Actes de JEP-TALN-RECITAL 2024. 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés

Mathieu Balaguer, Nihed Bendahman, Lydia-Mai Ho-Dac, Julie Mauclair, José G. Moreno, Julien Pinquier

35èmes Journées d'Études sur la Parole (JEP 2024)
31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024)
26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)

https://jep-taln2024.sciencesconf.org

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés

Mathieu BALAGUER, Nihed BENDAHMAN, Lydia-Mai HO-DAC, Julie MAUCLAIR, Jose G MORENO, Julien PINQUIER (Éds.)

Toulouse, France, 8 au 12 juillet 2024

Avec le soutien de

Actes JEP-TALN-RECITAL Toulouse 2024 – ISBN : 978-2-917490-37-2
Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles – ISBN : 978-2-917490-39-6

Préface

Organisée conjointement par les équipes de recherche IRIS, MELODI et SAMoVA de l'Institut de Recherche en Informatique de Toulouse (IRIT UMR 5505), l'équipe PLC du laboratoire Cognition, Langues, Langage, Ergonomie (CLLE UMR 5263) et l'axe neurocognition langagière, linguistique et phonétique cliniques du laboratoire de NeuroPsychoLinguistique (LNPL URI EA 4156), sous l'égide de l'Association Francophone de la Communication Parlée (AFCP) et l'Association pour le Traitement Automatique des Langues (ATALA), la conférence JEP-TALN-RECITAL 2024 regroupe :

— les 35ème Journées d'Études sur la Parole (JEP),
— la 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN),
— la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL).

Les conférences TALN et JEP sont un rendez-vous qui offre le plus important forum d'échange francophone aux acteurs universitaires et industriels des technologies de la langue et la parole. Pour cette édition, nous avons plus de 200 inscrits dont une grande partie des étudiants qui construisent le futur de la recherche francophone et assurent le relais de son développement.

En tant que conférenciers invités, nous aurons Véronique HOSTE de l'Université de Ghent, Laurent BESACIER de Naver Labs Europe et Catia CUCCHIARINI de l'Université de Radboud. Ces trois conférenciers qui représentent un large spectre de thématiques entre le texte et la parole vont aborder les dernières avancées de leurs domaines d'expertise.

Cette édition permet aussi de célébrer les 30 ans de TALN. À cette occasion, nous avons dédié une session spéciale dans le programme. La session a comme objectif de rappeler l'historique de la conférence avec l'intervention des participants qui ont participé à sa pérennité afin de mieux transmettre les enjeux de ce rassemblement à la communauté scientifique du traitement automatique des langues naturelles.

En termes des soumissions, pour TALN, 66 articles pour la conférence principale ont été soumis, dont respectivement 18 ont été acceptés pour une présentation orale et 30 pour une présentation sous forme de posters. Également, nous avons reçu 13 résumés des articles publiés lors de conférences internationales qui ont été acceptés pour une présentation en format poster. En ce qui concerne RECITAL, 11 articles ont été soumis dont 7 ont été acceptés. L'ensemble des soumissions acceptées seront présentées sous forme de posters et 3 d'entre elles donneront lieu à une présentation orale. Pour les JEP, 64 articles ont été soumis et 62 ont été acceptés (17 sous forme de présentation orale et 45 sous format poster). L'alternance de sessions communes entre TALN, JEP et RECITAL et de sessions plus spécifiques devraient permettre de susciter des échanges fructueux. En complément de la conférence principale, se tiennent les ateliers "Parole Spontanée", "Défi Fouille de Texte" (DEFT), "Jurisprudence Prédictive" (JP'24), "Evaluation des modèles génératifs" (EvalLLM) et l'activité HackaTAL 2024. Ces événements illustrent à la fois des tendances nouvelles présentes dans la communauté et des activités récurrentes.

Il convient d'exprimer une profonde reconnaissance envers toutes les personnes qui ont participé à faire vivre la conférence, d'un côté les auteurs de toutes les soumissions et de l'autre les membres de différents comités scientifiques de la conférence. Un remerciement très chaleureux aux relecteurs qui ont accepté une charge importante et qui ont fait des relectures d'urgence afin de faciliter le bon déroulement de la conférence. La bienveillance et l'expertise des comités de programme ont permis la constitution d'un programme riche en thématiques et d'un niveau scientifique correspondant aux attentes de la communauté. Il est également essentiel d'exprimer notre gratitude envers les sponsors et les organisations qui ont subventionné la conférence. Leur soutien financier a permis à cet événement scientifique de se réaliser dans les meilleures conditions, rappelant l'importance des aspects financiers dans la réussite de telles

initiatives. Finalement, un grand merci aux différentes équipes présentes pour le bon fonctionnement, notamment des équipes de l'ATALA, l'AFCP et le CPRS qui nous ont accompagnés dans les différentes étapes de l'organisation.

Jose G Moreno	Lydia-Mai Ho-Dac Nihed Bendahman	Julie Mauclair
Président de TALN	Présidentes de RECITAL	Présidente de JEP

Comités

Comité de Programme

— Rachel Bawden, Inria
— Leonor Becerra-Bonache, Laboratoire d'Informatique et Systèmes
— Delphine Bernhard, LiLPa, Université de Strasbourg
— Nathalie Camelin, LIUM — Université du Maine
— Marie Candito, Universtité Paris 7 / INRIA
— Vincent Claveau, Irisa
— Géraldine Damnati, Orange Labs
— Iris Eshkol-Taravella, University of Orléans
— Benoit Favre, Aix-Marseille Université
— Natalia Grabar, STL CNRS Université Lille 3
— Thierry Hamon, France
— Lydia-Mai Ho-Dac, CLLE
— Philippe Langlais, Canada
— Jose G Moreno, IRIT – Université Paul Sabatier
— Emmanuel Morin, Université de Nantes, LS2N
— Vincent Segonne, Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
— Christophe Servan, Qwant Research
— Anne Vilnat, LIMSI-CNRS

Comité de Relecture

— Maxime Amblard, Université de Lorraine
— Jean-Yves Antoine, Université Franois Rabelais de Tours
— Lauriane Aufrant, Inria
— Frederic Bechet, Aix Marseille Université - LIF
— Patrice Bellot, Aix-Marseille Université - CNRS (LIS)
— Asma Ben Abacha, Microsoft Health AI
— Timothée Bernard, Université Paris Cité
— Romaric Besançon, CEA LIST
— Philippe Blache, LPL, AMU
— Chloé Braud, IRIT - CNRS
— Remi Cardon, CENTAL, IL&C, Université Catholique de Louvain
— Maximin Coavoux, CNRS, Université Grenoble Alpes
— Matthieu Constant, Université de Lorraine, ATILF, CNRS
— Caio Corro, Université Paris-Saclay
— Benoît Crabbé, Paris 7 et INRIA
— Béatrice Daille, Laboratoire d'Informatique Nantes Atlantique (LINA)
— Gaël de Chalendar, CEA LIST
— Gaël Dias, Normandie University
— Taoufiq Dkaki, IRIT, Intitut de Recherche en Informatique de Toulouse
— Benamara Farah, Univ. Paul Sabatier, Toulouse and IPAL, Singapore
— Olivier Ferret, CEA List
— Karën Fort, Sorbonne Université
— Amel Fraisse, Université de Lille
— Thomas Francois, Université catholique de Louvain
— Sahar Ghannay, LISN lab
— Cyril Grouin, LISN

