 63th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27-August 1, 2025*, Association for Computational Linguistics, 2025, pp. 1–16. URL: https://doi.org/10.18653/v1/2022.acl-long.15.

Humor, requiring creativity and contextual understanding, is a hallmark of human intelligence, showcasing adaptability across linguistic scenarios. While recent advances in large language models (LLMs) demonstrate strong reasoning on various benchmarks, it remains unclear whether they truly adapt to new tasks like humans (i.e., generalize) or merely replicate memorized content. To explore this, we introduce Phunny, a new humor-based question-answering benchmark designed to assess LLMs' reasoning through carefully crafted puns. Our dataset is manually curated to ensure novelty and minimize data contamination, providing a robust evaluation of LLMs' linguistic comprehension. Experiments on pun comprehension, resolution, and generation reveal that most LLMs struggle with generalization, even on simple tasks, consistently underperforming the human baseline. Additionally, our detailed error analysis provides valuable insights to guide future research.

4 A. Cocchieri, L. Ragazzi, G. Tagliavini, A. Carbonaro, L. Tordi, and G. Moro, "Can Large Language Models Win the International Mathematical Games?" In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, Nov. 2025.

Recent advances in large language models (LLMs) have demonstrated strong mathematical reasoning abilities, even in visual contexts, with some models surpassing human performance on existing benchmarks. However, these benchmarks lack structured age categorization, clearly defined skill requirements, and—crucially—were not designed to assess human performance in international competitions. To address these limitations, we introduce MathGames, a new benchmark of 2,183 high-quality mathematical problems (both text-only and multimodal) in an open-ended format, sourced from an international mathematical games championships. Spanning seven age groups and a skill-based taxonomy, MathGames enables a structured evaluation of LLMs' mathematical and logical reasoning abilities. Our experiments reveal a substantial gap between state-of-the-art LLMs and human participants—even 11-year-olds consistently outperform some of the strongest models—highlighting the need for advancements. Further, our detailed error analysis offers valuable insights to guide future research. The data is publicly available at https://github.com/disi-unibo-nlp/math-games.

5 S. Fantazzini, G. Frisoni, G. Moro, L. Ragazzi, M. Ciccioni, and C. Sartori, "Magic Mirror on the Wall, Which is the Fairest Prompt of All? A Survey on Automatic Prompt Learning," in *ECAI 2025 - 28th European Conference on Artificial Intelligence, October 25 - 30, 2025, Bologna, Italy*, ser. Frontiers in Artificial Intelligence and Applications, IOS Press, 2025.

Prompts direct the behavior of a model by conditioning its outputs on carefully designed instructions and examples, similar to setting the trajectory of an arrow before release. More broadly, prompt learning is the research area that aims to solve downstream tasks by directly leveraging the knowledge acquired by language models at pre-training time, removing the need for expensive fine-tuning stages with potentially different objective functions. While manual prompt engineering has enabled both small and large language models to achieve superhuman performance on numerous benchmarks, it remains a

labor-intensive and suboptimal process. Recently, the field has shifted towards automating the search for prompts that effectively elicit the desired model responses. This survey presents the first systematic review of prompt learning for pre-trained language models operating on text inputs, with a particular focus on automatic methods. We critically analyze existing publications and organize them into a novel taxonomy, describing key aspects for practical usage. We finally discuss promising directions for future research. Our curated repository of annotated papers, continuously updated, is available at https://github.com/disi-unibo-nlp/awesome-prompt-learning.

