EU-DOCS for SmUCS

CO-SUPERVISED SUBJECT PROPOSAL FOR A DOCTORAL CONTRACT

Title of the thesis project Multi-disciplinary knowledge management of Coastal Area publications	
La Rochelle University Research Unit LIENSs - LIttoral ENvironnement et Sociétés L3i - Laboratoire Informatique, Image et Interaction	**Partner university** University of Helsinki **Cotutelle research unit** Learning Language Lab.
Name of the LRUniv supervisor Olivier de Viron	**Name of the co-supervisor** Roman Yangarber
Non-academic partner Siren Investigation	
Main research field: Natural Language Processing **Secondary research field**: Knowledge Management	
Keywords: Coastal areas, Natural language processing, deep learning, Scientific literature analysis, knowledge map	

Scientific description of the research project

Scientific context

Coastal areas are dynamic interfaces between natural systems and human societies, concentrating critical environmental and economic challenges. These regions face multiple pressures, including rising human populations, biodiversity collapse, sea level rise, extreme climate events, and shifting socio-economic and urban contexts (Barbier, 2015; Merkens et al., 2016; Neumann et al., 2015; Newton et al, 2020; Nichols et al. 2021, Spalding et al., 2014)

. Understanding coastal dynamics is essential for predicting the impacts of ongoing environmental upheavals on ecosystems and communities and for devising effective adaptation and remediation strategies.

The interdisciplinary nature of coastal systems presents a significant challenge in achieving a comprehensive understanding of their functioning (Cloern et al., 2016 ; Glavovic et al., 2015). This requires synthesizing and integrating knowledge from diverse fields and bridging conceptual gaps across disciplines. However, key concepts within one discipline often remain disconnected or unaligned with those from others. Consequently, current interdisciplinary efforts either focus narrowly on specific areas and interactions or result in an overly simplistic understanding of the broader system.

At the same time, the scientific literature on coastal areas is extensive, comprising nearly 70,000 publications. This project is part of a broader effort to structure and synthesize this vast body of knowledge by leveraging natural language processing (NLP), deep learning (Lane et al., 2019), and lexical statistical methods (Miner et al., 2012; Mendes et al., 2019) to analyze the entire corpus of publication abstracts. Building on existing work, a PhD student co-supervised by the project PIs has developed a neural network algorithm to identify coastal entities and their interactions at the document level (Delaunay et al., 2024). This algorithm adapts the ARDI method (Actors, Resources, Dynamics, and Interactions; Michel, 2011), originally used for participatory interdisciplinary modeling, to extract and connect relevant entities within individual abstracts using deep learning techniques.

Scientific objectives

The next challenge is to scale this understanding from the level of abstracts to a corpus-wide perspective, constructing a comprehensive knowledge map. Achieving this goal requires addressing disciplinary silos, where concepts may be expressed differently across fields or entirely absent in certain domains. For example, the concept of coastal area representation— how people perceive the area as beautiful, safe, or interesting—is central to humanities and influences societal behavior, but is rarely linked to ecological or geomorphological studies.

The project will address this challenge by hiring a PhD student to develop a novel algorithm that synthesizes document-level information into a global knowledge map. This algorithm will start from the functioning information detected at the level of abstracts, position it into a common framework, and link it to the other pieces of information.
The resulting knowledge map will reveal both well-established and underexplored connections between entities, as well as the interdisciplinary relationships underlying them. It is expected to highlight knowledge gaps and uncover previously unrecognized patterns or regularities. Crucially, this will have a dual impact: in the domain of coastal systems, it will result in a more advanced and a better integrated understanding. However, on a much broader scale, it will facilitate our understanding about linking concepts across widely diverse scientific disciplines that have previously lacked alignment.

Methodology

The project will proceed through the following steps:

1. Ontology construction: A coastal systems ontology will be created from the corpus of abstracts (Cimiano, 2016; Asim et al., 2018).
2. Ontology refinement: Disciplinary biases will be addressed by uniquely identifying each entity and linking it to external knowledge bases.
3. Interaction classification: Interactions between entities will be categorized into a streamlined functional interaction ontology to reduce complexity.
4. Relational graph development: A graph-based approach (e.g., relational graph techniques) will be used to model the connections between entities and across disciplines.

Challenges

Two main challenges are anticipated in this project: the scale of the corpus (see for instance Wand et al., 2020) and its interdisciplinary nature (Augenstein, 2017; Beltagy et al., 2019; Gabor et al, 2018).

