Doctoral project

INFORMATION RATING AND ANALYSIS OF KNOWLEDGE DYNAMICS. APPLICATION TO THE TEMPORAL MONITORING OF THE RELIABILITY OF BIBLIOGRAPHIC INFORMATION ON INSECTS AS VECTORS OF PLANT PATHOGENS.

Contacts
Claire Nédellec (claire.nedellec@inrae.fr), Nicolas Sauvion (Nicolas.sauvion@inrae.fr), Vincent Guigue (vincent.guigue@agroparistech.fr).

Affiliation

Doctoral school Sciences and Technologies of Information and Communication
Speciality Computer Science

Research laboratory MaIAGE – Mathématics and Informatics applied from Genome to Environnement
Bibliome group: acquisition and formalisation ok knowledge from texts
Pôle B. Data, Knowledge, Machine Learning and Interactions
Institution: Université Paris-Saclay – GS Computer Science

1st year
2024
Starting date
01/10/2024
Application deadline
30/06/2023 à 23h59
Interdiciplinary
Yes

Doctoral Project

Principal supervisor
NEDELLEC Claire - 36282
01 34 65 28 78
claire.nedellec@inrae.fr
HDR: oui

Supervision team
Claire NEDELLEC (MaIAGE-INRAE) [DR-HDR]
Nicolas SAUVION (PHIM-INRAE) [IR-HDR]
Vincent Guigue (MIA-Paris-Saclay, AgroParisTech) [PR- HDR]

Robert Bossy (IR, MaIAGE) will contribute to the supervision.
The student will be a member of the MaIAGE lab. (Jouy-en-Josas) and will periodically visit the MIA-Paris-Saclay and PHIM (Montpellier) labs.

Title

INFORMATION RATING AND ANALYSIS OF KNOWLEDGE DYNAMICS. APPLICATION TO THE TEMPORAL MONITORING OF THE RELIABILITY OF BIBLIOGRAPHIC INFORMATION ON INSECTS AS VECTORS OF PLANT PATHOGENS.

Keywords

Natural Language Processing, Information Extraction, Data Quality, Uncertainty, Language Models, Deep Learning, Time Series, Information Reliability, Biological Risk, Crop Health, Insect Vectors, Biological Interaction, Health monitoring.

Abstract
In biology, scientific intelligence covers a growing number of objects whose knowledge is evolving rapidly, with established or unestablished statements frequently challenged. Scientists are lacking in tools for extracting, structuring and assessing the relevance of published information to scientific knowledge. Relevance depends on both the novelty and validity of the information. These can be measured by linguistic, domain-specific (e.g. experimental protocol), documentary (e.g. citation network), and temporal (citation dynamics and modality) clues, as well as by their consistency with scientific knowledge of the field. The project aims to understand how information spreads through publication and to analyze the relationship between the dynamics of evolution and the reliability of information. Starting from its primary source, the analysis will list and qualify the way in which each piece of information is cited over time until a consensus is reached on its veracity, positive or negative.
The thesis will focus on a concrete case study, the evolution of knowledge on insect vectors of plant pathogens at the origin of major health crises in agriculture, in particular scientific knowledge on interactions between species and their geographical distribution, aiming at surveying biodiversity and anticipating epidemics.

Topic
Faced with the maelstrom of information accessible in real-time, whichever our field of interest, it is today more than ever necessary to equip ourselves with the means to separate the wheat from the chaff. The automatic verification of the veracity of information is a particularly challenging task from an algorithmic point of view: this thesis aims precisely at developing original approaches to the reliability of textual information by integrating both linguistic (NLP, language models) and dynamic (time series) dimensions [Le Naour, 2023a]. Starting from a language model, the challenge is to integrate the dynamic dimension into a versatile basic model capable of handling a wide variety of information centered around entities, events, or scientific claims. This information can come from a variety of sources, such as knowledge bases, scientific articles, or professional publications. Time-tracking has been studied on sequences of product reviews [Wang, 2011], message responses/transfers [Bourigault, 2016] and, of course, scientific citations [Šubelj, 2013]. The project aims to understand how information spreads in scientific publications and to analyze the link between the dynamics of evolution and the reliability of information. Starting with its primary source, the analysis will list and qualify the way in which each piece of information is cited over time until a consensus is reached on its veracity, whether positive or negative.