6 L. Molfetta, A. Cocchieri, S. Fantazzini, G. Frisoni, L. Ragazzi, and G. Moro, "FEAST: Retrieval-Augmented Multi-Hierarchical Food Classification for the FoodEx2 System," in *ECAI 2025 - 28th European Conference on Artificial Intelligence, October 25 - 30, 2025, Bologna, Italy*, ser. Frontiers in Artificial Intelligence and Applications, IOS Press, 2025.
Hierarchical text classification (HTC) and extreme multi-label classification (XML) tasks face compounded challenges from complex label interdependencies, data sparsity, and extreme output dimensions. These challenges are exemplified in the European Food Safety Authority's FoodEx2 system—a standardized food classification framework essential for food consumption monitoring and contaminant exposure assessment across Europe. FoodEx2 coding transforms natural language food descriptions into a set of codes from multiple standardized hierarchies, but faces implementation barriers due to its complex structure. Given a food description (e.g., 'organic yogurt'), the system identifies its base term ('yogurt'), all the applicable facet categories (e.g., 'production method'), and then, every relevant facet descriptors to each category (e.g., 'organic production'). While existing approaches perform adequately on well-balanced and semantically dense hierarchies, no work has been applied on the practical constraints imposed by the FoodEx2 system. The limited literature addressing such real-world scenarios further compounds these challenges. We propose FEAST (Food Embedding And Semantic Taxonomy), a novel retrieval-augmented framework that decomposes the FoodEx2 classification challenge into a three-stage approach: (1) base term identification, (2) multi-label facet prediction, and (3) facet descriptor assignment. By leveraging the system's hierarchical structure to guide training and performing deep metric learning, FEAST learns discriminative embeddings that mitigate data sparsity and improve generalization on rare and fine-grained labels. Evaluated on the multilingual FoodEx2 benchmark, FEAST outperforms the prior European's CNN baseline F1 scores by 12—38% on rare classes.

7 J.-F. Bussotti, L. Ragazzi, G. Frisoni, G. Moro, and P. Papotti, "Unknown Claims: Generation of Fact-Checking Training Examples from Unstructured and Structured Data," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12 105–12 122. DOI: 10.18653/v1/2024.emnlp-main.675.
Computational fact-checking (FC) relies on supervised models to verify claims based on given evidence, requiring a resource-intensive process to annotate large volumes of training data. We introduce Unown, a novel framework that generates training instances for FC systems automatically using both textual and tabular content. Unown selects relevant evidence and generates supporting and refuting claims with advanced negation artifacts. Designed to be flexible, Unown accommodates various strategies for evidence selection and claim generation, offering unparalleled adaptability. We comprehensively evaluate Unown on both text-only and table+text benchmarks, including Feverous, SciFact, and MMFC, a new multi-modal FC dataset. Our results prove that Unown examples are of comparable quality to expert-labeled data, even enabling models to achieve up to 5% higher accuracy. The code, data, and models are available at https://github.com/disi-unibo-nlp/unown.

8 G. Moro, L. Ragazzi, L. Valgimigli, F. Vincenzi, and D. Freddi, "Revelio: Interpretable Long-Form Question Answering," in *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR, Vienna, Austria, May 11, 2024*, OpenReview.net, 2024. URL: https://openreview.net/forum?id=fyvEJXsaQf.
The black-box architecture of pretrained language models (PLMs) hinders the interpretability of lengthy responses in long-form question answering (LFQA). Prior studies use knowledge graphs (KGs) to enhance output transparency, but mostly focus on non-generative or short-form QA. We present

Revelio, a new layer that maps PLM's inner working onto a KG walk. Tests on two LFQA datasets show that Revelio supports PLM-generated answers with reasoning paths presented as rationales while retaining performance and time akin to their vanilla counterparts.

9 L. Ragazzi, P. Italiani, G. Moro, and M. Panni, "What Are You Token About? Differentiable Perturbed Top-k Token Selection for Scientific Document Summarization," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds., Association for Computational Linguistics, 2024, pp. 9427–9440. doi: 10.18653/V1/2024.FINDINGS-ACL.561.
Scientific document summarization aims to condense complex and long articles in both technical and plain-language terms to facilitate the accessibility and dissemination of scientific findings. Existing datasets suffer from a deficiency in source heterogeneity, as their data predominantly stem from a single common resource, hindering effective model training and generalizability. First, we introduce SciLay, a novel dataset that includes documents from multiple natural science journals with expert-authored technical and lay summaries. Second, we propose PrunePert, a new transformer-based model that incorporates a differentiable perturbed top-k encoder layer to prune irrelevant tokens in end-to-end learning. Experimental results show that our model achieves a nearly 2x speed-up compared to a state-of-the-art linear transformer, remaining comparable in effectiveness. Additional examinations underscore the importance of employing a training dataset that includes different sources to enhance the generalizability of the models. Code is available at https://github.com/disi-unibo-nlp/sci-lay.

