Research Statement

Federico Ruggeri

Vision

The aim of my research is to devise Natural Language Processing (NLP) systems that **learn to generate, distill, and use knowledge from unstructured text**. Driven by an intrinsic desire to acquire new knowledge [Aristotle and Ross, 1933], humans have consistently developed tools for this purpose. Machine learning systems are no exception. These systems should be designed to provide new knowledge, which, accordingly, represents a valuable component for their evaluation and trustworthiness. In NLP, this requires the development of methods with the capability of **understanding** text to **distill** and **organize** novel structured knowledge. Moreover, these capabilities should be designed to guarantee desired properties of autonomous support systems, such as model transparency, efficiency, robustness to text representations, adaptability to different contexts, and computational scalability.

Concretely, we require a paradigm shift in how machine learning problems are formulated, going from the traditional approach of *"Given input $\mathcal{X}$, provide output $\mathcal{Y}$"*, where the focus is on $\mathcal{Y}$, to *"Given input $\mathcal{X}$ and knowledge $\mathcal{K}$, provide output $\mathcal{Y}$ and update $\mathcal{K}$"*, where the focus is **also** on $\mathcal{K}$. Here, the knowledge $\mathcal{K}$ refers to any information bit that contributes to the general human understanding of a given problem. Notably, this includes information about the task, the domain, and the model of interest. This paradigm shift implies that when we design a system for an NLP problem, our purpose and measurement of its success should not be limited to model accuracy. Instead, we should also account for the **reason** behind that accuracy. This translates to a model that not only explains its predictions, but also aggregates individual explanations to summarize its decision-making process into more general criteria. Altogether, these capabilities constitute important desiderata for several real-world applications where extracted knowledge represents the basis for understanding the underlying factors like human sociodynamics, dialogical interactions, and reasoning. Notable examples include Argument Mining, where understanding how humans convey and relate opinions is paramount [Lawrence and Reed, 2019], Hate Speech Detection, where knowledge about the cultural domain of content is necessary to discriminate hateful content [Yu et al., 2022], and Legal Analytics, where the decision-making process of legal experts is inherently grounded on factual knowledge [Xu et al., 2020].

During my research activity, I have taken several steps towards this ultimate goal of **learning with knowledge**, with applications in Argument Mining and Legal Analytics. I will continue these efforts by focusing on two main directions:

- **Unstructured Knowledge Integration.** The capability of models to leverage a large amount of unstructured textual knowledge to address specific problems.
- **Structured Knowledge Extraction from Text.** The capability of models to extract structured knowledge from raw text.

Unstructured Knowledge Integration

Introduction. Structured knowledge integration into machine learning models aims to overcome the shortcomings of purely data-driven approaches [von Rueden et al., 2023]. In particular, these include limited generalization capabilities when training data is scarce, lack of transparency, bias,

and lack of learning critical regulatory factors like fair behavior, security guidelines, and coherence concerning natural laws, which are often dictated by the given context. The context also affects how structured knowledge is represented and integrated into models. For instance, in NLP, popular representations are knowledge graphs, logic rules, and linguistic dependencies. In contrast, in other settings like physics or robotics, knowledge is often represented as mathematical equations, simulations, or human feedback. Albeit a valuable source of information to regulate the learning process of autonomous systems, in many scenarios, structured knowledge or its distillation process may be unavailable, prohibitively costly, or even unfeasible. For instance, in the legal domain, the motivations explaining why specific juridical actions have been taken may be articulated over multiple documents with several cross-references to legal laws and, thus, deeply entangled within domain-specific linguistic constructs and stylistic patterns. Here, correctly highlighting reasons for action may be prohibitively time-consuming, if possible at all. Therefore, since the beginning of my PhD, I have been actively working on directly using unstructured text as a form of knowledge due to the sheer amount of information in different domains. Formally, I denote this integration process as Unstructured Knowledge Integration (UKI).

