Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

*Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr*

Thibault BAÑERAS-ROUX : thibault.baneras.roux@gmail.com

Titre : Analyse et compréhension de l'évaluation des systèmes de reconnaissance automatique de la parole : vers des métriques intégrant la perception humaine

Mots-clés : reconnaissance automatique de la parole, métriques d'évaluation, perception humaine, sémantique.

Title: *Analysis and Understanding of the Evaluation of Automatic Speech Recognition Systems: Towards Metrics Integrating Human Perception*

Keywords: *automatic speech recognition, evaluation metrics, human perception, semantics.*

Thèse de doctorat en informatique, LS2N, UFR Sciences et techniques, Nantes Université, sous la direction de M. Richard Dufour (Pr, Nantes Université), Mme Jane Wottawa (MC, Le Mans Université) et M. Mickael Rouvier (MC HDR, Avignon Université). Thèse soutenue le 17/01/2025.

Jury : M. Richard Dufour (Pr, Nantes Université, directeur), Mme Irina Illina (MC HDR, université de Lorraine, rapporteuse), M. Cyril Grouin (IR HDR, université Paris-Saclay, rapporteur), M. Benjamin Lecouteux (Pr, université Grenoble Alpes, président), Mme Béatrice Daille (Pr, Nantes Université, examinatrice), Mme Martine Adda-Decker (DR, CNRS, examinatrice), Mme Jane Wottawa (MC, Le Mans Université, co-encadrante), M. Mickael Rouvier (MC HDR, Avignon Université, co-encadrant).

Résumé : *Le taux d'erreur de mots reste la métrique principale pour évaluer les systèmes de reconnaissance automatique de la parole (RAP), mais il ne reflète pas toujours la perception humaine. Cette thèse propose des métriques alternatives pour*

évaluer non seulement l'orthographe, mais aussi la grammaire, la sémantique et la phonétique. À travers le corpus HATS, annoté par 143 francophones, nous avons mesuré la corrélation entre ces métriques et les choix humains. Les résultats montrent que *SemDist*, basée sur les représentations sémantiques de BERT, est la plus pertinente, tandis que le taux d'erreur de mots se révèle peu performant. Une analyse des hyperparamètres des systèmes de RAP révèle que chaque métrique évalue des aspects distincts, soulignant l'importance d'une évaluation multimétrique. Enfin, pour rendre les métriques sémantiques plus compréhensibles, nous avons développé la méthode *minED*, qui identifie la gravité des erreurs et améliore l'interprétation des scores, offrant des outils précieux pour évaluer et perfectionner les systèmes RAP.

URL où le mémoire peut être téléchargé :

<https://theses.hal.science/tel-04931310>

Johanna CORDOVA : johanna.cordova@inalco.fr

Titre : Le quechua dans les outils numériques, un défi pour le TAL ? Développement de ressources linguistiques et numériques pour le quechua ancashino

Mots-clés : TAL, quechua, langue agglutinante, langue minoritaire, analyse morphologique, corpus aligné, corpus arboré, *Universal Dependencies*, OCR.

Title: *Quechua in Digital Tools: A Challenge for NLP? Development of Linguistic and Digital Resources for Ancash Quechua*

Keywords: *NLP, Quechua, agglutinative language, low-resourced language, morphological analysis, parallel corpora, treebank, Universal Dependencies, OCR.*

Thèse de doctorat en traitement automatique des langues, école doctorale Langues, littératures et sociétés du monde, institut national des langues et civilisations orientales, sous la direction de M. César Itier (Pr, institut national des langues et civilisations orientales) et M. Damien Nouvel (MC, institut national des langues et civilisations orientales). Thèse soutenue le 18/12/2024.

Jury : M. César Itier (Pr, institut national des langues et civilisations orientales, codirecteur), M. Damien Nouvel (MC, institut national des langues et civilisations orientales, codirecteur), Mme Capucine Boidin (Pr, université Sorbonne-Nouvelle, institut des hautes études de l'Amérique latine, IHEAL, centre de recherche et de documentation sur les Amériques, CREDA, présidente), Mme Kata Gábor (MC, institut national des langues et civilisations orientales, ERTIM, examinatrice), Mme Katharina Haude (CR, CNRS, Structure et dynamique des langues, SeDyl, UMR 8202, examinatrice), M. Elwin Huaman (*senior researcher*, Semantic Technology Institute Innsbruck, université d'Innsbruck, Autriche, examinateur), M. Sylvain Kahane (Pr, université Paris Nanterre, rapporteur), M. Matthias Urban (CR, CNRS, Dynamique du langage, DDL, rapporteur).