10 G. Moro, L. Ragazzi, and L. Valgimigli, "Carburacy: Summarization Models Tuning and Comparison in Eco-Sustainable Regimes with a Novel Carbon-Aware Accuracy," in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, B. Williams, Y. Chen, and J. Neville, Eds., AAAI Press, 2023, pp. 14 417–14 425. doi: 10.1609/AAAI.V37I12.26686.
Generative transformer-based models have reached cutting-edge performance in long document summarization. Nevertheless, this task is witnessing a paradigm shift in developing ever-increasingly computationally-hungry solutions, focusing on effectiveness while ignoring the economic, environmental, and social costs of yielding such results. Accordingly, such extensive resources impact climate change and raise barriers to small and medium organizations distinguished by low-resource regimes of hardware and data. As a result, this unsustainable trend has lifted many concerns in the community, which directs the primary efforts on the proposal of tools to monitor models' energy costs. Despite their importance, no evaluation measure considering models' eco-sustainability exists yet. In this work, we propose Carburacy, the first carbon-aware accuracy measure that captures both model effectiveness and eco-sustainability. We perform a comprehensive benchmark for long document summarization, comparing multiple state-of-the-art quadratic and linear transformers on several datasets under eco-sustainable regimes. Finally, thanks to Carburacy, we found optimal combinations of hyperparameters that let models be competitive in effectiveness with significantly lower costs.

11 G. Moro, L. Ragazzi, and L. Valgimigli, "Graph-Based Abstractive Summarization of Extracted Essential Knowledge for Low-Resource Scenarios," in *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, and R. Radulescu, Eds., ser. Frontiers in Artificial Intelligence and Applications, vol. 372, IOS Press, 2023, pp. 1747–1754. doi: 10.3233/FAIA230460.
Although current summarization models can process increasingly long text sequences, they still struggle to capture salient related information spread across the lengthy size of inputs with few labeled training instances. Today's research still relies on standard input truncation without considering graph-based modeling of multiple semantic units to summarize only crucial facets. This paper proposes G-Seek, a graph-based summarization of extracted essential knowledge. By representing the long source with a heterogeneous graph, our method extracts and provides salient sentences to an abstractive summarization model to generate the summary. Experimental results in low-resource

scenarios, distinguished by data scarcity, reveal that G-Seek consistently improves both the long- and multi-document summarization performance and accuracy across several datasets.

12 G. Moro, L. Ragazzi, L. Valgimigli, and L. Molfetta, "Retrieve-and-Rank End-to-End Summarization of Biomedical Studies," in *Similarity Search and Applications - 16th International Conference, SISAP 2023, A Coruña, Spain, October 9-11, 2023, Proceedings*, O. Pedreira and V. Estivill-Castro, Eds., ser. Lecture Notes in Computer Science, vol. 14289, Springer, 2023, pp. 64–78. DOI: 10.1007/978-3-031-46994-7_6.
An arduous biomedical task involves condensing evidence derived from multiple interrelated studies, given a context as input, to generate reviews or provide answers autonomously. We named this task context-aware multi-document summarization (CA-MDS). Existing state-of-the-art (SOTA) solutions require truncation of the input due to the high memory demands, resulting in the loss of meaningful content. To address this issue effectively, we propose a novel approach called Ramses, which employs a retrieve-and-rank technique for end-to-end summarization. The model acquires the ability to (i) index each document by modeling its semantic features, (ii) retrieve the most relevant ones, and (iii) generate a summary via token probability marginalization. To facilitate the evaluation, we introduce a new dataset, FAQsumC19, which includes the synthesizing of multiple supporting papers to answer questions related to Covid-19. Our experimental findings demonstrate that Ramses achieves notably superior ROUGE scores compared to state-of-the-art methodologies, including the establishment of a new SOTA for the generation of systematic literature reviews using MS2. Quality observation through human evaluation indicates that our model produces more informative responses than previous leading approaches.

