Traitement Automatique des Langues Naturelles (TALN) [1]

Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles.

Volume 1 : conférence principale

Yannick Estève, Tania Jiménez, Titouan Parcollet, Marcely Zanon Boito (Éds.)

Avignon, France, 27 juin au 1er juillet 2022

1. https://taln2022.univ-avignon.fr

Avec le soutien de

Préface

Bienvenue à TALN-RECITAL 2022

La 29ème édition de la conférence TALN et 24ème édition du RECITAL se déroulent cette année à Avignon, organisée conjointement par le Laboratoire d'Informatique et Systèmes (LIS) et le Laboratoire Informatique d'Avignon (LIA) sous l'égide de l'Association pour le Traitement Automatique des Langues (ATALA).

La thématique spéciale de la conférence est : « Vers un TAL inclusif. »

Nous accueillons les participants sur le campus Hannah Arendt d'Avignon Université du lundi 27 juin au vendredi 1er juillet 2022.

Après deux années minées par les contraintes sanitaires, nous sommes heureux de pouvoir organiser cet événement en mode présentiel, avec plus de 150 participants.

Le premier jour, lundi, est consacré à deux ateliers thématiques : « DEFT Fouille de texte » dédié cette année à la correction automatique de copies d'étudiants ; et la première édition de l'atelier « TAL et Humanités Numériques » qui applique les thèmes du TAL aux problématiques des corpus SHS.

Ce même jour se tiendra le tutoriel « Quelques étapes souvent omises dans la préparation de corpus ». Nous remercions les organisateurs de ces ateliers et du tutoriel pour leurs propositions, leur animation et l'organisation de ces événements.

Les quatre jours qui suivent associent dans des séances communes la conférence TALN et la conférence jeunes chercheurs RECITAL.

Le dernier jour, vendredi 1er juillet, nous avons le plaisir d'accueillir une table ronde autour de laquelle industriels et académiques échangeront autour des thèmes de la conférence. À la suite de cette table ronde, les démonstrations et posters permettent de continuer ses échanges.

Nous accueillons deux conférencières invitées : Seza Doğruöz, Professeure de la Faculté de philosophie et lettres à l'Université de Gand et Teresa Lynn, Research Fellow à The ADAPT Centre, Dublin City University. Nous sommes très fiers qu'elles aient accepté notre invitation.

TALN2022 et RECITAL2022 auront permis aux chercheurs de présenter leurs travaux sous forme de communications orales, de posters et de démonstrations. 62 papiers au total ont été soumis à TALN, 47 papiers (76%) ont été acceptés : 35 en présentation orale et 12 en poster. En ce qui concerne la conférence RECITAL, 15 articles ont été soumis, et 9 articles ont été acceptés.

Nous tenons à remercier chaleureusement les membres des comités de programme de TALN et de RECITAL, qui ont fait un travail formidable, ainsi que le comité permanent de la conférence (CPERM) et son président pour l'aide dans l'organisation de l'événement et vos conseils éclairés.

Un grand merci à nos sponsors : Avignon Université, Aix-Marseille Université, NAVER LABS, MOBIDYS et ORKIS, sans lesquels cette édition de TALN - RECITAL n'aurait pas pu avoir lieu.

Yannick Estève, Tania Jiménez, Titouan Parcollet, Marcely Zanon Boito

Comités

TALN 2022 est organisé conjointement par le Laboratoire d'Informatique et Systèmes (LIS) et le Laboratoire Informatique d'Avignon (LIA) sous l'égide de l'Association pour le Traitement Automatique des Langues (ATALA).