Formalizing UKI

Contributions. In [Ruggeri, 2022], I introduce a preliminary theoretical analysis of UKI, setting the foundation to **define** and **analyze** an unstructured textual knowledge integration process systematically. In particular, I decompose UKI into three main processes: *Integration*, *Mapping*, and *Validation*. Integration defines the mechanism by which we should employ unstructured knowledge. Mapping describes how knowledge, or a portion of it, should be linked to input data according to the integration process, while Validation verifies if the adopted integration and mapping processes do not violate knowledge-specific requirements. According to how these processes are defined in a given domain, knowledge can play different roles during learning, such as a source of complementary information (e.g., additional data), support for reasoning (e.g., schemes and task instructions), and constraints. Altogether, my analysis denotes that learning with textual knowledge is an articulated process where several factors have to be aligned consistently during learning, including how knowledge is represented, for what purpose is knowledge employed, how knowledge is integrated, and how knowledge is validated.

How to integrate knowledge

As a first step in studying how unstructured textual knowledge could be integrated into machine learning models, I explore **embedding-based** approaches for comparing knowledge with input texts. In [Lagioia et al., 2019], I employ memory neural networks [Sukhbaatar et al., 2015] to enrich a classifier of unfair clauses in online consumer contracts to encode legal explanations of unfairness, denoted as *rationales*. In particular, I explore knowledge integration through **embedding-based semantic similarity** where the model compares input texts with rationales provided by legal experts. This allows the model to provide the most prominent selected rationales to support its predictions. Results on a specific category of legal unfairness show that the integration process leads to classification improvement over purely data-driven baselines. In [Ruggeri et al., 2022], we then extend the work of [Lagioia et al., 2019] to account for different unfairness categories. In particular, we introduce ToS-100, a dataset containing consumer contracts from 100 popular service providers annotated for nine distinct unfairness categories, five of which are annotated with rationales. Moreover, we explore training memory networks by providing **supervision** on which rationales to select when targeting unfair clauses. Experiments show that rationale supervision leads to a significant improvement in accurately selecting relevant rationales for most unfairness categories without affecting classification performance. These results denote that the way unstructured knowledge is formulated affects the effectiveness of the integration process. To measure to what extent these observations hold when **scaling** the size of the knowledge base (up to hundreds

of entries), in [Ruggeri et al., 2024a] I develop adaptable learnable strategies to sample a small portion of unstructured knowledge based on (i) how semantically similar was the retrieved content to the given input data and (ii) to what extent the retrieved content improved the classification. Sampling enables working with large knowledge bases with limited memory overhead, and models with limited knowledge often reach performance comparable to counterparts accessing the whole knowledge base.

KNOWLEDGE FOR REASONING

In addition to **how** knowledge could be integrated, I explore **for what** purpose knowledge could be employed. Specifically, I focus on integrating knowledge as **support for reasoning**. In [Muti et al., 2024a], I, along with collaborators specialized in hate speech, inject reasoning schemes proper of argument mining for implicit misogyny detection. More precisely, we prompt Large Langue Models (LLMs) with the intent of extracting the rationale (denoted as the warrant in argumentative theory [Toulmin et al., 1979]) behind hate speech, thus making it explicit for analysis. This methodology allows us to assess that LLMs fall short of reasoning capabilities about misogynistic comments, and they mostly rely on their implicit knowledge. Similarly, in [Muti et al., 2024b], we formulate misogyny detection as a word sense disambiguation task, where pejorative epithets are first disambiguated by an ad-hoc model to provide more contextual information to the misogyny classifier.

KNOWLEDGE AS INSTRUCTIONS

Unstructured knowledge for reasoning can also be expressed as **task-specific instructions** like, for instance, annotation guidelines. In [Korre et al., 2024], we collect a corpus of 300+ definitions for hate speech and employed a method derived from semantic componential analysis to extract conceptual components from these definitions (e.g., religion, race, age, etc.). We then prompt LLMs to evaluate if definitions with different components allow for detecting examples characterized by those components as hate speech, indicating a sensitivity to the cultural context conveyed by the definition. Similarly, in [Ruggeri et al., 2024b], we propose an annotation methodology based on the prescriptive paradigm [Rottger et al., 2022] that enforces reporting which criteria of annotation guidelines were employed by annotators to label examples. We then evaluate if learning with guidelines resembling human annotators leads to superior classification performance and fine-grained model analysis.