Domain
The thesis will focus on a concrete case study, the evolution of knowledge on insect vectors of plant pathogens responsible for major health crises in agriculture [Marie-Jeanne, 2020]. The primary objective will be to analyze the bibliography on these insect vectors in order to better understand the dynamics of the reliability of the observations reported in the literature, aiming at building partially or fully automated monitoring tools on the subject.
The project responds to a double requirement: to take into account the evolution of knowledge [e.g. the very recent role of alder in the propagation of the grapevine Flavescence dorée (Malembic et al., 2020)], while disqualifying poorly extracted, unconfirmed or false knowledge [e.g. the ability of a leafhopper to transmit an apple phytoplasma in the laboratory (Hegab & El-Zohairy 1986), never since confirmed in the orchard (Fischnaller et al., 2020)]. The models will be also validated on academic corpora for the detection of review spam [Wang, 2011] and fake news [Shu, 2017].
This thesis aims to develop innovative models at the interface between text and time series, drawing on multimodal architectures of foundation models [Radford, 2021] as well as Implicit Neural Representation-type approaches [Le Naour, 2023b]. But the applicative stakes are also central: the combination of these analyses should enable significant advances in estimating the veracity of information.

Goals
Qualify the nature of quotations and analyze linguistic aspects: Distinguishing whether a citation reinforces or denies a piece of information requires a detailed analysis of citation contexts and linguistic markers that indicate the acceptance or rejection of a piece of information. This problem is part of the field of information extraction, in which language models have profoundly changed the state of the art [Taillé, 2020].
Projecting these analyses to thematic dimensions: linking linguistic markers to the facet of information and citation in question to understand which aspects of the information are praised or criticized [Zhang, 2019]. The

challenge is both to improve the performance of information qualification and to propose a new fine-grained scale of analysis.

Analyze the dynamics of information propagation: Understand how information evolves from its primary source, through its qualified replications, to consensus or challenge, while modeling the duration of this transition and fluctuations in information reliability. All aspects of the dynamics are important: trajectory, speed, etc. New time series modeling techniques allow us to consider approaches that are more interpretable [Le Naour, 2023a] or better adapted to irregular sampling frequencies [Le Naour, 2023b].

Developing multimodal models: Creating foundation models, derived from language models (to a greater or lesser extent) and combining different modalities (documents, knowledge graphs), increases analysis capabilities and opens up new applications. The challenge of this project is to apply this paradigm by crossing linguistic aspects with information dynamics to assess the reliability and nature of information in a more comprehensive framework. In this context, Implicit Neural Representation (INR) approaches provide an opportunity to model different temporal and frequency aspects [Le Naour, 2023b] and combine them with other linguistic descriptors.

Exploring different applications: Measuring the reliability of information is a challenge, from fake news and review spam to consensus analysis in scientific literature. The aim of this thesis is to build robust models that can be used in a variety of situations. In particular, we will analyze the bibliography on insect vectors of plant pathogens to better understand the dynamics of the reliability of reported observations: this is a critical application in health monitoring to model epidemiological risks [Morris, 2022].

The student will contribute to studying the specificity of psyllid-phytoplasm-fruit tree relationships to evaluate and develop the thesis proposals. They will adjust the weightings of the various criteria with the support of the project partners to take into account the time/risk trade-off of missing information. The relatively limited biological scope of the pathosystems selected as case studies will enable different scenarios to be explored in detail.

Context

This multidisciplinary thesis involves a number of supervisors, each with their own expertise. Claire Nedellec (DR INRAe, UMR MAIAGE) will play a central role in the supervision of the thesis, both for her expertise in information extraction methods and for her knowledge of the use case of insect vectors of plant pathogens. Vincent Guigue (PR AgroParisTech, UMR MIA-PS) will contribute his expertise in time series and his motivation to build a nested system at the interface with the textual modality. The involvement of Nicolas Sauvion (IR, HDR, UMR PHIM) in the supervision of the project means that we can look forward to major applications in the field of health monitoring and epidemiological risks. Robert Bossy works with Claire Nedellec on a daily basis and will strengthen the management team.