Comité local d'organisation

— Yannick Estève, LIA
— Tania Jiménez, LIA
— Teva Merlin, LIA
— Antoine Caubrière, LIA
— Arthur Amalvy, LIA
— Bassam Jabaian, LIA
— Carlos Ramisch, LIS
— Corinne Fredouille, LIA
— Gaëlle Laperrière, LIA
— Jarod Duret, LIA
— Juan-Manuel Torres-Moreno, LIA
— Marcely Zanon Boito, LIA
— Natalia Tomashenko, LIA
— Paul-Gauthier Noé, LIA
— Pierre Jourlin, LIA
— Salima Mdhaffar, LIA
— Stéphane Huet, LIA
— Sylvie Ros, LIS
— Titouan Parcollet, LIA
— Jean-François Bonastre, LIA

Comité de programme

— Benoit Favre, Aix-Marseille Université
— Caio Corro, Université Paris-Saclay
— Caroline Brun, Naver Labs Europe
— Delphine Bernhard, LiLPa, Université de Strasbourg
— Géraldine Damnati, Orange Labs
— Guillaume Wisniewski, LLF, Université de Paris
— Magalie Ochs, LSIS
— Marie Candito, Université Paris 7 / INRIA
— Maud Ehrmann, EPFL, DHLAB
— Natalia Grabar, STL CNRS Université Lille 3
— Philippe Langlais, Université de Montréal
— Philippe Muller, IRIT, Université Toulouse
— Thomas François, Université Catholique de Louvain
— Yannick Estève, LIA
— Yannick Parmentier, LORIA — Université de Lorraine

Comité de Lecture TALN

— Andon Tchechmedjiev, EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales
— Aurélie Névéol, Université Paris-Saclay, CNRS, LISN

— Béatrice Daille, Laboratoire d'Informatique Nantes Atlantique (LINA)
— Cédric Lopez, Emvista
— Céline Alec, Université de Caen-Normandie
— Charles Teissèdre, Synapse Développement
— Christine Jacquin, LS2N
— Cyril Grouin, LIMSI, CNRS, Université Paris-Saclay
— Damien Nouvel, INaLCO
— Denis Maurel, University Francois Rabelais Tours
— Didier Schwab, Université Grenoble Alpes
— Dominique Estival, Western Sydney University
— Emmanuel Morin, Université de Nantes, LS2N
— Eric Laporte, Université Gustave Eiffel
— François Yvon, LIMSI/CNRS et Université Paris-Sud
— Gaël Dias, Normandie University
— Gaël Guibon, Télécom Paris - SNCF
— Gwénolé Lecorvé, Orange
— Jean-Philippe Prost, Aix-Marseille Université
— Jian-Yun Nie, Université de Montréal
— Juan-Manuel Torres-Moreno, Laboratoire Informatique d'Avignon — Avignon Université
— Karën Fort, Sorbonne Université
— Laurent Besacier, Laboratoire d'Informatique de Grenoble
— Ludovic Tanguy, CLLE-ERSS
— Luka Nerima, Université de Genève
— Mathieu Dehouck, INRIA
— Matthieu Constant, Université de Lorraine, ATILF, CNRS
— Matthieu Labeau, Telecom Paris
— Maxime Amblard, Université de Lorraine
— Maximin Coavoux, CNRS, Université Grenoble Alpes
— Mikaela Keller, Université Lille 3 - Inria
— Nabil Hathout, CNRS
— Nathalie Camelin, LIUM — Université du Maine
— Nicolas Dugué, LIUM — Université du Maine
— Nicolas Hernandez, Université de Nantes — LINA CNRS UMR 6241
— Olivier Ferret, CEA List
— Olivier Hamon, Syllabs
— Pascal Amsili, Sorbonne Nouvelle
— Pascale Sébillot, IRISA
— Patrice Bellot, Aix-Marseille Université - CNRS (LSIS)
— Peggy Cellier, IRISA/INSA Rennes
— Philippe Blache, CNRS & Université de Provence
— Remi Cardon, CENTAL
— Richard Moot, CNRS (LIRMM) & Université de Montpellier
— Salima Mdhaffar, LIUM
— Solen Quiniou, LINA — Université de Nantes
— Stéphane Huet, Laboratoire Informatique d'Avignon — Avignon Université
— Sylvain Kahane, Modyco, Université Paris Ouest Nanterre & CNRS
— Sylvain Pogodalla, LORIA/INRIA Lorraine
— Thierry Charnois, LIPN CNRS University of PARIS 13
— Thierry Hamon, Université Paris-Saclay, CNRS, LIMSI & Université Sorbonne Paris Nord
— Véronique Moriceau, IRIT-CNRS
— Vincent Claveau, IRISA - CNRS

