18e Conférence en Recherche d'Information et Applications
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues
(CORIA-TALN) [1]

Actes de CORIA-TALN 2023.

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023

Florian Boudin, Béatrice Daille, Richard Dufour, Oumaima El Khettari, Maël Houbre, Léane Jourdan, Nihel Kooli (Éds.)

Paris, France, 5 au 9 juin 2023

1. https://coria-taln-2023.sciencesconf.org/

Avec le soutien de

Préface

Le nombre d'articles scientifiques produits chaque année ne cesse d'augmenter. Rien que dans l'archive ouverte `arXiv`, le nombre d'articles scientifiques déposés en 2022 s'élève à plus de 185 000, soit près de 500 dépôts chaque jour [2]. Face à cette explosion du volume de littérature scientifique, des solutions intelligentes sont nécessaires pour faciliter la recherche et la lecture des articles scientifiques et pour en analyser le contenu et y extraire des informations utiles aux chercheurs et aux applications qui les utilisent. De plus, l'avènement de la science ouverte et la disponibilité croissante des textes intégraux soulèvent de nouveaux enjeux pour le traitement automatisé des articles scientifiques et interroge sur l'utilisabilité des modèles de langues actuels. Comment analyser et rendre accessible les informations contenues dans les tables, les équations ou les figures sont autant de questions qui doivent être explorées.

L'atelier sur l'Analyse et la Recherche de Textes Scientifiques (ARTS) [3], qui se déroule le 5 juin 2023 pendant la conférence CORIA-TALN à Paris, se veut un lieu de rencontre et d'échange pour les chercheurs en Recherche d'Information (RI) et en Traitement Automatique des Langues (TAL) qui s'intéressent aux textes scientifiques. Douze communications écrites ont été acceptées, puis présentées sous la forme d'un poster pendant l'atelier.
Les travaux présentés portent sur une diversité de problématiques, allant de l'annotation et de la collecte de corpus, à la classification de documents, à la traduction automatique, ou encore la simplification de textes scientifiques.

Nous adressons des remerciements particuliers au conférencier invité, Mathieu Constant (ATILF, Université de Lorraine), qui nous a fait le plaisir de présenter ses travaux sur la *construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX*.

Le comité d'organisation remercie également l'ensemble des contributeurs à l'atelier, la diversité des travaux de recherche présentés montrant l'intérêt important, et croissant, que suscite ce domaine de recherche ouvert.

Nous souhaitons enfin remercier chaleureusement l'ensemble des membres du comité de programme scientifique pour leur aide importante quant à la relecture et à la sélection des papiers.

L'atelier ARTS est soutenu par le projet DGA-CNRS NaviTerm [4] (convention 2022 65 0079 CNRS Occitanie Ouest) ayant pour objectif d'accélérer la montée en compétence des chercheurs par la création automatisée de représentations navigables des connaissances scientifiques.

2. https://arxiv.org/stats/main
3. https://arts2023.sciencesconf.org
4. https://cnrs-naviterm.github.io/

Comités

Comité scientifique

— Sabine Barreaux (INIST, CNRS)
— Guillaume Cabanac (IRIT, Université Toulouse 3)
— Florian Boudin (LS2N, Nantes Université)
— Mathieu Constant (ATILF, Université de Lorraine)
— Béatrice Daille (LS2N, Nantes Université)
— Richard Dufour (LS2N, Nantes Université)
— Natalia Grabar (STL, Université de Lille)
— Thierry Hamon (LISN, Université Sorbonne Paris Nord)
— Evelyne Jacquey (ATILF, CNRS)
— Cyril Labbé (LIG, Université Grenoble Alpes)
— François Yvon (LISN, CNRS)

Comité d'organisation

— Florian Boudin (LS2N, Nantes Université)
— Béatrice Daille (LS2N, Nantes Université)
— Richard Dufour (LS2N, Nantes Université)
— Oumaima El Khettari (LS2N, Nantes Université)
— Maël Houbre (LS2N, Nantes Université)
— Léane Jourdan (LS2N, Nantes Université)
— Nihel Kooli (DGA)

Présentation invitée

Mathieu Constant (ATILF, Université de Lorraine)

Titre : Construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX

Résumé : La plateforme ISTEX (https://www.istex.fr/) permet d'accéder à une large base d'archives scientifiques comptant plus de 25 millions de documents de tous les grands domaines scientifiques. Les documents incluent non seulement les métadonnées mais aussi le texte plein, et ont été prétraités de manière homogène pour faciliter leur traitement automatique. Dans cet exposé, nous présenterons une initiative pour créer une dynamique de recherche en TAL et TDM autour de ces données. En particulier, nous présenterons les travaux en cours pour la construction d'un jeu de données dédié au TAL et la fouille de textes.

Table des matières

La pré-annotation automatique de textes cliniques comme support au dialogue avec les experts du domaine lors de la mise au point d'un schéma d'annotation **1**
Virgile Barthet, Marie-José Aroulanda, Laura Monceaux-Cachard, Christine Jacquin, Cyril Grouin, Johann Gutton, Guillaume Hocquet, Pascal De Groote, Michel Komajda, Emmanuel Morin, Pierre Zweigenbaum

MaTOS : Traduction automatique pour la science ouverte **8**
Maud Bénard, Alexandra Mestivier, Natalie Kubler, Lichao Zhu, Rachel Bawden, Eric De La Clergerie, Laurent Romary, Mathilde Huguin, Jean-François Nominé, Ziqian Peng, François Yvon

Projet NaviTerm : navigation terminologique pour une montée en compétence rapide et personnalisée sur un domaine de recherche **16**
Florian Boudin, Richard Dufour, Béatrice Daille

Annotation d'interactions hôte-microbiote dans des articles scientifiques par similarité sémantique avec une ontologie **21**
Oumaima El Khettari, Solen Quiniou, Samuel Chaffron

Quand des Non-Experts Recherchent des Textes Scientifiques Rapport sur l'action CLEF 2023 SimpleText **27**
Liana Ermakova, Stéphane Huet, Eric Sanjuan, Hosein Azarbonyad, Olivier Augereau, Jaap Kamps

Apprentissage de dépendances entre labels pour la classification multi-labels à l'aide de transformeurs **34**
Haytame Fallah, Elisabeth Murisasco, Emmanuel Bruno, Patrice Bellot

Elaboration d'un corpus d'apprentissage à partir d'articles de recherche en chimie **41**
Bénédicte Goujon

Classification de relation pour la génération de mots-clés absents **47**
Maël Houbre, Florian Boudin, Béatrice Daille

Le corpus « Machine Translation » : une exploration diachronique des (méta)données Istex **54**
Mathilde Huguin, Sabine Barreaux

CASIMIR : un Corpus d'Articles Scientifiques Intégrant les ModIfications et Révisions des auteurs **60**
Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez

MORFITT : Un corpus multi-labels d'articles scientifiques français dans le domaine biomédical **66**
Yanis Labrak, Mickael Rouvier, Richard Dufour

La détection de textes générés par des modèles de langue : une tâche complexe ? Une étude sur des textes académiques **71**
Vijini Liyanage, Davide Buscaldi

Construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX **79**

Constant Mathieu

What shall we read : the article or the citations ? - A case study on scientific language understanding **80**
Aman Sinha, Sam Bigeard, Marianne Clausel, Mathieu Constant