Method

The planned thesis program is as follows.

Material & Methods. The student will first familiarize themself with the biological question and the existing documentary resources and databases on the subject of vector transmission of plant pathogens. They will also familiarize themself with the example datasets (corpus), the information extraction workflow and knowledge graph provided.

Language models. The student will formalize the research challenge and propose novelty and relevance indicators and the NLP methods to predict their values for the corpus and information pieces extracted by the NLP workflow. These proposed indicators will be derived from a state-of-the-art study, an analysis of the approach taken by monitoring specialists, and an in-depth manual exploration of examples from the corpus. Among them, is the use of natural language processing (NLP) techniques to identify and qualify citations, and analyze linguistic and thematic aspects. Build a fine-tuning reference model capable of qualifying scientific citations and the various aspects addressed in a product review or social network message [Taillé, 2020; Tang, 2022].

Temporal modeling: Carry out the relevant literature review on the modeling of irregularly sampled time series. In order to build multimodal approaches, we will focus on representation learning systems, in particular INR., and on information propagation in social networks [Le Naour, 2023b].

Multimodal approaches: Analyze multimodal fusion strategies in the literature [Radford, 2021]. Develop models that integrate both linguistic and dynamic variables to predict information reliability and understand its evolution over time.

Performance. Different strategies will be tested according to the nature of the indicators - methodological, linguistic, biological, or bibliographical - and their expression in the corpus. Endogenous training and evaluation of the methods developed will be based on examples obtained through retrospective analysis of publications on biological interactions. Exogenous training and evaluation will be carried out by exploiting external knowledge bases. Although NLP metrics are numerous, analyzing the performance of generative models remains a challenge today [Zhang, 2019]. Assessing performance in new applications such as the evolution of information reliability in the scientific literature requires the development of metrics dedicated to the task.

Expected results

In a situation of massive and continuous information flow, the expected result is to support the acquisition of scientific knowledge from textual data in the most exhaustive and reliable way possible. The measure based on an assessment of its reliability and novelty, will facilitate sorting, interpretation and decision making. To answer this question, the student will use original AI methods to measure the novelty and plausibility of information based on (1) the domain expertise, and (2) a diachronic analysis of the publication of information and its comparison with consensus knowledge at a given point in time. This approach, which distinguishes the contributions of the different criteria to the decision, will provide the user with a detailed explanation of how the assigned likelihood score is computed.

Scientific, material and financial conditions

The student will have access to the computational and storage resources of the MaIAGE unit and to the Jean Zay and lab.IA GPUs for the deep learning approaches used for information extraction.
The student will split their time between MaIAGE (Jouy-en-Josas) and MIA-Paris-Saclay with regular visits to PHIM (Montpellier).
They will be affiliated to the Computer Science Graduate School at Paris-Saclay University and employed by INRAE. We offer a stimulating transdisciplinary research environment with many opportunities for in-house, national, and international collaborations and access to computing GPU resources and state-of-the-art research equipment. The gross salary per month for the three-year contract is 2 100 (in 2024) to 2300 (in 2026) including the social security package (healthcare, pensions, unemployment benefits).

Collaborations

Biologists with expertise in insect vectors and computer scientists with expertise in NLP will work closely together to formalize knowledge in the field, establish criteria for relevance and novelty in publications, develop methods, and validate results.

Objectifs de valorisation des travaux de recherche du doctorant : diffusion, publication et confidentialité, droit à la propriété intellectuelle,...:

NLP conferences (*Coling*, *EMNLP*, *LREC*, *BioNLP*, *ACL*) will be relevant depending on the results and the degree of originality of the methods. Software and datasets will be published in public repositories (*recherche.data.gouv*, *GitHub*) and data papers (*F1000Research*, *Data in Brief*). Software and data will be published under free licenses (Apache, *Creative Commons*).
Targeted bioinformatics journals, such as *Peer Community Math Comp Biol*, *BMC Bioinformatics*, will be appropriate for promoting methodological results. More general journals such as *Scientific Reports*, *Plos One*, *Journal of Pest Science*, *Plant Pathology* will be more appropriate to reach a broader audience of biologists.