— Gaël Guibon, Université de Lorraine - LORIA
— Nabil Hathout, CNRS
— Nicolas Hernandez, Nantes Université - LS2N CNRS UMR 6004
— Gilles Hubert, IRIT
— Luce Lefeuvre, DTIPG, SNCF
— Fabio Martínez Carrillo, Bivl2ab- Biomedical Imaging, vision and learning laboratory. Universidad Industrial de Santander
— Véronique Moriceau, IRIT Université Toulouse 3
— Philippe Muller, IRIT, Toulouse University
— Alexis Nasr, LIS
— Aurélie Névéol, Université Paris-Saclay, CNRS, LISN
— Jian-Yun Nie, University de Montreal
— Damien Nouvel, INaLCO
— Yannick Parmentier, LORIA - Université de Lorraine
— Patrick Paroubek, Universite Paris Sacle - CNRS
— Benjamin Piwowarski, CNRS / ISIR, Sorbonne Université
— Thierry Poibeau, LaTTiCe-CNRS
— Solen Quiniou, LS2N - Nantes Université
— Benoît Sagot, INRIA
— Djamé Seddah, Alpage/Université Paris la Sorbonne
— Nasredine Semmar, CEA
— Ludovic Tanguy, CLLE-ERSS
— Xavier Tannier, Sorbonne Université, INSERM, LIMICS
— Julien Tourille, CEA, LIST
— Guillaume Wisniewski, LLF - Université de Paris
— François Yvon, CNRS
— Pierre Zweigenbaum, Université Paris-Saclay, CNRS, LISN

Table des matières

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Apport de la structure de tours à l'identification automatique de genre textuel: un corpus annoté de sites web de tourisme en français*

Rémi Cardon[1] Trang Pham Tran Hanh[1,2]
Julien Zakhia Doueihi[1] Thomas François[1]
(1) CENTAL, IL&C, Université catholique de Louvain, Belgique
prenom.nom@uclouvain.be,
(2) Département de français, Université de Hanoï, Vietnam

RÉSUMÉ

Ce travail étudie la contribution de la structure de tours à l'identification automatique de genres textuels. Ce concept semble être peu exploité dans l'identification automatique du genre. Nous décrivons la collecte d'un corpus de sites web francophones relevant du domaine du tourisme et le processus d'annotation avec les informations de tours. Nous menons des expériences d'identification automatique du genre de texte avec notre corpus. Nos résultats montrent qu'ajouter l'information sur la structure de tours dans un modèle améliore ses performances pour l'identification automatique du genre, tout en réduisant le volume de données nécessaire et le besoin en ressource de calcul.

ABSTRACT

Contribution of Move Structure to Automatic Genre Identification : an Annotated Corpus of French Tourism Websites

The present work studies the contribution of the moves structure to automatic genre identification. This concept - well known in other branches of genre analysis - seems to have little application in automatic genre identification. We describe how we collect a corpus of websites in French related to tourism and annotate it with move structure. We conduct experiments on automatic genre identification with our corpus. Our results show that informing a model with move structure can increase its performance for automatic genre identification, and reduce the need for annotated data and computational power.

MOTS-CLÉS : identification automatique du genre de texte, analyse de genre, corpus, annotation.

KEYWORDS: automatic genre identification, genre analysis, corpus, annotation.

Références

CARDON R., PHAM T. T. H., ZAKHIA DOUEIHI J. & FRANÇOIS T. (2024). Contribution of move structure to automatic genre identification : An annotated corpus of French tourism websites. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 3916–3926, Torino, Italy : ELRA and ICCL.

∗. Ce document est un résumé de l'article (Cardon *et al.*, 2024).

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés, pages 1–1.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Caractérisation de la ville du futur dans un corpus de science-fiction

Sami Guembour[1, 2] Chuanming Dong[1, 3] Catherine Dominguès[1, 4]
(1) LASTIG, Univ Gustave Eiffel, ENSG, IGN, France
(2) Université Paris Cité, Paris, France
(3) ADEME, Agence de l'Environnement et de la Maîtrise de l'Energie, France
(4) I-SITE FUTURE, Paris, France
{sami.guembour,catherine.domingues}@ign.fr
dongchuanming@yahoo.com

RÉSUMÉ

Ce résumé présente le travail de recherche détaillé dans l'article (Guembour *et al.*, 2023). Ce travail s'intègre au projet PARoles de VIlleS (PARVIS, https ://parvis.hypotheses.org/) ; il se concentre sur la caractérisation de la ville du futur dans un corpus de romans de science-fiction et de fictions climatiques constitué par l'équipe PARVIS, en utilisant des techniques de traitement automatique des langues. L'objectif est d'analyser les images de la ville du futur à travers les éléments qu'elle contient (lieux, objets urbains, etc.) et les fonctions qui leur sont associées. Cependant, tous les romans du corpus ne parlent pas de la ville, d'où la nécessité d'identifier ceux dans lesquels elle constitue le cadre dans lequel s'intègrent l'action et les personnages. Pour cela, une ressource terminologique (Topalov *et al.*, 2010) permettant d'identifier les éléments spécifiques de la ville a été utilisée. Elle regroupe 533 mots, en majorité des noms, désignés par la forme OOC (abréviation de "Object Of the City" pour "Objets de la ville").

Un algorithme de clustering fondé sur les fréquences des OOC dans les romans est utilisé pour identifier ceux de la ville, et ainsi construire un nouveau corpus qui est spécifique à la ville. Les OOC ayant permis de construire ce nouveau corpus n'ont pas tous les mêmes fréquences et donc pas le même poids dans la description de la ville du futur. Un algorithme de co-clustering a été appliqué sur les fréquences des OOC dans le nouveau corpus afin d'identifier ceux ayant un poids important dans la description de la ville ; ces derniers sont appelés OOC discriminants. Un parseur de dépendance syntaxique est ensuite mis en oeuvre sur les phrases du nouveau corpus afin d'extraire les fonctions (verbes) associées aux OOC discriminants. Des algorithmes de clustering (et des techniques de réduction de dimension pour la visualisation des résultats) sont ensuite utilisés de manière itérative sur les cinq fonctions les plus fréquentes de chaque OOC discriminant pour regrouper ces OOC afin d'identifier d'éventuelles fonctions nouvelles ou anciennes associées à des OOC (anciens ou nouveaux objets de la ville).