Future Work. UKI is a broad concept, encapsulating several distinct concepts, from the definition and maintenance of unstructured knowledge to its integration and validation. So far, my research on UKI has touched only some aspects, often subject to several limitations derived by the case of study of interest, including fixed and partial knowledge, integration by comparison via embedding-based operators, limited datasets, and domains. Currently, I'm working on **extending the analysis** introduced in [Ruggeri, 2022] to account for and compare with several existing research areas. Among the many, in-context learning and prompting [Liu et al., 2023]. In particular, prompts are a peculiar case study since they mix input text with several forms of knowledge like textual constraints, task-specific instructions, and additional data (e.g., few-shot examples). Another research direction is using **structured knowledge extraction as a sub-process of UKI**, where unstructured text is first pre-processed to extract abstract and high-level representations to aid individual UKI processes like validation. To do so, I want to explore neuro-symbolic approaches [Besold et al., 2021] as a formal way to assess how and to what extent knowledge has been used by a model to address a downstream task.

Structured Knowledge Extraction From Text

Introduction. Structured knowledge extraction is a valuable component for defining high-level reasoning capabilities. For instance, suppose the problem of integrating knowledge in the form of explanations (as in online consumer contracts) into a machine learning model. To do so, we would require a method that exposes the connection between input text and selected explanations, allowing answering the question: *For what reasons was input X connected to knowledge K?* Structured knowledge, like knowledge graphs and logic rules, is characterized by a semantic frame that is interpretable by design: the semantics of the nodes and edges in a graph are known, and upon those concepts, articulated operations like graph traversal are built. This is a desired property when dealing with machine learning models since we can use extracted knowledge as a proxy to interpret their decisions. In this context, my research focuses on the automatic extraction of structured knowledge from unstructured text to discover novel insights about tasks and what knowledge models acquired specifically through data-driven learning.

Extracting rationales

Contributions. A first step towards **knowledge discovery** is represented by automatically **extracting textual patterns** from unstructured text. In particular, patterns are portions of input texts highlighted by the model to perform predictions and are often denoted as *highlights* or *rationales* [Wiegreffe and Marasovic, 2021]. Inspired by previous work on integrating tree kernel methods for text classification [Lippi and Torroni, 2015], in [Ruggeri et al., 2021], I develop regularization methods to define neural classifiers capable of extracting structured rationales from text for Argument Mining. In particular, structured rationales are in the form of parsing trees derived from input texts to classify. The ultimate goal of this study is to identify whether argumentative components like claims (i.e., opinions) do follow specific syntactic or semantic patterns in a given domain, a hypothesis that has been discussed in related work [Shnarch et al., 2017, Lauscher et al., 2022]. While leading to performance improvements, the developed models required pre-processing steps for extracting rationales. To account for this issue and devise a general-purpose methodology, in [Ruggeri and Signorelli, 2024], we propose GenSPP, an effective method for extracting interpretable rationales by design. In particular, we train GenSPP through genetic-based optimization to overcome existing optimization issues of select-then-predict (SPP) architectures [Lei et al., 2016, Yu et al., 2021], resulting in more accurate and robust models.

Textual knowledge graphs

Knowledge is not necessarily input-specific as in rationalization models, and extracted content bits may be connected to each other **through semantic relations** like contextual dependency. A first step in this direction is the work of [Ruggeri et al., 2023], where I introduce ArgSciChat, a corpus of argumentative dialogues on scientific texts with the intent of developing chatbots capable of retrieving content from documents and discussing it through text generation. Subsequently, in [Ruggeri, 2022], I enrich ArgSciChat with argumentative component annotations with the intent of constructing an argumentative knowledge graph from these annotations corresponding to a specific dialogue state. This is achieved by developing ad-hoc argumentative classifiers and employing argumentative annotations as a knowledge base to guide the information retrieval process without relying on any other supervision.