Funding

INRAE grant

Profile

The student will have a strong background in AI, NLP, and knowledge representation acquired at the Master's level. Significant work experience or training in biology is a plus. S/he will have solid computer development skills.

Level in French A1

Level in English B2
References

Supervisor references

Chaix E., Deléger L., Bossy R., & Nédellec C. (2019). Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology* 63-75 https://doi.org/10.1016/j.fm.2018.04.011

Ferré A., Bossy R., Ba M., Deléger L., Lavergne T., Zweigenbaum P., & Nédellec C. (2020). Handling Entity Normalization with no Annotated Corpus: Weakly Supervised Methods Based on Distributional Representation and Ontological Information. *Proc. of* LREC-2020, 1959–1966

Le Naour E., Agoua G., Baskiotis N. & Guigue V. (2023) Interpretable time series neural representation for classification purposes, 10th IEEE International Conference on Data Science and Advanced Analytics

Le Naour, E.., Serrano, L., Migus, L., Yin, Y., Agoua, G., Baskiotis, N., & Guigue, V. (2023). Time series continuous modeling for imputation and forecasting with implicit neural representations. arXiv preprint arXiv:2306.05880.

Marie-Jeanne V., Bonnot F., Thébaud G., Peccoud J., Labonne G., & Sauvion N. (2020). Multi-scale spatial genetic structure of the vector-borne pathogen 'Candidatus phytoplasma prunorum' in orchards and in wild habitats. *Scientific Reports* 10, 5002

Morris C.E., Geniaux G., Nédellec C., Sauvion N., & Soubeyrand S. (2022). One Health concepts and challenges for surveillance, forecasting and mitigation of plant disease beyond the traditional scope of crop production. *Plant Pathology*, 71, 86-97

Sauvion N., Peccoud J., Meynard C., Ouvrard D. (2021). Occurrence data for the two Cacopsylla pruni cryptic species (Hemiptera: Psylloidea) *Biodiversity Data Journal* 9, pp.e68860. ff10.3897/BDJ.9.e68860

Simon É., Guigue V., & Piwowarski B. (2019). Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1378-1387).

Taillé B., Guigue V., & Gallinari P. (2020). Contextualized embeddings in named-entity recognition: An empirical study on generalization. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020*, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42 (pp. 383-391). Springer International Publishing.

Taillé B., Guigue V., Scoutheeten G., & Gallinari P. (2020). Let's Stop Incorrect Comparisons in End-to-end Relation Extraction!. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3689-3701).

Taillé B., Guigue V., Scoutheeten G., & Gallinari P. (2021). Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10438-10449).

Tang A., Deléger L., Bossy R., Zweigenbaum P., & Nédellec C. (2022). Do syntactic trees enhance domain-specific BERT models for relation extraction? *Database*, 2022, https://doi.org/10.1093/database/baac070.

External references

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 596-606).

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, 9(1), 2158244019829575

Alomar, O., Batlle, A., Brunetti, J. M., García, R., Gil, R., Granollers, T., ... & Virgili-Gomà, J. (2016). Development and testing of the media monitoring tool med is YS for the monitoring, early identification and reporting of existing and emerging plant health threats. *EFSA Supporting Publications*, 13(12), 1118E

Arsevska E., Valentin S., Rabatel J., De Goër de Hervé J., Falala S., Lancelot R., Roche M. (2018). Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PloS One*, 13 (8) e0199960.

Barboza P, Vaillant L, Le Strat Y, Hartley DM, Nelson NP, Mawudeku A, et al. (2014) Factors Influencing Performance of Internet-Based Biosurveillance Systems Used in Epidemic Intelligence for Early Detection of Infectious Diseases Outbreaks. *PLoS ONE* 9(3): e90536. https://doi.org/10.1371/journal.pone.0090536

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv*:1903.10676.