Résumé : *Les langues quechuas constituent l'une des familles linguistiques amérindiennes comptant le plus grand nombre de locuteurs natifs. Au Pérou, selon le recensement de 2017, 13,9% de la population a le quechua pour première langue et environ 20% le parle. Pourtant, cette macrolangue est presque totalement absente des usages numériques. En traitement automatique des langues (TAL), c'est une langue peu dotée, avec une forte disparité de ressources selon la variété de quechua considérée. L'objectif de cette thèse est de développer un ensemble d'outils fondamentaux pour le traitement automatique d'une variété du quechua central, le quechua ancashino, parlé par environ 400 000 personnes, et en danger d'extinction d'après la classification de l'UNESCO. Ce processus comporte trois étapes : la première est la numérisation des ressources disponibles dans cette variété (dictionnaires, corpus écrits), qui permet de constituer le plus grand corpus bilingue quechua-espagnol aligné par phrases disponible à ce jour pour le traitement automatique (6 120 phrases pour 38 500 tokens en quechua). La seconde étape, indispensable pour une langue agglutinante comme le quechua, est l'implémentation d'un analyseur morphologique permettant d'identifier, à partir d'une forme de surface, les morphes et leurs morphèmes correspondants, ainsi que leur étiquette suivant le formalisme de Leipzig. Outre l'analyse au niveau du mot, nous proposons une méthode basée sur les CRF pour sélectionner les annotations correctes en fonction du contexte phrasique lorsque plusieurs analyses sont possibles. Nous terminons par l'élaboration d'un corpus arboré pour l'analyse en morphosyntaxe, suivant le formalisme Universal Dependencies. Ces développements s'accompagnent d'une description approfondie de la morphologie du quechua ancashino, largement sous-étudiée. Les ressources développées sont distribuées sous licence libre pour des travaux ultérieurs en TAL, mais aussi adaptées pour une utilisation grand public à travers des applications telles qu'un moteur de recherche permettant d'interroger l'ensemble des dictionnaires ou un outil d'analyse morphologique en ligne. Dans un contexte global de valorisation des langues originaires et alors que d'ambitieuses politiques liées aux droits linguistiques sont en cours de déploiement dans les pays de l'aire andine, nous faisons l'hypothèse que la présence du quechua dans les technologies constitue un important levier pour renforcer sa pratique et faciliter son enseignement.*

URL où le mémoire peut être téléchargé :

<https://theses.hal.science/tel-04989396>

Nicolas GUTEHRLÉ : nicolas.gutehrle@univ-fcomte.fr

Titre : Extraction d'informations appliquée aux documents non structurés pour la valorisation de périodiques historiques : application au patrimoine de la région Bourgogne Franche-Comté en France

Mots-clés : extraction d'information, gestion des connaissances, annotation sémantique, interfaces de recherche, exploitation et exploration des documents, humanités numériques.

Title: *Information Extraction from Unstructured Documents for the Valorisation of Historical Periodicals: Application to the Heritage of the Bourgogne Franche-Comté Region in France*

Keywords: *information extraction, knowledge management, semantic annotation, search interfaces, exploration and exploitation of documents, digital humanities.*

Thèse de doctorat en sciences du langage, mention traitement automatique des langues, centre de recherches interdisciplinaires et transculturelles, CRIT, UR 3224, UFR Sciences du langage, de l'homme et de la société, université de Franche-Comté, Besançon, sous la direction de Mme Iana Atanassova (MC HDR, université de Franche-Comté, centre de recherches interdisciplinaires et transculturelles, CRIT, UR 3224, Besançon). Thèse soutenue le 21/06/2024.

Jury : Mme Iana Atanassova (MC HDR, université de Franche-Comté, centre de recherches interdisciplinaires et transculturelles, CRIT, UR 3224, Besançon, directrice), M. Antoine Doucet (Pr, La Rochelle Université, laboratoire Informatique, image et interaction, L3i, président), M. Bruno Bachimont (Pr, université de technologie de Compiègne, Connaissance, organisation et systèmes techniques, Costech, rapporteur), M. Pavel Pecina (MC, Charles University, Institute of Formal and Applied Linguistics, ÚFAL, Prague, République tchèque, rapporteur), M. Jean-Charles Lamirel (MC, université de Strasbourg, laboratoire lorrain de recherche en informatique et ses applications, LORIA, SYNALP, Nancy, examinateur), M. Mohand Boughanem (Pr, université Paul Sabatier, institut de recherche en informatique de Toulouse, IRIT, examinateur), M. Adam Jatowt (Pr, Universität Innsbruck, Data Science Group, Autriche, examinateur).

