Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Projet NaviTerm : navigation terminologique pour une montée en compétence rapide et personnalisée sur un domaine de recherche

Florian Boudin Richard Dufour Béatrice Daille
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
prenom.nom@univ-nantes.fr

RÉSUMÉ

Cet article présente le projet NaviTerm dont l'objectif est d'accélérer la montée en compétence des chercheurs sur un domaine de recherche par la création automatique de représentations terminologiques synthétiques et navigables des connaissances scientifiques.

ABSTRACT

The NaviTerm project : terminological navigation for assisting researchers in gaining expertise on a new research domain.

This article presents the NaviTerm project, whose objective is to assist scholars in gaining new skills in a field of research by automatically producing concise and navigable terminological representations of scientific knowledge.

MOTS-CLÉS : Ici une liste de mots-clés en français.

KEYWORDS: Here a list of keywords in English.

1 Introduction

L'explosion de la production scientifique mondiale amène les chercheurs à revoir en profondeur leur démarche de veille scientifique. Malgré l'émergence de moteurs de recherche dédiés (e.g. Google Scholar, Semantic Scholar, PubMed), parcourir la littérature pour monter en compétence sur un domaine de recherche est de plus en plus laborieux et chronophage. À cela s'ajoute l'opacité des algorithmes utilisés par ces moteurs de recherche qui impose de s'interroger sur la pertinence des résultats retournés (Martín-Martín *et al.*, 2018). Faciliter l'accès aux nouvelles connaissances scientifiques de manière efficace et transparente est donc, plus que jamais, un enjeu majeur pour la recherche scientifique. Cet article présente le projet NaviTerm, financé dans le cadre d'un accord entre le CNRS et l'Agence de l'Innovation et de la Défense (AID), qui apporte une réponse à cet enjeu sous l'angle de la terminologie et de la navigation documentaire.

Plus précisément, le projet NaviTerm porte sur le développement de méthodes automatisées pour extraire, structurer et ordonner par importance les termes d'une collection d'articles scientifiques relevant d'un domaine de recherche. Ces termes dressent une cartographie des connaissances scientifiques et constituent une interface d'accès naturel efficace au contenu des articles. Associés à une méthodologie de recherche à facettes ou à une interface de navigation, ils offrent un moyen intuitif et rapide de repérer les articles clés d'un domaine. Pour illustrer ce point, un exemple de navigation par termes-clés pour l'accès aux connaissances scientifiques est présenté dans la Figure 1. Cet exemple

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 16–20.

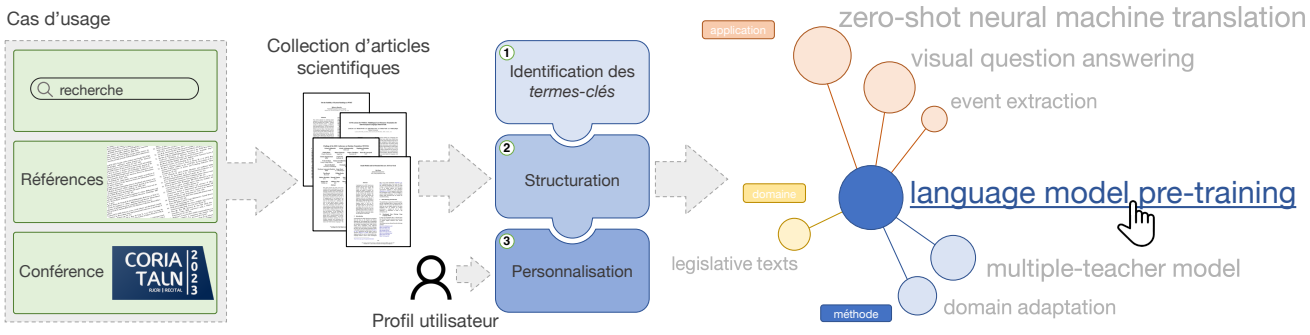

FIGURE 1 – Illustration du processus de création de représentation terminologique navigable à partir d'une collection d'articles scientifiques. Trois cas d'usage sont présentés : monter en compétence à partir des résultats d'une recherche, à partir des références citées dans un article, ou à partir des actes d'une conférence.

permet également de mettre en avant un élément important du projet NaviTerm qui est la structuration automatique des termes, ici en fonction des catégories [application , domaine , methode]. L'objectif de cette structuration est de faciliter la navigation dans la collection d'articles scientifiques et de proposer des listes filtrées et priorisées des articles clés.