13 G. Moro and L. Ragazzi, "Semantic Self-Segmentation for Abstractive Summarization of Long Documents in Low-Resource Regimes," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 11 085–11 093. DOI: 10.1609/AAAI.V36I10.21357.
The quadratic memory complexity of transformers prevents long document summarization in low computational resource scenarios. State-of-the-art models need to apply input truncation, thus discarding and ignoring potential summary-relevant contents, leading to a performance drop. Furthermore, this loss is generally destructive for semantic text analytics in high-impact domains such as the legal one. In this paper, we propose a novel semantic self-segmentation (Se3) approach for long document summarization to address the critical problems of low-resource regimes, namely to process inputs longer than the GPU memory capacity and produce accurate summaries despite the availability of only a few dozens of training instances. Se3 segments a long input into semantically coherent chunks, allowing transformers to summarize very long documents without truncation by summarizing each chunk and concatenating the results. Experimental outcomes show the approach significantly improves the performance of abstractive summarization transformers, even with just a dozen of labeled data, achieving new state-of-the-art results on two legal datasets of different domains and contents. Finally, we report ablation studies to evaluate each contribution of the components of our method to the performance gain.

14 G. Moro, L. Ragazzi, L. Valgimigli, and D. Freddi, "Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Association for Computational Linguistics, 2022, pp. 180–189. DOI: 10.18653/V1/2022.ACL-LONG.15.
Although current state-of-the-art Transformer-based solutions succeeded in a wide range for single-document NLP tasks, they still struggle to address multi-input tasks such as multi-document summarization. Many solutions truncate the inputs, thus ignoring potential summary-relevant contents, which is unacceptable in the medical domain where each information can be vital. Others leverage linear model approximations to apply multi-input concatenation, worsening the results because all information is considered, even if it is conflicting or noisy with respect to a shared background. Despite the importance and social impact of medicine, there are no ad-hoc solutions for multi-document summarization. For this reason, we propose a novel discriminative marginalized probabilistic method (Damen) trained to discriminate critical information from a cluster of

topic-related medical documents and generate a multi-document summary via token probability marginalization. Results prove we outperform the previous state-of-the-art on a biomedical dataset for multi-document summarization of systematic literature reviews. Moreover, we perform extensive ablation studies to motivate the design choices and prove the importance of each module of our method.

Scientific Activities

Participation in Research Groups

2020 - *now* — **UniboNLP Research Group**. Since my M.S. degree, I've been working with the UniBoNLP group at the University of Bologna, led by Prof. Gianluca Moro. The team–comprising Post Docs, Ph.D. students, and faculty–focuses on cutting-edge deep learning for NLP, especially in impactful areas like medicine and law, with over 50 publications in the last 5 years. We explore trends such as LLMs, GNNs, neuro-symbolic AI, prompt learning, and explainability.

Certifications

2025 — **Finetuning Large Language Models**. Awarded by DeepLearning.AI. Certificate.

AI Agents Fundamentals. Awarded by Hugging Face. Certificate.

2022 — **NLP Specialization**. Awarded by DeepLearning.AI. Certificate.

Speaker (Paper Presentation)

2026 — **The 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026, Rabat, Morocco, In-person**; *"Are We Done With Medical Multiple-Choice Benchmarks?*.

2025 — **The 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, In-person**; *"What do you call a dog that is incontrovertibly true? Dogma": Testing LLM Generalization through Humor*.

