PROPOSITION DE STAGE MASTER 2 IEV 2024

Extraction d'Information sur les maladies transmises par vecteurs chez les plantes

Information extraction on vector-borne diseases in plants from the literature

Durée : 5 à 6 mois entre février et août 2024
Lieu : MaIAGE, Centre de recherche de Jouy en Josas, France et visites chez les partenaires
Financement : gratification de stage selon les montants réglementaires
Mots clefs : Traitement Automatique de la Langue, apprentissage automatique, deep learning, extraction d'information, épidémiologie, insecte vecteur

Profil

- Master 2 en INFORMATIQUE orienté Traitement Automatique des Langues et/ou Apprentissage automatique
- Expérience de *deep learning*
- Expérience en TAL et/ou utilisation de la bibliothèque HuggingFace
- Maîtrise de l'anglais ou français courant.
- Compétences techniques requises : Python et/ou Java
- Intérêt pour les applications en biologie et le travail interdisciplinaire.

Candidature

cv, lettre de motivation et relevés de notes L3, M1, M2 sont à adresser aux encadrants :

- Claire Nédellec (équipe Bibliome, unité MaIAGE, INRAE), Jouy-en-Josas. claire.nedellec@inrae.fr
- Vincent Guigue (équipe EkINOCs, unité MIA-Paris-Saclay, AgroParisTech) Palaiseau. vincent.guigue@agrosparistech.fr
- Nicolas Sauvion (équipe FORISK, unité PHIM, INRAE), Montpellier. nicolas.sauvion@inrae.fr

Les équipes Bibliome et EkINocs sont spécialistes du Traitement Automatique de la Langue (TAL) et de l'extraction d'information, l'équipe Forisk est spécialiste des insectes vecteurs.

Sujet du stage

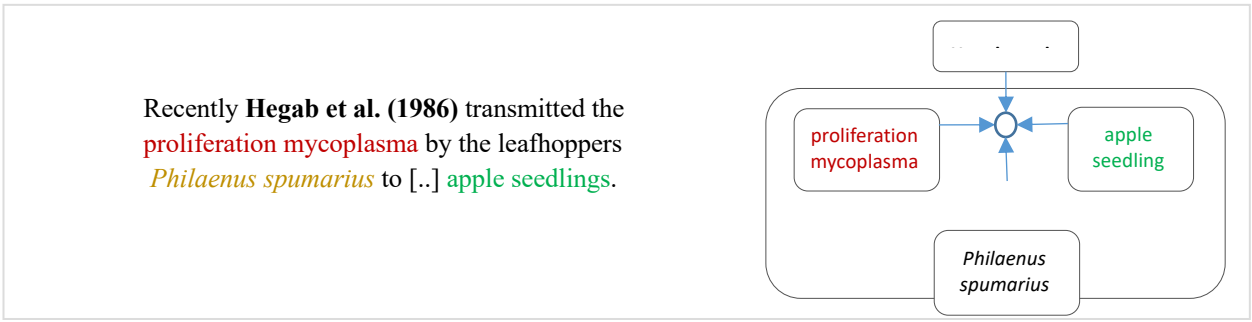
Contexte

Les phytoplasmes sont des bactéries qui causent des maladies d'arbres fruitiers dont les impacts économiques sont très importants en Europe [Hadidi et al., 2011]. Ces bactéries pathogènes s'attaquent à différents types de plantes de la famille des rosacées (Prunus, pommiers et poiriers). Les bactéries peuvent être transmises d'une plante à l'autre par des insectes piqueurs suceurs, des psylles du genre *Cacopsylla*. Ces bactéries et leurs insectes vecteurs sont endémiques en Europe. Ils sont largement présents dans les vergers ainsi que dans les habitats sauvages, ce qui limite leur contrôle et, par conséquent l'endiguement des maladies dont ils sont responsables. Les psylles vecteurs sont aujourd'hui contrôlés principalement par des insecticides, mais l'évolution des pratiques agricoles pourraient être, voire sont déjà, la source de nouvelles émergences de maladies. En effet, la réduction de l'utilisation des pesticides en accord avec le plan EcoPhyto en France et les nouvelles réglementations européennes moins contraignantes en terme de surveillance facilite leur dissémination.
Les efforts de la recherche pour mieux comprendre la biologie et l'écologie des psylles vecteurs (ou potentiellement vecteurs) de phytoplasmes visent à proposer de nouveaux moyens d'anticipation et de contrôle du risque épidémiologique. Malgré ces travaux, la connaissance des interactions biologiques de ces bactéries, insectes et plantes est incomplète et mal établie, notamment en raison du très grand nombre de publications.

