BIBLIOME : Acquisition et Formalisation de Connaissances à partir de Textes

Robert Bossy, Arnaud Ferré, Équipe Bibliome, Claire Nédellec, Louise Deleger

Équipe Bibliome : Acquisition et Formalisation de Connaissances à partir de Textes

Équipe Bibliome / Unité MaIAGE
Université Paris-Saclay / Centre de recherche INRAE de Jouy-en-Josas
http://maiage.inra.fr/

Claire NÉDELLEC
claire.nedellec@inra.fr

Robert BOSSY
robert.bossy@inra.fr

Louise DELÉGER
louise.deleger@inra.fr

Arnaud FERRÉ
arnaud.ferre@inra.fr

Domaine de recherche

L'équipe Bibliome développe des méthodes d'extraction et de formalisation d'information à partir de textes écrits. Ces méthodes identifient et formalisent des informations et connaissances précises dans de larges corpus de documents de genres divers et les mettent en relation, faisant appel à des méthodes de traitement automatique de la langue et d'apprentissage automatique. Les principaux travaux concernent trois sujets,

1. L'apprentissage automatique pour la reconnaissance et la formalisation d'entités et de relations ;
2. La conception de terminologies et d'ontologies ;
3. L'intégration et l'évaluation des méthodes dans une infrastructure partagée.

Nos recherches sont guidées par des besoins applicatifs qui permettent de valider nos méthodes et d'identifier les objectifs prioritaires dans des domaines variés de la biologie, microbiolologie, génétique et phénotypes des plantes et des animaux d'élevage.

Méthodes développées

Les méthodes en Intelligence Artificielle développées par l'équipe Bibliome traitent deux étapes clés, l'extraction et l'annotation des entités du texte par des concepts d'ontologie et l'extraction de relations formelles entre ces entités. Pour étudier des phénomènes scientifiques en sciences du vivant dispersés dans une grande quantité de documents, nos travaux ont pour objectif de compenser le petit nombre d'occurrences par des approches dites *knowledge intensive*, combinant analyse linguistique computationnelle, connaissance du domaine sous forme de lexiques et d'ontologie et apprentissage automatique, facilitant ainsi la généralisation des méthodes et leur adaptation à de nouvelles questions.

Par exemple, l'équipe Bibliome développe la méthode HONOR [1] qui intègre deux méthodes complémentaires pour la détection et le rattachement de termes du texte à des concepts d'une ontologie. La méthode ToMap [2] exploite la structure syntaxique et les similarités de forme des termes. La méthode CONTES [3] associe par apprentissage automatique les représentations vectorielles (*embeddings*) et la structure hiérarchique des ontologies. Nos méthodes pour l'extraction de relation combinent analyse linguistique profonde (résolution d'anaphore et dépendances syntaxiques) et méthodes d'apprentissage à noyau (*shortest path dependency kernel*) [4].

Domaine d'application

Nos domaines d'application en science de la vie, agriculture et alimentation sont variés par exemple, microbiologie [5], biologie végétale [6] et animale [7] sur des thèmes divers tels que la régulation génétique [8], la biodiversité microbienne [9], les phénotypes [10], l'épidémiologie végétale, santé humaine [11] et l'analyse bibliométrique [12]. Nos projets applicatifs en extraction d'information suivent un schéma récurrent : définir un modèle pour la représentation formelle des informations, construire un corpus pertinent de documents scientifiques, adapter ou concevoir les nomenclatures,

terminologies et ontologies nécessaires, annoter manuellement les corpus de référence, concevoir des workflows d'entraînement et de prédiction d'entités et de relations, puis lier les prédictions à des données de référence du domaine d'application.

Construction de ressources sémantiques partagées

Nous publions les ressources sous licence ouverte, principalement des corpus annotés (BioNLP-ST[1]) et des ontologies (AgroPortal[2]). Les corpus de référence annotés manuellement sont nécessaires pour entraîner et évaluer des méthodes d'extraction d'information dans les domaines spécialisés de l'INRA où elles sont rares ou inexistantes.

Nous concevons également des modèles formels et des ontologies qui permettent de normaliser les informations extraites du texte et les rattacher ensuite à des données issues d'autres sources dans un cadre de *Linked Open Data*.

Nos projets de construction de ressources, corpus et ontologies, sont mis en oeuvre grâce aux outils logiciels collaboratifs que nous développons et qui favorisent les échanges entre les participants avec des compétences diverses : biologie, traitement automatique de la langue, information scientifique et technique et ingénierie de la connaissance. Nous valorisons les corpus annotés et ontologie dans l'organisation régulière de Shared Task internationaux (BioNLP Open Shared Task) [13].

Développement logiciel

L'équipe développe la suite logicielle Alvis de conception de workflow de text mining à partir d'outils et de contenus pour l'extraction d'information. Elle facilite la mise en place d'expériences, la reproductibilité, la mutualisation des résultats au sein de l'équipe et le transfert. Nous contribuons à l'infrastructure européenne OpenMinTeD[3] de text mining, en particulier sur le volet interopérabilité avec l'apport d'une bibliothèque d'outils de Traitement Automatique de la Langue (AlvisNLP[4]) et services pour les sciences de la vie. Les services associés d'annotation (AlvisAE [14]), de visualisation et de recherche d'information (AlvisIR[5]) permettent de visualiser et de communiquer les résultats des traitements aux applications tierce comme l'application Florilege[6].