2024 — **Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, Florida, In-person**; *Unknown Claims: Generation of Fact-Checking Training Examples from Unstructured and Structured Data*.

2023 — **Association for the Advancement of Artificial Intelligence, AAAI 2023, Washington DC, USA, In-person**; *Carburacy: Summarization Models Tuning and Comparison in Eco-Sustainable Regimes with a Novel Carbon-Aware Accuracy*.

2022 — **Association for the Advancement of Artificial Intelligence, AAAI 2022, Virtual**; *Semantic Self-Segmentation for Abstractive Summarization of Long Documents in Low-Resource Regimes*.

Reviewing for Journals

2026 — **Multimedia Tools and Applications**; SJR Class: Q1.

Engineering Applications of Artificial Intelligence; SJR Class: Q1.

Neurocomputing; SJR Class: Q1.

2025 — **IEEE/ACM Transactions on Audio, Speech and Language Processing**; SJR Class: Q1.

Knowledge and Information Systems; SJR Class: Q2.

Artificial Intelligence and Law; SJR Class: Q1; 2 manuscripts.

Discover Computing; SJR Class: Q2.

IEEE Transactions on Knowledge and Data Engineering; SJR Class: Q1.

Expert Systems with Applications; SJR Class: Q1; 2 manuscripts.

Scientific Activities (continued)

- **Regenerative Biomaterials**; SJR Class: Q1.
- **Machine Learning**; SJR Class: Q1; 3 manuscripts.
- **Language Resources and Evaluation**; SJR Class: Q1.
- **Scientific Reports**; SJR Class: Q1; 4 manuscripts.
- **PeerJ Computer Science**; SJR Class: Q1.
- **International Journal of Data Science and Analytics**; SJR Class: Q2; 2 manuscripts.
- **Results in Engineering**; SJR Class: Q1.

2024
- **Computers, Materials & Continua**; SJR Class: Q2.
- **Natural Language Processing Journal**.
- **Semantic Web Journal**; SJR Class: Q2.
- **Artificial Intelligence and Law**; SJR Class: Q1.
- **Neurocomputing**; SJR Class: Q1.

2023
- **IEEE Transactions on Systems, Man and Cybernetics: Systems**; SJR Class: Q1.

2020
- **Artificial Intelligence and Law**; SJR Class: Q1.

Reviewing for Conferences

2026
- **ACL ARR 2026 January (Association for Computational Linguistics, ACL 2026)**; CORE Rank: A*; 4 manuscripts.

2025
- **ACL ARR 2025 October (European Chapter of the Association for Computational Linguistics, EACL 2026)**; CORE Rank: A; 9 manuscripts.
- **International Conference on Learning Representations, ICLR 2026**; CORE Rank: A*; 5 manuscripts.
- **ACL ARR 2025 July**; 2 manuscripts.
- **ACL ARR 2025 May (Empirical Methods on Natural Language Processing, EMNLP 2025)**; CORE Rank: A*; 6 manuscripts.
- **ACL ARR 2025 February (Association for Computational Linguistics, ACL 2025)**; CORE Rank: A*; 5 manuscripts.
- **Extended Semantic Web Conference, ESWC 2025**; CORE Rank: A; 3 manuscripts.

2024
- **International Conference on Learning Representations, ICLR 2025**; CORE Rank: A*; 3 manuscripts.
- **International Semantic Web Conference, ISWC 2024**; CORE Rank: A*.
- **Extended Semantic Web Conference, ESWC 2024**; CORE Rank: A.

2023
- **Association for the Advancement of Artificial Intelligence, AAAI 2024**; CORE Rank: A*.
- **ACM Symposium on Applied Computing, ACM SAC 2024**; CORE Rank: A.
- **ACL ARR 2023 October (Empirical Methods on Natural Language Processing, EMNLP 2023)**; CORE Rank: A*; 4 manuscripts.
- **Neural Information Processing Systems, NeurIPS 2023**; CORE Rank: A*.