Résumé : *Ces dernières années, les bibliothèques et archives ont entrepris de nombreuses campagnes de numérisation afin d'élargir l'accès du public à leurs collections d'archives. Cependant, le défi de promouvoir le contenu des collections et de rendre ces ressources accessibles reste entier. La numérisation produit souvent un contenu non structuré dans lequel il est difficile de naviguer, tandis que les interfaces qui s'appuient sur des requêtes basées sur des mots-clés pour accéder aux documents d'archives peuvent fournir aux utilisateurs des résultats non pertinents. Afin d'exploiter le potentiel des « Big Data of the Past », notion introduite par Kaplan et di Lenardo en 2017, il est essentiel de développer des méthodes et des cadres pour structurer le contenu textuel des documents, dans le but d'en améliorer l'exploration et l'exploitation. Dans ce contexte, la présente thèse de doctorat aborde le problème du traitement des documents historiques numérisés, en se concentrant sur l'extraction des entités nommées et des relations afin de créer des interfaces pour l'exploitation efficace des données textuelles historiques. Premièrement, nous proposons une nouvelle méthode pour déterminer la structure logique des journaux historiques en utilisant une approche à base de règles. Deuxièmement, nous présentons une méthode pour extraire les entités et les relations concernant les personnes et les lieux mentionnés dans les textes. Notre approche s'intitule Extensible, Lightweight and Interpretable Joint*

Extraction of Relations and Entities (*ELIJERE*). Elle est basée sur des ressources linguistiques obtenues par supervision distante. Enfin, nous proposons un cadre général pour l'étude de l'expression d'informations spatiales dans les documents, et un autre cadre pour l'application des méthodes de TimeLine Summarisation à des collections de documents. Nous montrons comment ces méthodes peuvent être appliquées pour produire des interfaces sémantiquement riches, telles que des frises chronologiques et des cartes, qui permettent au grand public une lecture proche ou distante de ces collections.

URL où le mémoire peut être téléchargé :

<https://theses.hal.science/tel-04719778>

Thi Phuong Hang LE : hangtp.le@gmail.com

Titre : Architectures et techniques d'entraînement pour la traduction parole-texte multilingue

Mots-clés : multilinguisme, traduction automatique, traduction automatique de la parole, *transformer*, traduction parole-texte, traduction.

Title: *Model Architectures and Training Techniques for Multilingual Speech-to-Text Translation*

Keywords: *multilingualism, automatic speech recognition, ASR, speech-to-text translation, ST, transformer, translation.*

Thèse de doctorat en mathématiques et informatique, laboratoire d'informatique de Grenoble, LIG, école doctorale Mathématiques, sciences et technologies de l'information, université Grenoble Alpes, sous la direction de M. Didier Schwab (Pr, université Grenoble Alpes) et M. Benjamin Lecouteux (Pr, université Grenoble Alpes). Thèse soutenue le 25/03/2024.

Jury : M. Didier Schwab (Pr, université Grenoble Alpes, codirecteur), M. Benjamin Lecouteux (Pr, université Grenoble Alpes, codirecteur), Mme Caroline Rossi (Pr, université Grenoble Alpes, présidente), M. Frédéric Béchet (Pr, université Aix-Marseille, rapporteur), M. François Yvon (DR, CNRS, rapporteur), M. Laurent Besacier (*principal scientist*, HDR, NaverLabs Europe, examinateur), M. Juan Pino (*research scientist*, Meta AI Research, examinateur).

Résumé : *Speech-to-text translation (ST) consists in translating a speech audio input in one language into a text output in another language. This task is highly challenging due to its multimodal and multilingual nature (the former involves both speech and text modalities, while the latter involves more than one language). In this thesis, we make three major contributions spanning two primary research areas of ST, namely model architectures and training techniques.*

First, in terms of model architectures, we introduce the dual-decoder transformer, a new model architecture that jointly performs automatic speech recognition (ASR) and multilingual ST. Our model consists of two decoders, each responsible for one task (ASR or ST), that can interact with each other through a novel dual-attention mechanism. This design allows the decoders to specialize in their respective tasks while being helpful to each other. We propose two variants, called the parallel and cross dual-decoder transformers, corresponding to two different levels of dependencies between the decoders. The proposed model also generalizes existing approaches using two independent or weakly tied decoders. Experiments on standard benchmarks show that our models outperform previous work in terms of translation performance under both bilingual and multilingual settings.

Second, in terms of training techniques, we propose a parameter-efficient fine-tuning approach based on adapter modules. We show that language-specific adapters can enable a fully trained multilingual ST model to be further specialized in each language pair. With these adapter modules, one can efficiently obtain a single multilingual ST system that outperforms the original multilingual model as well as multiple bilingual systems while maintaining a low storage cost and simplicity in deployment. In addition, we show that adapters can also be used to connect available pre-trained models such as an ASR model and a multilingual denoising auto-encoder to form strong multilingual ST systems.