L'objectif central du projet NaviTerm est le développement et la mise à disposition de nouveaux outils de recherche bibliographique qui intègrent les dernières avancées du TAL. La montée en compétence scientifique et technique est une problématique transversale et interdisciplinaire, dont les applications potentielles dans l'académique et l'industrie sont nombreuses. Le projet NaviTerm se distingue de l'existant sur deux aspects : d'abord d'un point de vue méthodologique en s'éloignant du paradigme dominant de l'apprentissage automatique supervisé de bout-en-bout au profit d'approches faiblement ou non supervisées de l'état-de-l'art, associées à des algorithmes facilement interprétables ; ensuite d'un point de vue applicatif avec la personnalisation des représentations terminologiques et des résultats de recherche pour accélérer la montée en compétences des utilisateurs. La section suivante rappelle les objectifs du projet et présente les trois verrous scientifiques que nous chercherons à lever.

2 Verrous scientifiques et solutions envisagées

Le projet NaviTerm s'attaque à la problématique de plus en plus prégnante de l'accès rapide aux connaissances scientifiques dans le contexte actuel d'explosion du nombre des publications. Ses objectifs sont doubles : 1) la création automatisée de représentations terminologiques navigables des connaissances scientifiques d'un domaine ; et 2) la personnalisation des représentations produites pour accélérer et maîtriser la montée en compétence des utilisateurs. Pour cela, nous tenterons de lever les trois verrous scientifiques suivants :

1. **Identifier les termes « clés » d'une collection de documents** ; les méthodes d'extraction terminologique existantes permettent d'identifier avec précision les termes spécialisés relevant d'un domaine mais n'évaluent pas leur importance vis-à-vis de ce dernier (Hazem *et al.*, 2020). Pour cela, nous proposons d'étendre ces méthodes avec des techniques non supervisées d'ordonnancement de texte utilisées pour l'extraction de mots-clés (Mihalcea & Tarau, 2004; Bougouin *et al.*, 2013). Il s'agira d'explorer comment ces techniques peuvent être étendues du

niveau de granularité du document à celui de la collection et d'appréhender les problématiques qui en découlent (e.g. passage à l'échelle, évaluation). Nous motivons ce choix méthodologique par une volonté de transparence et de confiance dans les algorithmes utilisés pour le projet.

2. **Structurer les connaissances terminologiques** ; les termes « clés » retenus devront être catégorisés (e.g. application, méthode, ensemble de données), référencés (i.e. identifiant (DOI) et position(s) d'occurrence dans l'article) et structurés (e.g. associer un ensemble de données avec une application). La solution envisagée pour cela est d'explorer comment les récentes méthodes de co-apprentissage (joint learning) fondées sur des réseaux convolutifs sur graphes (Sun *et al.*, 2019) peuvent être adaptées dans un cadre faiblement supervisé pour la détection de relations terminologiques entre termes « clés ». Dans un second temps, il s'agira d'étudier comment inférer les catégories et les relations entre les termes pour permettre la transférabilité à d'autres domaines et types de données. A cet égard, les travaux actuels sur l'extraction non-supervisée de relations (Tran *et al.*, 2020; Yuan & Eldardiry, 2021) constituent une piste de recherche intéressante.