2022
- **Association for the Advancement of Artificial Intelligence, AAAI 2023**; CORE Rank: A*.

Chair

2025
- **Area Chair: ACL ARR 2025 October (European Chapter of the Association for Computational Linguistics, EACL 2026)**; CORE Rank: A.

Scientific Activities (continued)

2023 **Session Chair: Association for the Advancement of Artificial Intelligence, AAAI 2023**, Washington DC, USA; CORE Rank: A*.

Teaching

Invited Talks

2024 *Retrieval-based Chatbot*. "Data Intensive Applications" B.Sc. course, Computer Science and Engineering, University of Bologna, April 3, 2024.

2023 *Automatic Text Summarization: from Theory to Practice*. "Data Mining, Text Mining and Big Data Analytics" M.Sc. course, Artificial Intelligence, University of Bologna, December 15, 2023.

Long Document Summarization in Low-Resource Regimes. "Data Mining, Text Mining and Big Data Analytics" M.Sc. course, Artificial Intelligence, University of Bologna, October 18, 2023.

Retrieval-based Italian Chatbot. "Data Intensive Applications" B.Sc. course, Computer Science and Engineering, University of Bologna, June 8, 2023.

Long and Multi-Document Abstractive Summarization in Low-Resource Regimes. "EURECOM", March 21, 2023, Sophia Antipolis, France.

2022 *Long and Multi-Document Abstractive Summarization in Low-Resource Regimes*. "Data Mining" M.Sc. course, Computer Science and Engineering, University of Bologna, December 1, 2022.

2021 *Long Document Summarization in Low-resource Regimes with Applications in the Legal Domain*. "Data Mining" M.Sc. course, Computer Science and Engineering, University of Bologna, December 16, 2021.

2020 *Natural Language Processing for Automatic Text Summarization: an Overview*. "Data Mining" M.Sc. course, Computer Science and Engineering, University of Bologna, December 12, 2020.

Thesis with Co-Supervisor Role

2025 *A Systematic Review on Automatic Prompt Learning*. Candidate: Mario Ciccioni, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Giacomo Frisoni. B.Sc. in Computer Science and Engineering, University of Bologna.

Legal Lay Summarization: Exploring Techniques and Introducing the LegalEase Dataset. Candidate: Leonardo David Matteo Magnani, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*. B.Sc. in Computer Science and Engineering, University of Bologna.

Tecniche di Machine learning per la Gestione e il Monitoraggio delle Emissioni Odorigene negli Allevamenti Avicoli. Candidate: Francesco Filippini, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*. B.Sc. in Computer Science and Engineering, University of Bologna.

Valutazione del Quoziente Intellettivo di Large Language Model Multimodali. Candidate: Eduard Toni Alexandru, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Giacomo Frisoni. B.Sc. in Computer Science and Engineering, University of Bologna.

2024 *Agriveritas: Chatbot Generativo per il Supporto Normativo allo Sviluppo dell'Agricoltura Sostenibile*. Candidate: Jacopo Pesaresi, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Lorenzo Molfetta. B.Sc. in Computer Science and Engineering, University of Bologna.

Benchmarking e Prompt Tuning di Large Language Model per la Generazione di Codice. Candidate: Luca Bighini, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Giacomo Frisoni. B.Sc. in Computer Science and Engineering, University of Bologna.

Teaching (continued)