1. **Scale of the corpus**: With approximately 70,000 abstracts, the corpus to analyze is huge and composed of short texts. Natural Language Processing (NLP) on a large corpus of short texts presents unique challenges, particularly in validating the proposed solutions (Le & Mikolov, 2014). The diversity and volume of data complicate comprehensive model evaluation, making it difficult to detect systematic biases or errors. Moreover, short texts often provide limited context, which can significantly impact the accuracy of semantic or thematic analyses. The inherent variability in a large number of brief texts can obscure subtle trends or important relationships, requiring sophisticated statistical and sampling approaches for reliable validation. Finally, balancing generalization and specificity becomes crucial, as models must be robust enough to handle corpus diversity while capturing the nuanced characteristics of each individual short text.

To tackle the challenges of processing large corpora of short texts in NLP, a comprehensive and multi-pronged approach is proposed. First, the algorithm will be tested on smaller sub-corpora and specific targeted questions, adhering to best practices in model evaluation (Devlin et al., 2018). Input from a panel of disciplinary experts will play a crucial role in validating results and refining methods, in line with the collaborative approach employed by Wang et al. (2020) for the CORD-19 data set.

The methodology also includes implementing stratified cross-validation to ensure representative test subsets (Sechidis et al., 2011), conducting progressive evaluations to pinpoint performance degradation, and performing qualitative error analysis to understand limitations (Ribeiro et al., 2020). Benchmarking against established methods using standard datasets (Agirre et al., 2015) and conducting robustness tests will further evaluate the algorithm's reliability.

To enhance effectiveness, domain-specific evaluation metrics will be developed (Reimers & Gurevych, 2019), and the algorithm will undergo iterative refinement based on expert feedback. Interactive visualizations, adhering to best practices in NLP data visualization (Liu et al., 2019), will be designed to facilitate the interpretation of results. Lastly, long-term evaluation will ensure the algorithm's stability and consistency over time, particularly as the corpus grows.

This structured and adaptive framework aims to address the complexities inherent in NLP applications on large and heterogeneous datasets effectively.

2. **Interdisciplinary complexity**: Natural Language Processing on interdisciplinary corpora presents significant challenges due to the diverse nature of the content. The variability in terminology, writing styles, and conceptual frameworks across disciplines can lead to ambiguity and misinterpretation by NLP models (Augenstein et al., 2017). Domain- specific jargon and technical language often have different meanings in various fields, making it difficult for general-purpose NLP tools to accurately process the text (Beltagy et al., 2019). Additionally, the interconnectedness of concepts across disciplines can create complex semantic relationships that are challenging to capture and represent computationally (Gábor et al., 2018). The lack of standardized vocabularies and ontologies across different fields further complicates the task of entity recognition and relation extraction (Wang et al., 2020). Moreover, the varying levels of abstraction and specificity in different disciplines can affect the performance of text classification and topic modeling algorithms (Jurgens et al., 2018). These challenges necessitate the development of more robust and adaptable NLP techniques that can effectively handle the heterogeneity and complexity of interdisciplinary corpora.

To *address the challenges of NLP on interdisciplinary corpora, a multi-faceted approach is proposed. Ontology construction and refinement with expert input will be crucial in mitigating terminological and conceptual ambiguities (Pesquita et al., 2014). The inherent redundancies in large corpora are expected to resolve some ambiguities automatically, while statistical analyses may reveal new insights into interdisciplinary connections (Yan et al., 2012).*

Additionally, employing domain adaptation techniques can help models better handle discipline-specific nuances (Beltagy et al., 2019). Implementing a hierarchical classification approach can capture both broad interdisciplinary themes and specific domain concepts (Silla & Freitas, 2011). Utilizing transfer learning methods can leverage knowledge from well-resourced domains to improve performance in less- studied interdisciplinary areas (Ruder et al., 2019).

Furthermore, incorporating active learning strategies can efficiently utilize expert knowledge to refine models iteratively (Settles, 2009). Developing interdisciplinary benchmarks and evaluation metrics will be essential for assessing model performance across diverse domains (Wang et al., 2020). Finally, employing interpretable AI techniques can provide insights into model decision-making, facilitating better alignment with expert knowledge and interdisciplinary understanding (Ribeiro et al., 2016).

Expected outcomes

The project aims to produce an interdisciplinary knowledge map of coastal systems that bridges disciplinary divides, connects entities across domains, and reveals overlooked relationships. This map will advance understanding of the interactions between ecological, social, and technical systems, identify blind spots, and generate new hypotheses for research.

The identification of knowledge gaps may guide future research, fostering interdisciplinary collaborations to address underexplored interactions or blind spots. The map could also enhance the accessibility of scientific knowledge by synthesizing complex interdisciplinary data into actionable insights. For example, connecting societal concepts like tourism attractiveness to ecological data could inform strategies for sustainable coastal development. Similarly, revealing overlooked connections between disciplines might inspire novel solutions to pressing coastal issues, such as biodiversity loss or climate adaptation.