Les résultats montrent que la ville du futur (dans le corpus PARVIS), comme celle d'aujourd'hui, vise à répondre principalement à deux problématiques principales : la circulation et l'habitation.

MOTS-CLÉS : TAL - corpus - clustering - réduction de dimensions - ville - urbain du futur - science-fiction.

KEYWORDS: NLP - corpus - clustering - dimension reduction - city - future urban - science fiction..

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés, pages 2–3.

Références

GUEMBOUR S., DONG C. & DOMINGUÈS C. (2023). Characterization of the city of the future from a science fiction corpus. In E. MÉTAIS, F. MEZIANE, V. SUGUMARAN, W. MANNING & S. REIFF-MARGANIEC, Éds., *Natural Language Processing and Information Systems*, p. 313–325, Cham : Springer Nature Switzerland.

TOPALOV C., DE LILLE L. C., DEPAULE J.-C. & MARIN B. (2010). *L'aventure des mots de la ville*. Paris, France : Robert Laffont.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

ChiCA: un corpus de conversations face-à-face vs. Zoom entre enfants et parents

Dhia-Elhak Goumri[1] Abhishek Agrawal[1] Mitja Nikolaus[2]
Duc Thang Vu Hong[1] Kübra Bodur[3] Elias Semmar[1] Cassandre Armand[4]
Chiara Mazzocconi[3, 5] Shreejata Gupta[3, 4] Laurent Prévot[3] Benoit Favre[1]
Leonor Becerra-Bonache[1] Abdellah Fourtassi[1]
(1) Aix Marseille Univ, CNRS, LIS, Marseille, France
(2) CerCo, CNRS, Toulouse, France
(3) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
(4) Aix Marseille Univ, CNRS, CRPN, Marseille, France
(5) Aix Marseille Univ, INSERM, INS, Marseille, France
{first_name.last_name}@univ-amu.fr

RÉSUMÉ

Les études existantes sur la parole en interaction naturelle se sont principalement concentrées sur les deux extrémités du spectre développemental, c'est-à-dire la petite enfance et l'âge adulte, laissant un vide dans nos connaissances sur la manière dont se déroule le développement, en particulier pendant l'age scolaire (6 à 11 ans). Le travail actuel contribue à combler cette lacune en introduisant un corpus développemental de conversations entre enfants et parents à domicile, impliquant des groupes d'enfants âgés de 7, 9 et 11 ans dont la langue maternelle est le français. Chaque dyade a été enregistrée deux fois : une fois en face-à-face et une fois en utilisant des appels vidéo par ordinateur. Pour les paramètres en face-à-face, nous avons capitalisé sur les progrès récents en matière de technologie de suivi oculaire mobile et de détection des mouvements de la tête pour optimiser le caractère naturel des enregistrements, nous permettant d'obtenir à la fois des données précises et écologiquement valides. De plus, nous avons contourné les difficultés de l'annotation manuelle en nous appuyant, dans la mesure du possible, sur des outils automatiques de traitement de la parole et de vision par ordinateur. Enfin, pour démontrer la richesse de ce corpus pour l'étude du développement communicatif de l'enfant, nous fournissons des analyses préliminaires comparant plusieurs mesures de la dynamique conversationnelle entre l'enfant et le parent selon l'âge, la modalité et le support communicatif. Nous espérons que le travail actuel ouvrira la voie à de futures découvertes sur les propriétés et les mécanismes du développement communicatif multimodal pendant l'age scolaire de l'enfant.

ABSTRACT

A Developmental Corpus of Child-Caregiver's Face-to-face vs. Computer-mediated Conversations in Middle Childhood

Existing studies of naturally occurring talk-in-interaction have largely focused on the two ends of the developmental spectrum, i.e., early childhood and adulthood, leaving a gap in our knowledge about how development unfolds, especially across middle childhood. The current work contributes to filling this gap by introducing a developmental corpus of child-caregiver conversations *at home*, involving groups of children aged 7, 9, and 11 years old. Each dyad was recorded twice : once in a face-to-face setting and once using computer-mediated video calls. For the face-to-face settings, we capitalized

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 4–5.

on recent advances in mobile, lightweight eye-tracking and head motion detection technology to optimize the naturalness of the recordings, allowing us to obtain both precise and ecologically valid data. Further, we mitigated the challenges of manual annotation by relying – to the extent possible – on automatic tools in speech processing and computer vision. Finally, to demonstrate the richness of this corpus for the study of child communicative development, we provide preliminary analyses comparing several measures of child-caregiver conversational dynamics across developmental age, modality, and communicative medium. We hope the current work will pave the way for future discoveries into the properties and mechanisms of multimodal communicative development across middle childhood.

MOTS-CLÉS : Corpus développemental, Conversations enfant-soignant, Enfance intermédiaire.

KEYWORDS: Developmental Corpus, Child-Caregiver Conversations, Middle Childhood.

Références

GOUMRI D. E., AGRAWAL A., NIKOLAUS M., VU H. D. T., BODUR K., EMMAR E., ARMAND C., MAZZOCCONI C., GUPTA S., PRÉVOT L. *et al.* (2024). Chica : A developmental corpus of child-caregiver's face-to-face vs. video call conversations in middle childhood. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 3153–3164.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Évaluer les modèles de langue pré-entraînés avec des propriétés de hiérarchie

Jesús Lovón-Melgarejo[1] Jose G. Moreno[1] Romaric Besançon[2] Olivier Ferret[2]
Lynda Tamine[1]
(1) Université Paul Sabatier, IRIT, Toulouse, France
(2) Université Paris-Saclay, CEA, List, Palaiseau, France
{jesus.lovon,jose.moreno,tamine}@irit.fr,
{romaric.besancon,olivier.ferret}@cea.fr

RÉSUMÉ

Étant donné que les modèles de langue pré-entraînés (PLM) constituent la pierre angulaire des modèles de recherche d'informations les plus récents, la façon dont ils encodent la connaissance sémantique est particulièrement importante. Cependant, on s'est peu intéressé à la capacité des PLM à capturer la connaissance sémantique hiérarchique. Traditionnellement, l'évaluation de ces connaissances codées dans les PLM s'appuie sur leurs performances lors d'évaluations dépendantes de la tâche, basées sur des tâches proxy telles que la détection d'hyperonymes. Malheureusement, cette approche ignore potentiellement d'autres relations taxonomiques implicites et complexes. Dans ce travail, nous proposons une méthode d'évaluation indépendante de la tâche, capable d'évaluer dans quelle mesure les PLM peuvent capturer des relations taxonomiques complexes, telles que les ancêtres et les frères et sœurs. Cette évaluation, basée sur des propriétés intrinsèques capturant ces relations, montre que les connaissances lexico-sémantiques codées implicitement dans les PLM ne capturent pas toujours les relations hiérarchiques. Nous démontrons en outre que les propriétés proposées peuvent être injectées dans les PLM pour améliorer leur compréhension de la hiérarchie. Grâce à des évaluations portant sur la reconstruction de taxonomies, la découverte d'hyperonymes et la compréhension de lecture, nous montrons que la connaissance de la hiérarchie est modérément transférable entre les tâches, mais pas de manière systématique. Ceci est le résumé de l'article "Probing Pretrained Language Models with Hierarchy Properties" publié à ECIR 2024 (Lovón-Melgarejo *et al.*, 2024).