Future Work. My current contributions mainly focus on extracting input-specific information as a proxy for interpretability. This is only a first step towards solutions capable of digesting such information into global and summarized content for knowledge discovery, which, so far, requires an active human intervention through qualitative analysis of extracted content. For this reason, I aim to extend GenSPP [Ruggeri and Signorelli, 2024] to include **aggregation mechanisms** to distill high-level information, mainly taking inspiration from several research areas

like topic modeling [Zhao et al., 2021], prototype networks [Snell et al., 2017], and neural clustering [Pakman et al., 2020]. The aggregated information is an abstract and robust representation of input textual content that can potentially benefit task-specific performance, similar to how linguistic dependencies like parse trees and abstract meaning representation graphs are applied in NLP [Bevilacqua et al., 2021]. Moreover, I aim to extend SPP architectures to **extract multiple patterns** from a single input text, where each pattern is viewed as an individual abstract concept [Poeta et al., 2023] with its own relations and structure. This is where I also plan to extend the work of [Ruggeri et al., 2021] to **replace attention-based** extractions with hard and stable selections as in GenSPP. Lastly, as my ultimate goal, I intend to evaluate how these developed knowledge extraction methods can be **integrated into a UKI** pipeline. For instance, we could define a mapping mechanism based on comparison as done with memory networks [Lagioia et al., 2019, Ruggeri et al., 2024a] that operates on extracted abstract content.

References

[Aristotle and Ross, 1933] Aristotle and Ross, W. D. (1933). *Metaphysics*, volume 1. Harvard University Press Cambridge, MA.

[Besold et al., 2021] Besold, T. R., d'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P. M., Hitzler, P., Kühnberger, K., Lamb, L. C., Lima, P. M. V., de Penning, L., Pinkas, G., Poon, H., and Zaverucha, G. (2021). Neural-symbolic learning and reasoning: A survey and interpretation. In Hitzler, P. and Sarker, M. K., editors, *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*, pages 1–51. IOS Press.

[Bevilacqua et al., 2021] Bevilacqua, M., Blloshmi, R., and Navigli, R. (2021). One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12564–12573. AAAI Press.

[Korre et al., 2024] Korre, K., Muti, A., Ruggeri, F., and Barrón-Cedeño, A. (2024). Untangling hate speech definitions: A semantic componential analysis across cultures and domains. *NAACL*. **In Press.**

[Lagioia et al., 2019] Lagioia, F., Ruggeri, F., Drazewski, K., Lippi, M., Micklitz, H., Torroni, P., and Sartor, G. (2019). Deep learning for detecting and explaining unfairness in consumer contracts. In Araszkiewicz, M. and Rodríguez-Doncel, V., editors, *Legal Knowledge and Information Systems - JURIX 2019: The Thirty-second Annual Conference, Madrid, Spain, December 11-13, 2019*, volume 322 of *Frontiers in Artificial Intelligence and Applications*, pages 43–52. IOS Press.

[Lauscher et al., 2022] Lauscher, A., Wachsmuth, H., Gurevych, I., and Glavaš, G. (2022). Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

[Lawrence and Reed, 2019] Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

[Lei et al., 2016] Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical*

Methods in Natural Language Processing, pages 107–117, Austin, Texas. Association for Computational Linguistics.

[Lippi and Torroni, 2015] Lippi, M. and Torroni, P. (2015). Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 185–191. AAAI Press.

[Liu et al., 2023] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

[Muti et al., 2024a] Muti, A., Ruggeri, F., Khatib, K. A., Barrón-Cedeño, A., and Caselli, T. (2024a). Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

[Muti et al., 2024b] Muti, A., Ruggeri, F., Toraman, C., Barrón-Cedeño, A., Algherini, S., Musetti, L., Ronchi, S., Saretto, G., and Zapparoli, C. (2024b). PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12700–12711, Torino, Italy. ELRA and ICCL.