2023
- *Neural Self-Supervised Information Retrieval: An Efficient and Effective Solution in Large Document Corpora*. Candidate: Samuele Marino, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Cristiano Casadei, Lorenzo Valgimigli. December, 2021. M.Sc. in Artificial Intelligence, University of Bologna.
- *Explaining Generative Model for Long-form Question Answering with Reasoning Graph*. Candidate: Fabian Vincenzi, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Lorenzo Valgimigli. M.Sc. in Artificial Intelligence, University of Bologna.
- *Summarization Astrattiva di Lunghi Articoli Scientifici mediante Estrazione di Frammenti Rilevanti*. Candidate: Filippo Di Pietro, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Paolo Italiani. B.Sc. in Computer Science and Engineering, University of Bologna.
- *Survey on Few-Shot Summarization*. Candidate: Emanuele Artegiani, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Giacomo Frisoni. B.Sc. in Computer Science and Engineering, University of Bologna.
- *Graph Neural Network Benchmark per la Selezione di Contenuto Rilevante nella Low-Resource Summarization*. Candidate: Riccardo Fiorani, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Lorenzo Valgimigli. B.Sc. in Computer Science and Engineering, University of Bologna.
- *Sci-Lay: Un Nuovo Dataset per Long Document Summarization Scientifica e Divulgativa di Studi Biomedici*. Candidate: Mattia Panni, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Paolo Italiani, Giacomo Frisoni. B.Sc. in Computer Science and Engineering, University of Bologna.
- *Generazione di Riassunti Fattuali Mediante Parsing Semantico*. Candidate: Luca Grandi, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*, Giacomo Frisoni. B.Sc. in Computer Science and Engineering, University of Bologna.

2022
- *Sviluppo di Metodi di Soft Labeling per la Multi-Document Summarization in Ambito Legale*. Candidate: Luca Rubboli, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*. B.Sc. in Computer Science and Engineering, University of Bologna.
- *Sviluppo di Retrieval-based Chatbot per l'Italiano con Transformer*. Candidate: Luca Morlino, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*. B.Sc. in Computer Science and Engineering, University of Bologna.
- *LAWSU-IT: Un Nuovo Dataset Giudiziario Italiano per Long Document Summarization con Baseline Estrattive e Astrattive*. Candidate: Stefano Guidi, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*. B.Sc. in Computer Science and Engineering, University of Bologna.

2021
- *Sintesi Generativa Multi-documento con Discriminazione della Rilevanza Mediante Probabilità Marginale: Una soluzione Neurale End-to-End per la Letteratura Medica*. Candidate: Davide Freddi, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*. B.Sc. in Computer Science and Engineering, University of Bologna.
- *Abstractive Long Document Summarization: Studio e Sperimentazione di Modelli Generativi Retrieval-Augmented*. Candidate: Veronika Folin, Supervisor: Gianluca Moro, Co-supervisor: *Luca Ragazzi*. B.Sc. in Computer Science and Engineering, University of Bologna.

Projects

Machine Learning

2020
- Design and implementation of a deep learning project on time series forecasting.
- Design and implementation of a text mining project on article knowledge discovery.

2018
- Design and implementation of a machine learning project on gold market trend.

Projects (continued)

Data Science

2020
- Design and implementation of a big data project on accident severity analysis.
- Design and implementation of a semantic web project on the expansion and modeling of an ontology for road management by an Ego vehicle.

Web Development

2019
- Design and implementation of the AlmaNotes web application.

2018
- Design and implementation of the Jestergest web app for the bachelor thesis.

Software Engineering

2020
- Design of a project management project on the development of a basketball application.

2019
- Design and implementation of concurrent and distributed programming projects.
- Design and implementation of a new compiler.

2018
- Design and implementation of a high-performance computing project.
- Design and implementation of embedded systems and internet of things projects.

2017
- Design and implementation of the GeoQuiz geography quiz game.
- Design and implementation of a database project.

Extracurricular

Sports Activities

2026
- **Bocce**, B series.

2025
- **Bocce**, C series.

2024
- **Bocce**, D series.

2004 – 2024
- **Basketball**. Shooting guard. Website

2018 – 2020
- **Scuba Diving**. Advanced level (3 patents).

Sports Achievements

2024
- **National Academic Bocce Championship First Place**. First position for the bocce's national academic championship, September 2024, Civitanova Marche, Italy.

2023
- **National Academic Bocce Championship First Place**. First position out of 42 participants for the bocce's national academic championship, September 2023, Padova, Italy.