ABSTRACT

Probing Pretrained Language Models with Hierarchy Properties

Since Pretrained Language Models (PLMs) are the cornerstone of the most recent Information Retrieval models, the way they encode semantic knowledge is particularly important. However, little attention has been given to studying the PLMs' capability to capture hierarchical semantic knowledge. Traditionally, evaluating such knowledge encoded in PLMs relies on their performance on task-dependent evaluations based on proxy tasks, such as hypernymy detection. Unfortunately, this approach potentially ignores other implicit and complex taxonomic relations. In this work, we propose a task-agnostic evaluation method able to evaluate to what extent PLMs can capture complex taxonomy relations, such as ancestors and siblings. This evaluation, based on intrinsic properties capturing these relations, shows that the lexico-semantic knowledge implicitly encoded in PLMs does not always capture hierarchical relations. We further demonstrate that the proposed properties can be injected into PLMs to improve their understanding of hierarchy. Through evaluations on taxonomy reconstruction, hypernym discovery and reading comprehension tasks, we show that knowledge about

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés, pages 6–7.

hierarchy is moderately but not systematically transferable across tasks. This is the summary of the published paper "Probing Pretrained Language Models with Hierarchy Properties" at ECIR 2024 (Lovón-Melgarejo *et al.*, 2024).

MOTS-CLÉS : modèles neuronaux de langue, relations taxonomiques, évaluation .

KEYWORDS: neural language models, taxonomic relations, evaluation .

Références

LOVÓN-MELGAREJO J., MORENO J. G., BESANÇON R., FERRET O. & TAMINE L. (2024). Probing pretrained language models with hierarchy properties. In *Advances in Information Retrieval : 46th European Conference on IR Research, ECIR 2024, Glasgow, Scotland, March 24–March 28, 2024* : Springer International Publishing.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Exploration d'approches hybrides pour la lisibilité : expériences sur la complémentarité entre les traits linguistiques et les transformers.

Rodrigo Wilkens[1] Patrick Watrin[1] Rémi Cardon[1]
Alice Pintard[1] Isabelle Gribomont[1,2] Thomas François[1]
(1) CENTAL, IL&C, Université catholique de Louvain, Belgique
(2) Royal Library of Belgium (KBR)
{rodrigo.wilkens, patrick.watrin, remi.cardon,alice.pintard,
isabelle.gribomont, thomas.francois}@uclouvain.be

RÉSUMÉ

Les architectures d'apprentissage automatique reposant sur la définition de traits linguistiques ont connu un succès important dans le domaine de l'évaluation automatique de la lisibilité des textes (ARA) et ont permis de faire se rencontrer informatique et théorie psycholinguistique. Toutefois, les récents développements se sont tournés vers l'apprentissage profond et les réseaux de neurones. Dans cet article, nous cherchons à réconcilier les deux approches. Nous présentons une comparaison systématique de 6 architectures hybrides (appliquées à plusieurs langues et publics) que nous comparons à ces deux approches concurrentes. Les diverses expériences réalisées ont clairement mis en évidence deux méthodes d'hybridation : *Soft-Labeling* et concaténation simple. Ces deux architectures sont également plus efficaces lorsque les données d'entraînement sont réduites. Cette étude est la première à comparer systématiquement différentes architectures hybrides et à étudier leurs performances dans plusieurs tâches de lisibilité.

ABSTRACT

Exploring hybrid approaches to readability : experiments on the complementarity between linguistic features and transformers [1]

Linguistic features have a strong contribution in the context of the automatic assessment of text readability (ARA). They have been one of the anchors between the computational and theoretical models. With the development in the ARA field, the research moved to Deep Learning (DL). In an attempt to reconcile the mixed results reported in this context, we present a systematic comparison of 6 hybrid approaches along with standard Machine Learning and DL approaches, on 4 corpora (different languages and target audiences). The various experiments clearly highlighted two rather simple hybridization methods (soft label and simple concatenation). They also appear to be the most robust on smaller datasets and across various tasks and languages. This study stands out as the first to systematically compare different architectures and approaches to feature hybridization in DL, as well as comparing performance in terms of two languages and two target audiences of the text, which leads to a clearer pattern of results.

MOTS-CLÉS : évaluation de la lisibilité, modèles hybrides, soft-labeling.

KEYWORDS: readabilty assessment, hybrid models, soft-labeling.

1. Référence de la publication : Wilkens, R., Watrin, P., Cardon, R., Pintard, A., Gribomont, I., & François, T. (2024, March). Exploring hybrid approaches to readability : experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics : EACL 2024* (pp. 2316-2331).

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 8–8.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Jargon : Une suite de modèles de langues et de référentiels d'évaluation pour les domaines spécialisés du français (Accepté à LREC/Coling 2024)

Vincent Segonne[1] Aidan Mannion [2,3] Laura Cristina Alonzo Canul [2] Alexandre Audibert[2] Xingyu Liu[2,4] Cécile Macaire[2] Adrien Pupier[2] Yongxin Zhou[2] Mathilde Aguiar[5] Felix Herron[2,6] Magali Norré[7,8] Massih-Reza Amini[2] Pierrette Bouillon[8] Iris Eshkol-Taravella[9] Emmanuelle Esperança-Rodier[2] Thomas François[7] Lorraine Goeuriot[2] Jérôme Goulian[2] Mathieu Lafourcade[10] Benjamin Lecouteux[2] François Portet[2] Fabien Ringeval[2] Vincent Vandeghinste[11,12] Maximin Coavoux[2] Marco Dinarelli[2] Didier Schwab[2]

(1) Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
(2) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(3) EPOS SAS, France
(4) Shesmet, France
(5) Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
(6) Laboratoire d'Analyse et de Modélisation de Systèmes d'Aide à la Décision (LAMSADE)
(7) CENTAL, IL&C, UCLouvain, Belgique
(8) Faculty of Translation and Interpreting, University of Geneva, Suisse
(9) MoDyCo, UPL, Univ Paris Nanterre, France
(10) LIRMM, Univ Montpellier, France
(11) Instituut voor de Nederlandse Taal, Pays-Bas
(12) KU Leuven, Belgique