Dagan I., Dolan B., Magnini B., and Roth D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.
Doan S., Ngo Q.-H., Kawazoe A., Collier N. (2008) Global Health Monitor. A Web-based System for Detecting and Mapping Infectious Diseases, *Proc. of the International Joint Conference on Natural Language Processing* (IJCNLP), pages 951-956 https://www.aclweb.org/anthology/I08-2140.pdf
European Food Safety Authority (EFSA), Delbianco A, Gibin D, Pasinato L, Boscia D, Morelli M. (2021). Update of the *Xylella* spp. host plant database - systematic literature search up to 31 December 2021. *EFSA J.* 2022 Jun 15;. doi: 10.2903/j.efsa.2022.7356. PMID: 35734284; PMC9198695.
Fan A., Piktus A., Petroni F., Wenzek, G. Saeidi M., Vlachos A., Bordes A., and Riedel S. (2020). Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 7147–7161, Online. Association for Computational Linguistics.
Färber, M., & Rettinger, A. (2013). A semantic wiki for novelty search on documents. In *Proc. of the 13th Dutch-Belgian Workshop on Information Retrieval* (pp. 60-61).
Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., & Bhattacharyya, P. (2022). Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, 48(1), 77-117.
Guo Z., Schlichtkrull M., Vlachos A. (2022) A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*; 10 178–206. doi: https://doi.org/10.1162/tacl_a_00454
Hao, B., Zhu, H., & Paschalidis, I. C. (2020). Enhancing clinical BERT embedding using a biomedical knowledge base. In *28th International Conference on Computational Linguistics* (COLING 2020).
Huttunen, S. (2020). *Information Extraction and linguistic characteristics of texts: exploring scenarios and text types*. PhD dissertation, University of Helsinki, Finland.
Ji, H., & Grishman, R. (2011, June). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1148-1158).
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.
Malembic-Maher, S., Desqué, D., Khalil, D., Salar, P., Bergey, B., Danet, J. L., ... & Foissac, X. (2020). When a Palearctic bacterium meets a Nearctic insect vector: Genetic and ecological insights into the emergence of the grapevine *Flavescence dorée* epidemics in Europe. *PLoS pathogens*, *16*(3), e1007967.
Sarrouti, M., Abacha, A. B., M'rabet, Y., & Demner-Fushman, D. (2021). Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics*: EMNLP 2021 (pp. 3499-3512).
Ouvrard, D. (2022) Psyl'list - The World Psylloidea Database. http://www.hemiptera-databases.com/psyllist - searched on 26 September 2022 doi:10.5519/0029634
Rees, E., Ng, V., Gachon, P., Mawudeku, A., McKenney, D., Pedlar, J., ... & Knox, J. (2019). Early detection and prediction of infectious disease outbreaks. *CCDR*, 45(5).
Abdelrahman, N. S. (2020). *Text Mining for Precision Medicine: Natural Language Processing, Machine Learning and Information Extraction for Knowledge Discovery in the Health Domain* (Doctoral dissertation, Utrecht University).
Sauvion N (2020) Compilation of occurrence data for two psyllid species of the *Cacopsylla pruni* complex (Hemiptera: Psylloidea). 10.15454/VC9UR5, Portail Data INRAE.
Steinberger R, Fuart F, Pouliquen B, Van Der Goot E. (2008) MedISys: A Multilingual Media Monitoring Tool for Medical Intelligence and Early Warning. In Conference Proceedings: Global Risk Forum GRF Davos *Proceedings of the International Disaster and Risk Conference*. Davos (Switzerland). p. 612-614. JRC45523
Thorne J. and Vlachos A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
Vlachos A. and Riedel S. (2015). Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics. https:// doi.org/10.18653/v1/D15-1312
Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 7534–7550, Online. Association for Computational Linguistics.
Wang, S., Durrett, G., & Erk, K. (2018). Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv*:1804.00619.
Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18

Xu, B., Wang, Q., Lyu, Y., Dai, D., Zhang, Y., & Mao, Z. (2023, July). S2ynRE: Two-stage Self-training with Synthetic data for Low-resource Relation Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 8186-8207).

Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Laurent-Walter Goix. 2022. PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22).* Association for Computing Machinery, New York, NY, USA, 561–571. https://doi.org/10.1145/3511808.3557422