The project's methods and results are expected to set a precedent for applying NLP and deep learning to complex social-ecological systems, broadening its impact beyond coastal studies. These

refinements position the project as a cornerstone for both advancing scientific understanding and delivering tangible benefits to stakeholders managing vulnerable coastal environments.

Finally, the resulting knowledge map will serve as a decision-support tool for coastal management, helping policymakers identify critical areas of intervention and prioritize resources.

Bibliography

- Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., ... & Wiebe, J. (2015). SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 252-263.

- Asim, M.-N., et al. (2018). A survey of ontology learning techniques And applications, Database, article I 101; doi:10.1093/database/bay101.

- Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). "SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications." Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 546-555.

- Barbier, E. B. (2015). "Climate change impacts on rural poverty in low-elevation coastal zones." Estuarine, Coastal and Shelf Science, 165, A1-A13.

- Beltagy, I., Lo, K., & Cohan, A. (2019). "SciBERT: A Pretrained Language Model for Scientific Text." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3615-3620. (Ce travail aborde les difficultés de l'adaptation des modèles de langue aux textes scientifiques multidisciplinaires.)

- Cimiano P., Ontology Learning and Population from Text, Springer, 350 pp. , 2006.

- Cloern, J. E., et al. (2016). "Human activities and climate variability drive fast-paced change across the world's estuarine–coastal ecosystems." Global Change Biology, 22(2), 513-529.

- Delaunay J. et al. CoastTerm: A Corpus for Multidisciplinary Term Extraction in Coastal Scientific Literature. (2024) Lect. Not.in Comp. Sc.e, 15048, pp. 97 – 109, 10.1007/978-3-031-70563-2_8.

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

- Gábor, K., Buscaldi, D., Schumann, A. K., QasemiZadeh, B., Zargayouna, H., & Charnois,

T. (2018). "SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers." Proceedings of The 12th International Workshop on Semantic Evaluation, 679-688.

- Glavovic, B. C., et al. (2015). "Living on the Margin in the Anthropocene: Engagement arenas for sustainability research and action at the ocean–land interface." Current Opinion in Environmental Sustainability, 14, 232-238.

- Jurgens, D., et al. (2018). Measuring the Evolution of a Scientific Field through Citation Frames.

- Lane et al., Natural language processing in action: understanding, analyzing, and generating text with Python, Manning, 2019.

- Le, Q., & Mikolov, T. (2014). "Distributed representations of sentences and documents." International conference on machine learning (pp. 1188-1196). PMLR.

- Liu, S., Wang, X., Collins, C., Dou, W., Ouyang, F., El-Assady, M., ... & Liu, S. (2019). Bridging text visualization and mining: A survey. IEEE transactions on visualization and computer graphics, 25(9), 2897-2917.

- Mendes et al., Mapping pharmacy journals: A lexicographic analysis. Res. Soc. Adm. Pharm. 15, 1464–1471 (2019).

- Merkens, J. L., et al. (2016). "Gridded population projections for the coastal zone under the Shared Socioeconomic Pathways." Global and Planetary Change, 145, 57-66.

- Miner et al., Practical text mining and statistical analysis for non-structured text data applications Academic Press, 2012

- Michel, ARDI: a co-construction method for participatory modeling in natural resources management, Ecol. Soc. 16, 1, 44, 2011.

- Neumann, B., et al. (2015). "Future coastal population growth and exposure to sea- level rise and coastal flooding - a global assessment." PloS one, 10(3), e0118571.

- Newton, A., et al. (2020). "Anthropogenic, Direct Pressures on Coastal Wetlands." Frontiers in Ecology and Evolution, 8, 144.

- Nicholls, R. J., et al. (2021). "A global analysis of subsidence, relative sea-level change and coastal flood exposure." Nature Climate Change, 11(4), 338-342.

- Pesquita, C., et al. (2014). The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. Journal of Biomedical Semantics, 5(1), 4.Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.

- Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4902–4912.

- Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. Machine Learning and Knowledge Discovery in Databases, 145-158.

- Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, 22(1-2), 31-72.

- Spalding, M. D., et al. (2014). "The role of ecosystems in coastal protection: Adapting to climate change and coastal hazards." Ocean & Coastal Management, 90, 50-57.

- Settles, B. (2009). Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences.

- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., ... & Kohlmeier, S. (2020). "CORD-19: The Covid-19 Open Research Dataset." ArXiv, abs/2004.10706.

- Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. Journal of the American Society for Information Science and Technology, 63(7), 1313-1326.