Finally, as a second contribution in training techniques, we propose an effective supervised pre-training method to address the so-called speech-text modality gap, a well-known major challenge in ST. Our method combines the connectionist temporal classification loss and optimal transport in a Siamese-like model. This model is composed of two encoders, one for acoustic inputs and the other for textual inputs, which are trained such that they produce representations that are close to each other in the Wasserstein space. Extensive experiments on standard benchmarks show that our pre-training method applied to the vanilla encoder-decoder transformer achieves state-of-the-art performance under the no-external-data setting, and performs on par with recent strong multi-task learning systems trained with external data. Finally, our method can also be applied on top of these multi-task systems, leading to further improvements for these models.

URL où le mémoire peut être téléchargé :

<https://theses.hal.science/tel-04726713>

Anna PAPPÀ : apappa@univ-paris8.fr

Titre : Contributions à la création de corpus et de modèles d'apprentissage profond pour les données textuelles multilingues

Mots-clés : création de corpus, corpus multilingues, annotation sémantique, analyse textuelle, apprentissage profond, méthodes hybrides.

Title: *Contributions to Corpus Creation and Deep Learning Models for Multilingual Textual Data*

Keywords: *corpus creation, multilingual corpora, semantic annotation, textual analysis, deep learning, hybrid methods.*

Habilitation à diriger des recherches en informatique, laboratoire d'intelligence artificielle et sémantique des données, LIASD, UFR des sciences et des technologies du numérique, département Programmation et informatique fondamentale, université Paris 8. Habilitation soutenue le 22/04/2024.

Jury : M. Nicolas Jouandeau (Pr, université Paris 8, rapporteur), M. Christian Boitet (Pr émérite, université Grenoble Alpes, rapporteur), M. Benoît Crabbé (Pr, université Paris Cité, rapporteur), M. Max Silberstein (Pr, université de Franche-Comté, examinateur), Mme Tita Kyriacopoulou (Pr, université Gustave Eiffel, examinatrice), M. Tristan Cazenave (Pr, université Paris Dauphine, examinateur), M. Jean-Jacques Bourdin (Pr, université Paris 8, examinateur).

Résumé : *Cette habilitation à diriger des recherches se situe à l'intersection de l'informatique et de la linguistique, synthétisant près d'une décennie de travaux qui explorent cette interface multidisciplinaire. Elle est structurée autour de trois axes principaux, chacun apportant des contributions au domaine du traitement automatique des langues (TAL).*

Le premier axe présente la création de corpus multilingues et thématiques, spécifiquement dédiés à l'analyse des opinions et des aspects. Les méthodologies et les outils développés visent à réduire le bruit et les irrégularités dans les données, en se concentrant sur des avis d'utilisateurs provenant de diverses plateformes en ligne. Ces corpus multilingues sans bruit servent de base solide pour les expérimentations subséquentes et permettent une analyse plus précise et pertinente des sentiments exprimés dans différentes langues.

Le deuxième axe explore l'analyse de sentiment par le biais de modèles hybrides, combinant des réseaux neuronaux convolutifs (CNN) et récurrents (RNN) tels que les LSTM (Long Short-Term Memory). Ces modèles atteignent une précision élevée, oscillant entre 90 % et 100 % sur divers corpus multilingues non annotés. L'utilisation de réseaux neuronaux convolutifs hiérarchiques (ConvNet) et de réseaux neuronaux récurrents permet de relever les défis de la prédiction de la polarité et de la classification thématique, améliorant ainsi les performances et la robustesse des systèmes de TAL.

Le troisième axe aborde l'annotation d'aspects en utilisant une architecture combinée BiLSTM-CNN-CRF (Bidirectional Long Short-Term Memory-Convolutional Neural Network-Conditional Random Field), enrichie par des techniques d'apprentissage profond telles que l'apprentissage actif et l'apprentissage par transfert. Ces méthodes se sont avérées particulièrement efficaces pour améliorer les performances des modèles dans des contextes de rareté de données ou de langues peu représentées, en

permettant une annotation plus précise et fiable des aspects dans des textes multilingues.

En synthèse, cette habilitation constitue une contribution pertinente à la fusion des méthodologies informatiques et linguistiques. Elle exploite des jeux de données sur mesure et des architectures de modèles hybrides pour résoudre des défis complexes en annotation sémantique et en analyse multilingue. Ce travail ouvre également la voie à des recherches futures pour affiner ces méthodologies dans des scénarios encore plus exigeants ou moins étudiés, et propose des perspectives prometteuses pour le développement de nouveaux outils et techniques en TAL. Les résultats obtenus démontrent l'importance et l'efficacité de l'approche multidisciplinaire dans le traitement et l'analyse des données textuelles.

URL où le mémoire peut être téléchargé :

<https://univ-paris8.hal.science/tel-04595140>