Les modèles de langue préentraînés (PLM) constituent aujourd'hui *de facto* l'épine dorsale de la plupart des systèmes de traitement automatique des langues. Dans cet article, nous présentons Jargon, une famille de PLMs pour des domaines spécialisés du français, en nous focalisant sur trois domaines : la parole transcrite, le domaine clinique / biomédical, et le domaine juridique. Nous utilisons une architecture de transformeur basée sur des méthodes computationnellement efficaces (LinFormer) puisque ces domaines impliquent souvent le traitement de longs documents. Nous évaluons et comparons nos modèles à des modèles de l'état de l'art sur un ensemble varié de tâches et de corpus d'évaluation, dont certains sont introduits dans notre article. Nous rassemblons les jeux de données dans un nouveau référentiel d'évaluation en langue française pour ces trois domaines. Nous comparons également diverses configurations d'entraînement : préentraînement prolongé en apprentissage autosupervisé sur les données spécialisées, préentraînement à partir de zéro, ainsi que préentraînement mono et multi-domaines. Nos expérimentations approfondies dans des domaines spécialisés montrent qu'il est possible d'atteindre des performances compétitives en aval, même lors d'un préentraînement avec le mécanisme d'attention approximatif de LinFormer. Pour une reproductibilité totale, nous publions les modèles et les données de préentraînement, ainsi que les corpus utilisés.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés, pages 9–10.

ABSTRACT

Jargon : A Suite of Language Models and Evaluation Tasks for French Specialized Domains (Accepted LREC/Coling 2024)

Pretrained Masked Language Models (PLMs) are the de facto backbone of most state-of-the-art NLP systems. In this paper, we introduce a family of domain-specific pretrained PLMs for French, focusing on three important applications : the domains of transcribed speech, medicine, and law. We use a transformer architecture based on efficient methods (LinFormer) to maximise their utility, since these domains often involve processing long documents. We evaluate and compare our models to state-of-the-art models on a diverse set of tasks and datasets, some of which are introduced in this paper. We gather the datasets into a new French-language evaluation benchmark for these three domains. We also compare various training configurations : continued pretraining, pretraining from scratch, as well as single- and multi-domain pretraining. Extensive domain-specific experiments show that it is possible to attain competitive downstream performance even when pre-training with the approximative LinFormer attention mechanism. For full reproducibility, we release the models and pretraining data, as well as contributed datasets.

MOTS-CLÉS : Autoapprentissage, modèles de langue préentraînés, référentiels d'évaluation, Traitement Automatique de la langue biomédicale et clinique, Traitement Automatique de documents légaux, transcription automatique.

KEYWORDS: Self-supervised learning, pretrained language models, evaluation benchmark, biomedical document processing, legal document processing, speech transcription.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

LOCOST: Modèles Espace-État pour le Résumé Abstractif de Documents Longs

accepté à EACL 2024

Florian Le Bronnec[1, 2] Song Duong[1, 6] Alexandre Allauzen[2]
Vincent Guigue[5] Alberto Lumbreras[6] Laure Soulier[1] Patrick Gallinari[1, 6]
(1) Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
(2) Miles Team, Université Paris-Dauphine, Université PSL, CNRS, LAMSADE, 75016 Paris, France
(5) AgroParisTech, UMR MIA-PS, Palaiseau, France
(6) Criteo AI Lab, Paris, France
florian.le-bronnec@dauphine.psl.eu, s.duong@criteo.com

RÉSUMÉ

Les modèles espace-état constituent une alternative peu coûteuse en termes de complexité de calcul aux transformeurs pour le codage de longues séquences et la capture de longues dépendances. Nous proposons LOCOST : une architecture encodeur-décodeur basée sur des modèles espace-état pour la génération de textes conditionnels avec de longues entrées contextuelles. Avec une complexité de calcul de $\mathcal{O}(L \log L)$, cette architecture peut traiter des séquences beaucoup plus longues que les modèles de référence qui sont basés sur des modèles d'attention parcimonieux. Nous évaluons notre modèle sur une série de tâches de résumé abstractif de longs documents. Le modèle atteint un niveau de performance qui est 93-96% comparable aux transformeurs parcimonieux les plus performants de la même taille tout en économisant jusqu'à 50% de mémoire pendant l'apprentissage et jusqu'à 87% pendant l'inférence. En outre, LOCOST traite efficacement les entrées dépassant *600K* tokens au moment de l'inférence, établissant de nouveaux résultats de référence sur le résumé de livre complet et ouvrant de nouvelles perspectives pour le traitement des entrées longues.

ABSTRACT

LOCOST : State-Space Models for Long Document Abstractive Summarization.

State-space models are a low-complexity alternative to transformers for encoding long sequences and capturing long-term dependencies. We propose LOCOST : an encoder-decoder architecture based on state-space models for conditional text generation with long context inputs. With a computational complexity of $\mathcal{O}(L \log L)$, this architecture can handle significantly longer sequences than state-of-the-art models that are based on sparse attention patterns. We evaluate our model on a series of long document abstractive summarization tasks. The model reaches a performance level that is 93-96% comparable to the top-performing sparse transformers of the same size while saving up to 50% memory during training and up to 87% during inference. Additionally, LOCOST effectively handles inputs exceeding *600K* tokens at inference time, setting new state-of-the-art results on full-book summarization and opening new perspectives for long input processing.

MOTS-CLÉS : modèles espace-état, résumé abstractif de documents longs.

KEYWORDS: state-space models, long document abstractive summarization.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 11–11.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

La subjectivité dans le journalisme québécois et belge : transfert de connaissance inter-médias et inter-cultures

Article publié dans *Proceedings of the 17th International Conference on Statistical Analysis of Textual Data*

Louis Escouflaire[1, 2] Antonin Descampe[2] Antoine Venant[3] Cédrick Fairon[1]
(1) CENTAL, UCLouvain, Belgique
(2) ORM-EJL, UCLouvain, Belgique
(3) OLST, Université de Montréal, Canada
[louis.escouflaire|antonin.descampe|cedrick.fairon]@uclouvain.be, antoine.venant@umontreal.ca