3. **Personnaliser les connaissances d'un domaine** ; les représentations terminologiques construites à partir des termes « clés » retenus devront être adaptées pour permettre une montée en compétences plus efficace et rapide. Nous explorerons des stratégies de filtrage (ou de tri) pour mettre en exergue les connaissances importantes selon différents prismes : par rapport à un profil utilisateur (e.g. une bibliographie personnelle) ; par rapport à un événement scientifique (e.g. une conférence du domaine), par rapport à une temporalité (e.g. 5 dernières années) ; par rapport à la saillance d'un événement (e.g. information récente qui prend de l'ampleur, approche récurrente dans les documents). Pour cela, nous envisageons d'adapter le fonctionnement des méthodes neuronales de recommandation de tags (Hassan *et al.*, 2018) à notre problématique, puis dans une forme très exploratoire d'étudier comment personnaliser l'interface de navigation avec comme point de départ les travaux sur les interfaces de navigation exploratoire par mots-clés (Shukla & Hoeber, 2021) et les interfaces de visualisations selon une échelle temporelle, à la manière de graphes dynamiques.

3 Discussion et perspectives

Monter en compétence sur un domaine de recherche est une tâche de plus en plus complexe, la faute à un volume de littérature scientifique en pleine croissance. Accéder aux articles scientifiques les plus pertinents d'un domaine de recherche suppose une connaissance préalable de sa terminologie, ce qui n'est à l'évidence pas le cas des chercheurs qui souhaitent se former. L'objectif du projet NaviTerm est de lever cette difficulté en offrant un accès direct aux articles scientifiques au travers de représentations terminologiques construites automatiquement. L'accès au contenu des articles ne constitue évidemment qu'une première étape dans le chemin vers la montée en compétence et soulève de nouvelles questions de recherche. Par exemple, comment faciliter et accélérer la lecture des articles scientifiques (Head *et al.*, 2021; Fok *et al.*, 2022) ou comment construire un parcours pour acquérir des connaissances sont autant de questions qu'il conviendra d'examiner à l'avenir.

Remerciements

Ce travail est financé dans cadre du projet AID-CNRS NaviTerm (convention 2022 65 0079 CNRS Occitanie Ouest).

Références

BOUGOUIN A., BOUDIN F. & DAILLE B. (2013). TopicRank : Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 543–551, Nagoya, Japan : Asian Federation of Natural Language Processing.

FOK R., KAMBHAMETTU H., SOLDAINI L., BRAGG J., LO K., HEARST M. A., HEAD A. & WELD D. S. (2022). Scim : Intelligent skimming support for scientific papers. *arXiv preprint arXiv :2205.04561*.

HASSAN H. A. M., SANSONETTI G., GASPARETTI F. & MICARELLI A. (2018). Semantic-based tag recommendation in scientific bookmarking systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, p. 465–469, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3240323.3240409.

HAZEM A., BOUHANDI M., BOUDIN F. & DAILLE B. (2020). TermEval 2020 : TALN-LS2N system for automatic term extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*, p. 95–100, Marseille, France : European Language Resources Association.

HEAD A., LO K., KANG D., FOK R., SKJONSBERG S., WELD D. S. & HEARST M. A. (2021). Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, p. 1–18.

MARTÍN-MARTÍN A., ORDUNA-MALEA E., THELWALL M. & DELGADO LÓPEZ-CÓZAR E. (2018). Google scholar, web of science, and scopus : A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, **12**(4), 1160–1177. DOI : https ://doi.org/10.1016/j.joi.2018.09.002.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

SHUKLA S. & HOEBER O. (2021). Visually linked keywords to support exploratory browsing. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, p. 273–277, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3406522.3446037.

SUN C., GONG Y., WU Y., GONG M., JIANG D., LAN M., SUN S. & DUAN N. (2019). Joint type inference on entities and relations via graph convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1361–1370, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1131.

TRAN T. T., LE P. & ANANIADOU S. (2020). Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7498–7505, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.669.

YUAN C. & ELDARDIRY H. (2021). Unsupervised relation extraction : A variational autoencoder approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 1929–1938, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.147.