Le web a démultiplié les possibilités d'accès aux documents scientifiques y compris très anciens. L'extraction automatique d'informations contenues dans ce type de documents par des méthodes de TAL a fait ses preuves dans de nombreux domaines de la biologie, notamment l'extraction d'entités nommées, leur normalisation et leur mise en relation. Les progrès récents sont considérables grâce aux larges modèles de langue (LLMs) qui ont trouvé de nombreuses applications notamment dans le domaine biomédical. Le domaine de l'écologie, sujet de ce stage, soulève des questions d'intérêt pour la recherche en TAL. Tout d'abord, les interactions biologiques d'intérêt impliquent plusieurs participants, au moins un pathogène, un vecteur et une plante, l'extraction de relations n-aires est donc nécessaire. Les articles reprennent des informations publiées en les citant. Associer la source bibliographique (la référence) à l'information extraite est nécessaire pour caractériser l'information dans la perspective d'en estimer la pertinence.

Objectifs du stage

Le projet de Master porte sur l'extraction automatique de relations biologiques à partir de documents. Le stage ciblera en priorité trois espèces particulières de psylles vecteurs de bactéries pathogènes d'arbres fruitiers. Ce travail s'inscrit dans le cadre plus large d'un projet de thèse, sur la qualité et la nouveauté d'informations épidémiologiques [Nédellec et al. 2024], pour laquelle des candidats étudiants sont également recherchés. Les événements représentant les interactions biologiques entre microbe, insecte, plante et leurs lieux et dates d'observation sont dénotés dans les textes scientifiques par des formulations complexes variables qui portent fréquemment sur plusieurs phrases. L'enjeu sera d'extraire ces événements (voir figure) par des méthodes d'apprentissage profond (*deep learning*) avec un nombre limité d'exemples produits manuellement.



Nous faisons l'hypothèse qu'exploiter la connaissance disponible dans les domaines spécialisés par des LLMs peut pallier le nombre réduit de données d'entraînement annotées. Il s'agit ici de la base de connaissance *Global DataBase* de l'EPPO[1] et *Psyl'list* [Ouvrard, 2022]. La méthode KBPubMedBERT [Tang et al., 2023] pourra être une première solution à explorer, ainsi que des méthodes génératives [Xu et al., 2023], ou semi-supervisée [Genest et al., 2022]. La distance parfois élevée entre les arguments d'événements multiphrases dépasse les limites de modèles de langue (e.g. BERT [Devlin et al., 2019], SciBERT [Beltagy et al., 2019], BioBERT [Lee et al., 2020) et devra faire l'objet de propositions adaptées, par exemple de réseau neuronal de graphe (GNN) pour construire un graphe d'entités et capturer les interactions entre les entités à travers les phrases [Li et al. 2022].

Le rattachement aux événements extraits des sources bibliographiques à travers leur citation est un second objectif du stage. Le rattachement des entités et références a fait l'objet de travaux [Viswanathan et al. 2021]. Il s'agit ici de traiter le rattachement des références à des événements structurés.

Programme

La/le stagiaire réalisera un état de l'art des méthodes existantes d'extraction de relations n-aires et de

[1] https://gd.eppo.int/

citations. Il/Elle adaptera une de ces méthodes au sujet et proposera des extensions originales intégrées dans le workflow ESV. Robert Bossy (éq. Bibliome) formera et accompagnera la/le stagiaire dans l'utilisation d'AlvisNLP. Les prédictions seront évaluées par les méthodes standards du domaine (e.g. F-mesure, rappel, précision). Les entités de type citation feront l'objet d'un traitement particulier portant sur leur extraction et leur rattachement aux événements biologiques. Un article sera préparé en collaboration avec les co-encadrants en fonction des résultats obtenus.