RESUME

Cet article s'intéresse à la capacité de transfert des modèles de classification de texte dans le domaine journalistique, en particulier pour distinguer les articles d'opinion des articles d'information. A l'ère du numérique et des réseaux sociaux, les distinctions entre ces genres deviennent de plus en plus floues, augmentant l'importance de cette tâche de classification. Un corpus de 80 000 articles de presse provenant de huit médias, quatre québécois et quatre belges francophones, a été constitué. Pour identifier les thèmes des articles, une clusterisation a été appliquée sur les 10 000 articles issus de chaque média, assurant une distribution équilibrée des thèmes entre les deux genres *opinion* et *information*. Les données ont ensuite été utilisées pour entraîner (ou peaufiner) et évaluer deux types de modèles : CamemBERT (Martin et al., 2019), un modèle neuronal pré-entraîné, et un modèle de régression logistique basé sur des traits textuels. Dix versions différentes de chaque modèle sont entraînées : 8 versions 'mono-médias', chacune peaufinée sur l'ensemble d'entraînement du sous-corpus correspondant à un média, et deux versions 'multi-médias', l'une peaufinée sur 8000 articles québécois, l'autre sur les articles belges.
Les résultats montrent que les modèles CamemBERT surpassent significativement les modèles statistiques en termes de capacité de transfert (voir Figures 1 et 2). Les modèles CamemBERT montrent une plus grande exactitude, notamment sur les ensembles de test du même média que celui utilisé pour l'entraînement. Cependant, les modèles entraînés sur Le Journal de Montréal (JDM) sont particulièrement performants même sur d'autres ensembles de test, suggérant une distinction plus claire entre les genres journalistiques dans ce média. Les modèles CamemBERT multi-médias affichent également de bonnes performances. Le modèle québécois notamment obtient les meilleurs résultats en moyenne, indiquant qu'une diversité de sources améliore la généricité du modèle. Les modèles statistiques (mono- et multi-médias) montrent des performances globalement inférieures, avec des variations significatives selon les médias. Les textes québécois sont plus difficiles à classer pour ces modèles, suggérant des différences culturelles dans les pratiques journalistiques entre le Québec et la Belgique. L'analyse des traits révèle que l'importance de certains éléments textuels, comme les points d'exclamation et les marqueurs de temps relatifs, varient considérablement entre les modèles entraînés sur différents médias. Par exemple, les éditoriaux du JDM utilisent fréquemment des points d'exclamation, reflétant un style plus affirmé et polarisant. En revanche, les articles de La Presse présentent des particularités qui compliquent la généralisation de la tâche. En sommme, cette étude démontre la supériorité des modèles neuronaux comme CamemBERT pour la classification de textes journalistiques, notamment grâce à leur capacité de transfert, bien que les modèles basés sur des traits se distinguent par la transparence de leur 'raisonnement'. Elle met également en lumière des différences significatives entre les cultures journalistiques québécoises et belges.

MOTS-CLES : journalisme, grands modèles de langage, comparaison inter-culturelle, transfert de connaissance

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés, pages 12–13.

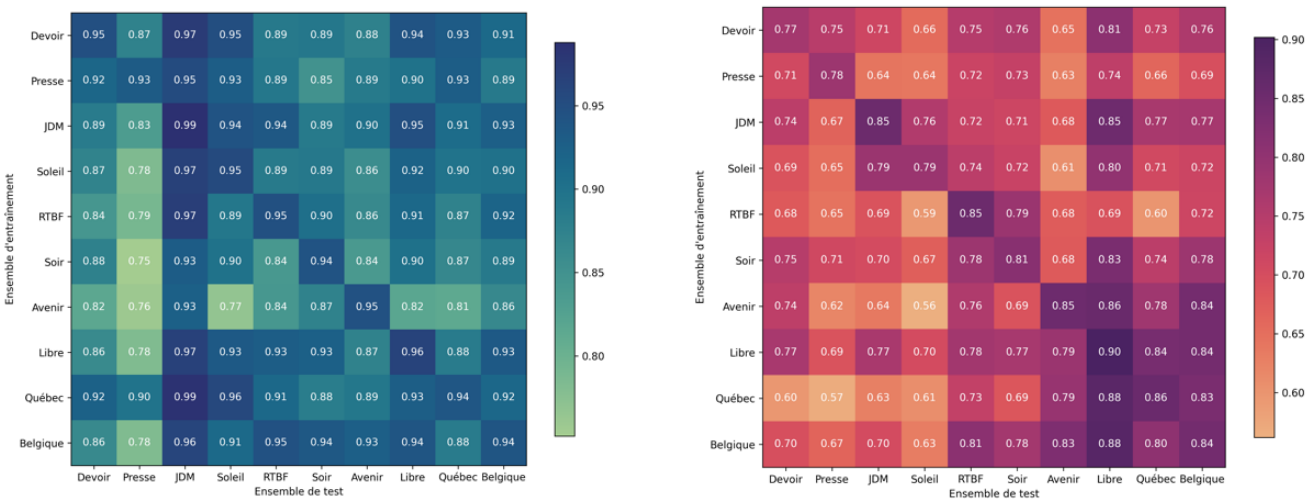

FIGURE 1 & 2 : Exactitudes sur les 10 ensembles de test des 10 modèles CamemBERT (à gauche) et des 10 modèles basés sur des traits (à droite).

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Le corpus BrainKT: Etudier l'instanciation du common ground par l'analyse des indices verbaux, gestuels et neurophysiologiques

Eliot Maës[1] Thierry Legou[2]
Leonor Becerra-Bonache[1] Philippe Blache[2]
(1) Aix Marseille Univ, CNRS, LIS, Marseille, France
(2) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
prenom.nom@lis-lab.fr, prenom.nom@univ-amu.fr

RÉSUMÉ

La quantité croissante de corpus multimodaux collectés permet de développer de nouvelles méthodes d'analyse de la conversation. Dans la très grande majorité des cas, ces corpus ne comprennent cependant que les enregistrements audio et vidéo, laissant de côté d'autres modalités plus difficiles à récupérer mais apportant un point de vue complémentaire sur la conversation, telle que l'activité cérébrale des locuteurs. Nous présentons donc BrainKT, un corpus de conversation naturelle en français, rassemblant les données audio, vidéo et signaux neurophysiologiques, collecté avec l'objectif d'étudier en profondeur les transmission d'information et l'instanciation du common ground. Pour chacune des conversations des 28 dyades (56 participants), les locuteurs devaient collaborer sur un jeu conversationnel (15min), et étaient ensuite libres de discuter du sujet de leur choix (15min). Pour chaque discussion, les données audio, vidéo, l'activité cérébrale (EEG par Biosemi 64) et physiologique (montre Empatica-E4) sont enregistrées. Cet article situe le corpus dans la littérature, présente le setup expérimental utilisé ainsi les difficultés rencontrées, et les différents niveaux d'annotations proposés pour le corpus.