[Pakman et al., 2020] Pakman, A., Wang, Y., Mitelut, C., Lee, J. H., and Paninski, L. (2020). Neural clustering processes. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7455–7465. PMLR.

[Poeta et al., 2023] Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., and Baralis, E. (2023). Concept-based explainable artificial intelligence: A survey. *CoRR*, abs/2312.12936.

[Rottger et al., 2022] Rottger, P., Vidgen, B., Hovy, D., and Pierrehumbert, J. (2022). Two contrasting data annotation paradigms for subjective NLP tasks. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

[Ruggeri, 2022] Ruggeri, F. (2022). *Towards Unstructured Knowledge Integration in Natural Language Processing*. PhD thesis, alma.

[Ruggeri et al., 2022] Ruggeri, F., Lagioia, F., Lippi, M., and Torroni, P. (2022). Detecting and explaining unfairness in consumer contracts through memory networks. *Artif. Intell. Law*, 30(1):59–92.

[Ruggeri et al., 2021] Ruggeri, F., Lippi, M., and Torroni, P. (2021). Tree-constrained graph neural networks for argument mining. *ArXiv*.

[Ruggeri et al., 2024a] Ruggeri, F., Lippi, M., and Torroni, P. (2024a). Combining transformers with natural language explanations. *arXiv*.

[Ruggeri et al., 2023] Ruggeri, F., Mesgar, M., and Gurevych, I. (2023). A dataset of argumentative dialogues on scientific papers. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada. Association for Computational Linguistics.

[Ruggeri et al., 2024b] Ruggeri, F., Misino, E., Muti, A., Korre, K., Torroni, P., and Barrón-Cedeño, A. (2024b). Let guidelines guide you: A prescriptive guideline-centered data annotation methodology. *ArXiv*.

[Ruggeri and Signorelli, 2024] Ruggeri, F. and Signorelli, G. (2024). Interlocking-free selective rationalization through genetic-based learning. *arXiv*.

[Shnarch et al., 2017] Shnarch, E., Levy, R., Raykar, V., and Slonim, N. (2017). GRASP: Rich patterns for argumentation mining. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1345–1350, Copenhagen, Denmark. Association for Computational Linguistics.

[Snell et al., 2017] Snell, J., Swersky, K., and Zemel, R. S. (2017). Prototypical networks for few-shot learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

[Sukhbaatar et al., 2015] Sukhbaatar, S., szlam, a., Weston, J., and Fergus, R. (2015). End-to-end memory networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

[Toulmin et al., 1979] Toulmin, S., Rieke, R. D., and Janik, A. (1979). *An introduction to reasoning*. Macmillan, New York.

[von Rueden et al., 2023] von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., and Schuecker, J. (2023). Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633.

[Wiegreffe and Marasovic, 2021] Wiegreffe, S. and Marasovic, A. (2021). Teach me to explain: A review of datasets for explainable natural language processing. In Vanschoren, J. and Yeung, S., editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

[Xu et al., 2020] Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., and Zhao, J. (2020). Distinguish confusing law articles for legal judgment prediction. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online. Association for Computational Linguistics.

[Yu et al., 2021] Yu, M., Zhang, Y., Chang, S., and Jaakkola, T. (2021). Understanding interlocking dynamics of cooperative rationalization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12822–12835. Curran Associates, Inc.

[Yu et al., 2022] Yu, X., Blanco, E., and Hong, L. (2022). Hate speech and counter speech detection: Conversational context does matter. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

[Zhao et al., 2021] Zhao, H., Phung, D. Q., Huynh, V., Jin, Y., Du, L., and Buntine, W. L. (2021). Topic modelling meets deep neural networks: A survey. In Zhou, Z., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4713–4720. ijcai.org.