Ressources

Seront mis à disposition les éléments nécessaires à la réalisation des objectifs du stage : (1) le workflow opérationnel ESV[2] sur la plateforme AlvisNLP d'extraction d'information d'entités, de normalisation et d'extraction de relations *binaires*, (2) la base de connaissance *Global DataBase* de l'EPPO, (3) un corpus de documents non annoté d'où les informations sont à extraire, (4) le corpus EPOP (*Epidemiomonitoring Of Plant*) annoté manuellement. Les moyens de calcul GPU du méso-centre de l'Université Paris-Saclay seront utilisés (e.g. Lab.IA).

Publications de l'équipe d'accueil relatives au sujet

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

Dérozier S, Bossy R, Deléger L, Ba M, Chaix E, Harlé O, Loux V., Falentin H., Nédellec C. (2023) Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. *PLoS ONE* 18(1): e0272473. https://doi.org/10.1371/journal.pone.0272473.

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

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 https://doi.org/10.1038/s41598-020-61908-0

MacLeod A. (…), Sauvion N., et al (2012) Pest risk assessment for the European Community plant health: a comparative approach with case studies. Supporting Publications 2012:EN-319 [1053 pp.] Available online: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2012.EN-319

Morris, C.E., Géniaux, G., Nédellec, C., Sauvion, N. & Soubeyrand, S. (2021) 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. https://doi.org/10.1111/ppa.13446

Nédellec C., Guigue V. and Sauvion N., *"Modèles de langue et analyse des informations incertaines : suivi temporel de la fiabilité de la bibliographie sur des insectes vecteurs de pathogènes des plantes."* Sujet de thèse 2024. N° de référence:51658. https://adum.fr/as/ed/voirproposition.pl?site=adumR&matricule_prop=51658

Peccoud J., Pleydell D.R.J., Sauvion N. (2018) A framework for estimating the effects of sequential reproductive barriers: implementation using Bayesian models with field data from cryptic species. *Evolution* 72-11 https://onlinelibrary.wiley.com/doi/pdf/10.1111/evo.13595

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 https://hal.archives-ouvertes.fr/hal-03230951v2

Simon, É., Guigue, V., & Piwowarski, B. (2019, July). Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses. In *Proceedings of the 57th*

[2] https://forgemia.inra.fr/bibliome/pesv-tm

Annual Meeting of the Association for Computational Linguistics (pp. 1378-1387).

Steffek, R., Follak, S., Sauvion, N., Labonne, G., MacLeod, A. (2012) Distribution of '*Candidatus* Phytoplasma prunorum' and its vector *Cacopsylla pruni* in European fruit growing areas: a review. EPPO Bulletin 42: 191-202. https://doi.org/10.1111/epp.2567

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, November). 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, November). 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).

Thébaud G., Yvon M., Alary R., Sauvion N. & Labonne G. (2009) Efficient transmission of '*Candidatus* Phytoplasma prunorum' is delayed by eight months due to a long latency in its host-alternating vector. *Phytopathology* 99: 265-273. https://apsjournals.apsnet.org/doi/10.1094/PHYTO-99-3-0265

Anfu Tang, Louise Deléger, Robert Bossy, Claire Nédellec, Pierre Zweigenbaum. Exploitation de plongements de graphes pour l'extraction de relations biomédicales. *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2023*, Paris, 5 au 9 juin 2023.

Références externes

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

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

Hadidi A, Barba M, Candresse T, Jelkmann W (2011) Virus and virus-like diseases of pome and stone fruits. The American Phytopathological Society Press, St Paul, Minnesota. [ISBN 978-0-89054-396-2] https://doi.org/10.1094/9780890545010

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.

Li, R., Zhong, J., Xue, Z., Dai, Q., & Li, X. (2022). Heterogenous affinity graph inference network for document-level relation extraction. *Knowledge-Based Systems*, *250*, 109146.

Ouvrard, D. (2022) Psyl'list - The World Psylloidea Database. http://www.hemiptera-databases.com/psyllist - searched on 26 September 2022 doi:10.5519/0029634

Viswanathan, V., Neubig, G., & Liu, P. (2021). Citationie: Leveraging the citation graph for scientific information extraction. *arXiv preprint arXiv:2106.01560*.

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).