ABSTRACT

Studying common ground instantiation using audio, video and brain behaviours : the BrainKT corpus

An increasing amount of multimodal recordings has been paving the way for the development of a more automatic way to study language and conversational interactions. However this data largely comprises of audio and video recordings, leaving aside other modalities that might complement this external view of the conversation but might be more difficult to collect in naturalistic setups, such as participants brain activity. In this context, we present BrainKT, a natural conversational corpus with audio, video and neuro-physiological signals, collected with the aim of studying information exchanges and common ground instantiation in conversation in a new, more in-depth way. We recorded conversations from 28 dyads (56 participants) during 30 minutes experiments where subjects were first tasked to collaborate on a joint information game, then freely drifted to the topic of their choice. During each session, audio and video were captured, along with the participants' neural signal (EEG with Biosemi 64) and their electrophysiological activity (with Empatica-E4). The paper situates this new type of resources in the literature, presents the experimental setup and describes the different kinds of annotations considered for the corpus.

MOTS-CLÉS : corpus, multimodalité, conversation naturelle, common ground, EEG.

KEYWORDS: corpus, multimodal, natural conversation, common ground, EEG.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés, pages 14–15.

Références

MAËS E., LEGOU T., BECERRA L. & BLACHE P. (2023). Studying common ground instantiation using audio, video and brain behaviours : the brainkt corpus. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, p. 691–702.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Rééquilibrer la distribution des labels tout en éliminant le temps d'attente inhérent dans l'apprentissage actif multi-label appliqué aux transformers

Maxime Arens[1, 2] Lucile Callebert[2] Jose G. Moreno[1] Mohand Boughanem[1]
(1) IRIT, Toulouse University, UMR 5505 CNRS, 31400 Toulouse, France
(2) Synapse Développement, 7 Boulevard de la Gare, 31500 Toulouse, France
maxime.arens@gmail.com

RÉSUMÉ

L'annotation des données est cruciale pour l'apprentissage automatique, notamment dans les domaines techniques, où la qualité et la quantité des données annotées affectent significativement l'efficacité des modèles entraînés. L'utilisation de personnel humain est coûteuse, surtout lors de l'annotation pour la classification multi-label, les instances pouvant être associées à plusieurs labels. L'apprentissage actif (AA) vise à réduire les coûts d'annotation en sélectionnant intelligemment des instances pour l'annotation, plutôt que de les annoter de manière aléatoire. L'attention récente portée aux transformers a mis en lumière le potentiel de l'AA dans ce contexte. Cependant, dans des environnements pratiques, la mise en œuvre de l'AA rencontre des défis pratiques. Notamment, le temps entre les cycles d'AA n'est pas mis à contribution par les annotateurs. Pour résoudre ce problème, nous examinons des méthodes alternatives de sélection d'instances, visant à maximiser l'efficacité de l'annotation en s'intégrant au processus de l'AA. Nous commençons par évaluer deux méthodes existantes, en utilisant respectivement un échantillonnage aléatoire et des informations de cycle d'AA périmées. Ensuite, nous proposons notre méthode novatrice basée sur l'annotation des instances pour rééquilibrer la distribution des labels. Notre approche atténue les biais, améliore les performances du modèle (jusqu'à une amélioration de 23% sur le score F1), réduit les disparités dépendantes de la stratégie (diminution d'environ 50% sur l'écart type) et diminue le déséquilibre des libellés (diminution de 30% sur le ratio moyen de déséquilibre). [1]

ABSTRACT

Rebalancing Label Distribution while Eliminating Inherent Waiting Time in Multi Label Active Learning applied to Transformers.

Data annotation is crucial for machine learning, notably in technical domains, where the quality and quantity of annotated data, significantly affect effectiveness of trained models. Employing humans is costly, especially when annotating for multi-label classification, as instances may bear multiple labels. Active Learning (AL) aims to alleviate annotation costs by intelligently selecting instances for annotation, rather than randomly annotating. Recent attention on transformers has spotlighted the potential of AL in this context. However, in practical settings, implementing AL faces challenges beyond theory. Notably, the gap between AL cycles presents idle time for annotators. To address this issue, we investigate alternative instance selection methods, aiming to maximize annotation efficiency by seamlessly integrating with the AL process. We begin by evaluating two existing methods in our transformer setting, employing respectively random sampling and outdated information. Following

1. Cet article a fait l'objet d'une publication en anglais à LREC-COLING 2024 (Arens *et al.*, 2024).

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 16–17.

this we propose our novel method based on annotating instances to rebalance label distribution. Our approach mitigates biases, enhances model performance (up to 23% improvement on f1score), reduces strategy-dependent disparities (decrease of nearly 50% on standard deviation) and reduces label imbalance (decrease of 30% on Mean Imbalance Ratio).

MOTS-CLÉS : apprentissage active, transformers, temps d'attente, distribution des labels.

KEYWORDS: active learning, transformers, wait time, label distribution.

Références

ARENS M., CALLEBERT L., BOUGHANEM M. & MORENO J. G. (2024). Rebalancing label distribution while eliminating inherent waiting time in multi label active learning applied to transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 13621–13632.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Sur les limites de l'identification par l'humain de textes générés automatiquement

Nadége Alavoine[1], Maximin Coavoux[2], Emmanuelle Esperança-Rodier[2], Romane Gallienne[3], Carlos-Emiliano González-Gallardo[4], Jérôme Goulian[2], Jose G. Moreno[5], Aurélie Névéol[6], Didier Schwab[2], Vincent Segonne[7] and Johanna Simoens[8]

(1) Université Paris-Saclay, LISN, Campus Universitaire bâtiment 507, Rue du Belvédère, 91400 Orsay, France
(2) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(3) Université Sorbonne Nouvelle, Lattice, CNRS, ENS-PSL, 1 rue Maurice Arnoux, 92120 Montrouge, France
(4) La Rochelle Université, L3i, 17000 La Rochelle, France
(5) University of Toulouse, IRIT, 31000 Toulouse, France
(6) LISN, Université Paris-Saclay, CNRS, 91403 Orsay, France
(7) Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
(8) Everteam, Bagneux, France
nadege.alavoine@universite-paris-saclay.fr,
{first.last}@univ-grenoble-alpes.fr
romane.gallienne@cnrs.fr, carlos.gonzalez_gallardo@univ-lr.fr,
jose.moreno@irit.fr, aurelie.neveol@lisn.upsaclay.fr
vincent.segonne@univ-ubs.fr, johanna.simoens@gmail.com

RÉSUMÉ

La génération de textes neuronaux fait l'objet d'une grande attention avec la publication de nouveaux outils tels que ChatGPT. La principale raison en est que la qualité du texte généré automatiquement peut être attribuée à un·e rédacteurice humain·e même quand l'évaluation est faite par un humain. Dans cet article, nous proposons un nouveau corpus en français et en anglais pour la tâche d'identification de textes générés automatiquement et nous menons une étude sur la façon dont les humains perçoivent ce texte. Nos résultats montrent, comme les travaux antérieurs à l'ère de ChatGPT, que les textes générés par des outils tels que ChatGPT partagent certaines caractéristiques communes mais qu'ils ne sont pas clairement identifiables, ce qui génère des perceptions différentes de ces textes par l'humain. Ceci est le résumé de l'article "Limitations of Human Identification of Automatically Generated Text" publié à LREC-COLING-2024 (Alavoine *et al.*, 2024).