Projets

Le projet H2020 OpenMinTeD d'infrastructure de text mining fait suite aux projets FP6 Alvis et BPI Quaero[7] pour le développement d'un environnement de développement d'outils et de service de text mining pour les spécialistes et non spécialistes. Notre participation au projet ANR D2KAB[8] approfondit ce thème à travers l'adaptabilité des méthodes de text mining à différents besoins et domaines et l'intégration avec des données hétérogènes impliquant des alignements sémantiques pour l'implémentation des principes FAIR dans un contexte de Science Ouverte.

Science Ouverte

L'équipe y participe activement à travers son implication dans les e-infrastructures ouvertes (projets H2020 OpenMinTeD et CoSO Visa TM[9]) et à des groupes de travail nationaux sur l'ouverture des publications au text mining[10]. Notre objectif est de faciliter l'appropriation des technologies de text mining pour la recherche scientifique dans une perspective de Science Ouverte permettant la mutualisation des ressources et la reproductibilité des résultats.

Références

[1] Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum, Claire Nédellec. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology, *The 11th international conference on Language Resources*

1 https://2019.bionlp-ost.org

2 http://agroportal.lirmm.fr

3 http://openminted.eu

4 https://bibliome.github.io/alvisnlp/

5 https://github.com/Bibliome/alvisir

6 http://migale.jouy.inra.fr/Florilege/

7 http://www.quaero.org

8 http://d2kab.mystrikingly.com

9 https://visatm.inist.fr

10 https://www.ouvrirlascience.fr/guide-dapplication-de-la-loi-numerique/

and Evaluation (LREC-2018), European Language Resources Association (ELRA) publisher, Miyazaki, mai 2018.

[2] Zorana Ratkovic, Wiktoria Golik, Pierre Warnier, "Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach". *BMC Bioinformatics* 2012, 13(Suppl 11):S8, 26 June 2012.

[3] Arnaud Ferré, Pierre Zweigenbaum, Claire Nédellec, Representation of complex terms in a vector space structured by an ontology for a normalization task. In *Proceedings of the BioNLP 2017 Workshop*, Association for Computational Linguistics, Vancouver, 8 pages, Canada 2017.

[4] Dialekti Valsamou, *Information Extraction for the Seed Development Regulatory Networks of* Arabidopsis thaliana. Thèse de doctorat en Informatique, ED STIC, Université Paris-Sud, 17 janvier 2107.

[5] Estelle Chaix, Louise Deléger, Robert Bossy, Claire Nédellec "Text mining tools for extracting information about microbial biodiversity in food" *Food Microbiology*, 2018.

[6] Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Deléger, Pierre Zweigenbaum, Philippe Bessières, Loïc Lepiniec, Claire Nédellec. Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task. In *Proceedings of the BioNLP Shared Task 2016 Workshop*, Association for Computational Linguistics, Berlin, Allemagne 2016.

[7] Pierre-Yves Le Bail, Jérôme Bugeon, Olivier Dameron, Alice Fatet, Wiktoria Golik, Jean-François Hocquette, Catherine Hurtaud, Isabelle Hue, C. Jondreville, Léa Joret, Marie-Christine Meunier-Salaün, Jean Vernet, Claire Nédellec, Mathieu Reichstadt, Philippe Chemineau. Un langage de référence pour le phénotypage des animaux d'élevage : l'ontologie ATOL, *INRA Prod. Anim*., 2014, 27 (3), 195-208.

[8] Robert Bossy, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Maarten van de Guchte, Philippe Bessières, Claire Nédellec "BioNLP Shared Task - The Bacteria Track". *BMC Bioinformatics* 13(Suppl 11):S3, juin 2012. 10.1186/1471-2105-13-S11-S3.

[9] Claire Nédellec, Robert Bossy, Estelle Chaix, Louise Deléger. Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. In *Proceedings of the 4th International Microbial Diversity Conference*. pp. 221-227, ed. Marco Gobetti. Baris, Pub. Simtra. ISBN 978-88-943010-0-7, Bari, October 2017. arXiv:1805.04107

[10] Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktoria Golik, Pierre Sourdille. Information Extraction from Bibliography for Marker Assisted Selection in Wheat. In *proceedings of Metadata and Semantics for Agriculture, Food & Environment (AgroSEM'14),* special track of the 8th Metadata and Semantics Research Conference (MTSR'14), Springer Communications in Computer and Information Science, Series Volume 478, Karlsruhe, pp 301-313, Allemagne, 2014. DOI: 10.1007/978-3-319-13674-5_28

[11] Leornardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat et Aurélie Névéol. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Lang Resources & Evaluation* 52, 571–601 (2018) doi:10.1007/s10579-017-9382-y

[12] Pascale Avril, Emilie Bernard, Maryse Corvaisier, Agnès Girard, Wiktoria Golik, Claire Nédellec, Marie-Laure Touzé, Nathaële Wacrenier, Analyser la production scientifique d'un département de recherche : construction d'une termino-ontologie par des documentalistes, p 1-12, *Cahier des Techniques*, INRA février 2017.

[13] Kim Jin-Dong, Nédellec Claire, Bossy Robert, Deléger Louise. *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks 2019*, EMNLP-IJCNLP 2019, Hong-Kong, nov 2019.

[14] Frédéric Papazian, Robert Bossy, Claire Nédellec, « AlvisAE: a collaborative Web text annotation editor for knowledge acquisition », *6th Linguistic Annotation Workshop (The LAW VI),* Jeju, Corée, 2012.