PhD student profile and skills required

A strong candidate would likely come from one of the following disciplines: Computer Science (especially NLP, AI, or machine learning), Computational Linguistics or Data Science.

The ideal candidate will have prior experience in natural language processing, particularly with scientific or interdisciplinary corpora, and a solid understanding of knowledge representation or ontology development. Familiarity with deep learning frameworks for text (e.g., BERT, SciBERT, Llama, GPT) and exposure to interdisciplinary research— especially in environmental or coastal contexts—is highly desirable.

The candidate should also possess an interdisciplinary mindset, strong critical thinking and communication skills, intellectual curiosity, adaptability to complex challenges, and the ability to work independently on exploratory research.

Scientific alignment with EU-DOCs for SmUCS objectives

The proposed research aims to develop a comprehensive knowledge map of coastal systems by applying advanced natural language processing (NLP) and deep learning techniques to an extensive corpus of scientific literature. This innovative approach holds transformative potential for addressing the challenges of sustainable development in anthropized coastal areas.

First, by bridging disciplinary silos and connecting concepts across diverse fields, the project fosters an integrated understanding of coastal systems. This holistic perspective is essential for devising sustainable solutions to the multifaceted challenges facing coastal zones. The resulting knowledge map will also identify critical gaps in existing research, highlighting underexplored connections and guiding future studies toward areas traditionally overlooked by discipline-specific approaches.

Furthermore, the project aims to transform complex interdisciplinary data into actionable insights, providing coastal managers and policymakers with decision-support tools. These tools will facilitate more informed and effective strategies for sustainable development. The enhanced understanding of coastal dynamics will also improve risk assessment capabilities, enabling more accurate predictions and mitigation of climate change impacts, sea-level rise, and other environmental pressures.

Additionally, by uncovering unexpected links between various components of coastal systems, the research may inspire innovative, cross-disciplinary solutions to pressing issues such as biodiversity loss, coastal erosion, and socio-economic challenges. It will also identify key leverage points to strengthen the adaptive capacity of both natural and human systems in response to rapid environmental changes.

Finally, by integrating ecological, social, and economic factors, the knowledge map will contribute to more sustainable resource management practices, balancing human needs with ecosystem health. This project thus represents a critical step toward advancing our understanding of anthropized coastal zones and promoting more effective and sustainable management practices in these vital areas.

Societal and economic challenges and contributions

By synthesizing vast interdisciplinary research, the project aims to empower policymakers and coastal managers with a more nuanced understanding of coastal dynamics, enabling more informed and effective decision-making. This can result in economic savings and improved societal outcomes. Moreover, the knowledge map will help identify sustainable development strategies that harmonize economic growth with environmental conservation—an essential balance for coastal economies reliant on tourism, fishing, and marine resources.

The project's integrated approach can significantly enhance climate change adaptation strategies, providing economically vital insights for coastal communities facing sea-level rise and extreme weather events. Additionally, by pinpointing knowledge gaps and underexplored connections, it can guide more efficient allocation of research funding and resources, amplifying the impact and cost-effectiveness of future studies.

Crucially, the knowledge map may act as a catalyst for innovation by revealing unexpected connections within coastal systems, inspiring new technologies or solutions to pressing issues with potential economic benefits. Enhanced risk assessment capabilities, derived from a comprehensive understanding of coastal dynamics, can further support industries, insurance providers, and investment decisions tied to these regions.

The project also seeks to improve stakeholder engagement by offering intuitive visual representations of complex systems, fostering clearer communication with local communities, businesses, and policymakers. This inclusivity can lead to more robust and participatory coastal management strategies. Furthermore, identifying intersectoral linkages could uncover opportunities for economic diversification in coastal areas, reducing vulnerability to sector-specific shocks.

Finally, the interdisciplinary approach will deepen the valuation of ecosystem services provided by coastal areas, supporting more sustainable economic models. The methodologies and insights developed in this project could establish the research team as a global leader in coastal studies, attracting international collaborations and funding, and bolstering global competitiveness in interdisciplinary research.

Partnership context

This project is part of a larger-scale effort to build a generic model of the coastal system. As such, he will benefit from interaction with the project team, with disciplinary experts of the coastal areas as well as with coastal territories stake holders.

The PhD student will also have the opportunity to work with Siren Investigation Company, specialized in data investigation and knowledge management. Siren Investigation focuses on research and development on methods using AI, automation and advanced search, Siren links data from open source, vendors and classified sources allowing investigators to surface and analyze risks and threats. Thus, this partnership will be very benefic for the student, as he will be able to spend at least 3 months in the company on a concrete subject dealing with his topic in a major actor of the domain.

HR EXCELLENCE IN RESEARCH