ABSTRACT

Here the title in English.

Neural text generation is receiving broad attention with the publication of new tools such as ChatGPT. The main reason for that is that the achieved quality of the generated text may be attributed to a human writer by the naked eye of a human evaluator. In this paper, we propose a new corpus in French and English for the task of recognising automatically generated texts and we conduct a study of how humans perceive the text. Our results show, as previous work before the ChatGPT era, that the generated texts by tools such as ChatGPT share some common characteristics but they are not clearly identifiable which generates different perceptions of these texts.

MOTS-CLÉS : identification humaine, génération de texte avec des modèles neuronaux, ChatGPT.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 18–19.

KEYWORDS: human identifying, neural text generation, ChatGPT.

Références

ALAVOINE N., COAVOUX M., ESPERANÇA-RODIER E., GALLIENNE R., GALLARDO C. G., GOULIAN J., MORENO J. G., NEVEOL A., SCHWAB D., SEGONNE V. *et al.* (2024). Limitations of human identification of automatically generated text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 10511–10516.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Un corpus multimodal alignant parole, transcription et séquences de pictogrammes dédié à la traduction automatique de la parole vers des pictogrammes (Accepté à LREC/Coling 2024)

Cécile Macaire[1] Chloé Dion[1] Jordan Arrigo[1] Claire Lemaire[1,2]
Emmanuelle Esperança-Rodier[1] Benjamin Lecouteux[1] Didier Schwab[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(2) LAIRDIL, IUT, Univ. Paul Sabatier, 115 B rte de Narbonne, 31077 Toulouse, France
first.last@univ-grenoble-alpes.fr

RÉSUMÉ

La traduction automatique de la parole vers des pictogrammes peut faciliter la communication entre des soignants et des personnes souffrant de troubles du langage. Cependant, il n'existe pas de formalisme de traduction établi, ni d'ensembles de données accessibles au public pour l'entraînement de systèmes de traduction de la parole vers des pictogrammes. Cet article présente le premier ensemble de données alignant de la parole, du texte et des pictogrammes. Ce corpus comprend plus de 230 heures de parole. Nous discutons de nos choix pour créer une grammaire adaptée à des séquences de pictogrammes. Cette dernière s'articule autour de règles et d'un vocabulaire restreint. La grammaire résulte d'une étude linguistique approfondie des ressources extraites du site Web d'ARASAAC. Nous avons ensuite validé ces règles à l'issue de multiples phases de post-édition par des annotateurs experts. Le corpus proposé est ensuite utilisé pour entraîner un système en cascade traduisant la parole vers des pictogrammes. L'ensemble du corpus est disponible gratuitement sur le site web d'Ortolang sous une licence non commerciale. Il s'agit d'un point de départ pour la recherche portant sur la traduction automatique de la parole vers des pictogrammes.

ABSTRACT

A Multimodal Corpus of Aligned Speech, Text, and Pictogram Sequences for Speech-to-Pictogram Machine Translation

The automatic translation of spoken language into pictogram units can facilitate communication involving individuals with language impairments. However, there is no established translation formalism or publicly available datasets for training end-to-end speech translation systems. This paper introduces the first aligned speech, text, and pictogram translation dataset ever created in any language. We provide a French dataset that contains 230 hours of speech resources. We create a rule-based pictogram grammar with a restricted vocabulary and include a discussion of the strategic decisions involved. It takes advantage of an in-depth linguistic study of resources taken from the ARASAAC website. We validate these rules through multiple post-editing phases by expert annotators. The constructed dataset is then used to experiment with a Speech-to-Pictogram cascade model, which employs state-of-the-art Automatic Speech Recognition models. The dataset is freely available under a non-commercial licence. This marks a starting point to conduct research into the automatic translation of speech into pictogram units.

MOTS-CLÉS : Pictogrammes, Reconnaissance Automatique de la Parole, Traduction Automatique.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 2 : traductions d'articles publiés, pages 20–21.

KEYWORDS: Pictograms, Automatic Speech Recognition, Machine Translation.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Une approche zero-shot pour localiser les transferts d'informations en conversation naturelle *

Eliot Maës[1] Hossam Boudraa[1, 2]
Leonor Becerra-Bonache[1] Philippe Blache[3]
(1) Aix Marseille Univ, CNRS, LIS, Marseille, France
(2) Department of Computer Science, Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University, Fez, Morocco
(3) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
prenom.nom@lis-lab.fr, prenom.nom@univ-amu.fr

RÉSUMÉ

Les théories de l'interaction suggèrent que l'émergence d'une compréhension mutuelle entre les locuteurs en conversation naturelle dépend de la construction d'une base de connaissances partagée (*common ground*), mais n'explicitent ni le choix ni les circonstances de la mémorisation de ces informations. Des travaux antérieurs utilisant les métriques dérivées de la théorie de l'information pour analyser la dynamique d'échange d'information ne fournissent pas de moyen efficace de localiser les informations qui entreront dans le *common ground*. Nous proposons une nouvelle méthode basée sur la segmentation automatique d'une conversation en thèmes qui sont ensuite résumés. L'emplacement des transferts d'informations est finalement obtenu en calculant la distance entre le résumé du thème et les différents énoncés produits par un locuteur. Nous évaluons deux larges modèles de langue (LLMs) sur cette méthode, sur le corpus conversationnel français Paco-Cheese.

ABSTRACT

Did You Get It ? A Zero-Shot Approach To Locate Information Transfers In Conversations

Interaction theories suggest that the emergence of mutual understanding between speakers in natural conversations depends on the construction of a shared knowledge base (common ground), but the details of which information and the circumstances under which it is memorized are not explained by any model. Previous works have looked at metrics derived from Information Theory to quantify the dynamics of information exchanged between participants, but do not provide an efficient way to locate information that will enter the common ground. We propose a new method based on the segmentation of a conversation into themes followed by their summarization. We then obtain the location of information transfers by computing the distance between the theme summary and the different utterances produced by a speaker. We evaluate two Large Language Models (LLMs) on this pipeline, on the French conversational corpus Paco-Cheese.

MOTS-CLÉS : Conversation Naturelle, Résumé, Localisation d'informations, Segmentation Thématique, LLMs.

KEYWORDS: Natural Conversation, Summarization, Information Location, Thematic Segmentation, LLMs.

∗. MAËS E., BOUDRAA H., BLACHE P. & BECERRA-BONACHE L. (2024). Did you get it ? a zero-shot approach to locate information transfers in conversations. In *LREC-COLING 2024 - The 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation.*

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 22–22.