— Xavier Tannier, Sorbonne Université, INSERM, LIMICS
— Yves Bestgen, F.R.S-FNRS et UCL
— Yves Lepage, Waseda University

Présentations invitées

• A. Seza Doğruöz

Professeure — Faculté de philosophie et lettres, Université de Gand

Titre / Title : "Multilingualism 101" for Computational Linguists : Challenges & Opportunities for Multilingual & Inclusive Language Technologies

Résumé / Abstract : A multilingual is someone who speaks more than one language for communication in his/her daily life. Multilingualism is commonly practiced at the individual or societal levels in most parts of the world (e.g., Europe, Asia, Africa). However, current language technologies are usually built around monolingual assumptions which do not always reflect the linguistic variation in real world communication. The first goal of this talk is to familiarize the computational linguistics community with the key concepts around multilingualism. Secondly, I will discuss the challenges and opportunities for building linguistically diverse, inclusive and multilingual language technologies through giving examples from high and low resource languages and their speakers/users.

Short bio : A. Seza Doğruöz is a faculty at Ghent University. Her research focuses on analyzing multilingual language use and linguistic variation between humans and developing language technologies accordingly. Her publications cover the themes related to Computational Sociolinguistics, Multilingualism & Linguistic Variation, Curating Language Resources (especially for low resource languages) and Open Domain Dialog Systems. Examples of her recent publications are :

- Doğruöz, A.S., Sitaram, S. (2022). Language Technologies for Low Resource Languages : Sociolinguistic and Multilingual Insights. LREC'22/SIGUL.
- Doğruöz, A.S., Sitaram, S., Bullock, B.E., Toribio, A.J. (2021). "A Survey of Code-switching : Linguistic and Social Perspectives for Language Technologies", Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), Bangkok, Thailand. (Nominated for the Best Paper Award)
- Doğruöz, A.S., Skantze, G. (2021). "How open are the conversations with open-domain chatbots ? A proposal for speech-event based evaluation". Proceedings of the 22nd Annual Meeting of Special Interest Group on Discourse and Dialogue (SIGDIAL2021), Singapore.

• Teresa Lynn

Research Fellow — The ADAPT Centre, Dublin City University

Titre / Title : The challenges of working with low-resourced languages in NLP : An Irish story.

Résumé / Abstract : Language technology underpins many of the applications and platforms that enable our digitally enhanced lives (virtual assistants, search engines, translation tools, spell-checkers, language learning tools, etc.). Yet these advances do not benefit all languages equally. Due to a lack of sufficient language technologies, users often need to revert to using another (better supported) language. Such a language shift plays a major role in the risk of digital extinction, i.e. an eventual decline in language use due to lack of technological support.
This talk focuses on Irish — an official EU language, and considered a low-resourced language in terms of digital support. "Low-resourced" not only means that there is a severe lack of speech and language applications available for Irish speakers to use, but it also means that the fundamental tools and language

resources required to build these technologies are also lacking. Irish is also in a precarious position while it competes alongside the most technologically supported language in the world — English. I will talk about our work at Dublin City University where we are taking some steps towards addressing this risk of digital extinction through the development of NLP tools and resources for Irish.

Short bio : Dr. Teresa Lynn is a Research Fellow at the ADAPT Centre in Dublin City University. She was awarded her PhD in 2015 from both DCU and Macquarie University Sydney, and was a recipient of the 2014-2015 Enterprise Ireland Fulbright Award, carrying out research at Saint Louis University, Missouri, USA.
Her main research interests lie in developing tools and resources for Irish language technology. She is the principal investigator on the GaelTech project, funded by the Irish Government Department of the Gaeltacht, which covers various research topics in Irish NLP. She is also a core member of the European Language Equality project and Ireland's National Anchor Point for the ELRC (European Language Resource Coordination), overseeing national data collection for Irish machine translation. Other projects include ELRI (European Language Resource Infrastructure), the National Digital Plan for Irish and the Universal Dependencies Project. Her research covers treebank development, syntactic parsing, social media NLP and multiword expressions. Teresa also worked in industry for several years, namely in the areas of localisation, NLP and machine translation.

Table des matières

Première partie

Travaux originaux