AfIA

Association française
pour l'Intelligence Artificielle

Dossier N° 2

***P**anorama **F**rançais de la **R**echerche en **T**echnologies du **L**angage **H**umain*

Collège TLH

Octobre 2021 N° 2

AfIA
Association française
pour l'Intelligence Artificielle

SOMMAIRE

DU DOSSIER

AfIA
Association française
pour l'Intelligence Artificielle

Dossier réalisé par

Gaël DIAS
GREYC UMR 6072
Université de Caen Normandie
gael.dias@unicaen.fr

AfIA
Association française
pour l'Intelligence Artificielle

Édito

Ce deuxième dossier vise à compléter la liste des équipes de recherche académiques et industrielles françaises menant des travaux à l'intersection du traitement automatique des langues (TAL), de la recherche d'information (RI), de la communication parlée (CP) et de l'intelligence artificielle (IA), domaine communément appelé technologies du langage humain (TLH). En particulier, les TLH proposent des méthodes permettant une communication homme-machine naturelle, pouvant s'étendre à une interaction homme-homme médiée. Ainsi, les TLH permettent d'analyser, d'interpréter et de produire des actes du langage écrit, parlé ou signé, mais aussi d'interagir avec des données langagières.

Suite à un deuxième appel à participation communiqué sur les listes de diffusion françaises des domaines de recherche des TLH, nous avons reçu 8 nouvelles contributions, totalisant ainsi 30 institutions, dont 27 laboratoires académiques et 3 entreprises, répartis sur 12 villes plus Paris et sa région (Figure 1). En plus des nouvelles contributions, les équipes présentes dans le premier dossier ont également actualisé leurs participations. Dans cette nouvelle version, nous avons changé la nomenclature de façon à ce que chaque équipe soit associée à son laboratoire.

La diversité des recherches présentées ainsi que la qualité et la quantité des contributions reçues démontrent à la fois une dynamique importante des TLH en France mais aussi un savoir-faire et des compétences reconnus à l'international. Notamment, il est très intéressant de remarquer la pluralité des approches scientifiques suivies, ce qui ne fait que renforcer une particularité nationale propice au foisonnement des idées.

Ce dossier ne se veut pas exhaustif mais a le mérite de rendre compte assez fidèlement du large spectre des thématiques abordées en TAL, RI et CP en France. Ainsi, si vous recherchez des spécialistes en (1) linguistique computationnelle, en veille d'information, en moteurs de recherche, en systèmes de questions réponses, en scientométrie, en web sémantique, en traduction automatique, en classification de textes, en analyse de sentiments et d'opinions, en génération de textes, en systèmes de recommandation, en synthèse et reconnaissance de parole, en agents conversationnels, en *forensic*, en simplification de textes, en grammaires formelles, en sémantique lexicale, en extraction d'information, en indexation, en ingénierie des documents ou en analyse des réseaux sociaux, dans (2) un cadre de données hétérogènes, multimodales, multilingues, sous-dotées ou complexes, pour (3) des applications en santé, en environnement, en biologie, en conservation du patrimoine, en agriculture, en handicap, en génétique ou en éducation, et dans (4) un cadre éventuellement pluri ou transdisciplinaire, alors vous trouverez un interlocuteur en France.

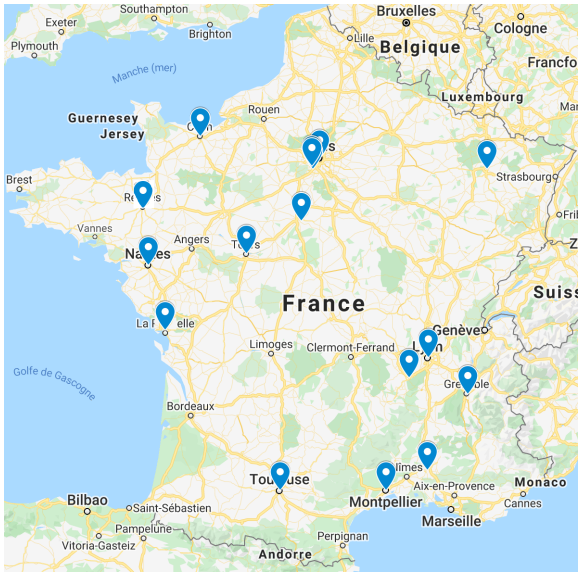

Fig. 1 : Cartographie des TLH en France.

Je tiens à remercier particulièrement tous les contributeurs de ce bulletin qui ont pris de leur temps et de leur énergie pour promouvoir leur discipline et informer la communauté de leurs recherches actuelles, ainsi que les membres du comité de pilotage du collège TLH pour leur soutien dans cette initiative.

J'espère que vous trouverez autant de plaisir à lire ce deuxième dossier que j'en ai pris à sa réalisation. Bonne lecture.

Gaël DIAS

AfIA
Association française
pour l'Intelligence Artificielle

MaIAGE/BIBLIOME : Acquisition et Formalisation de Connaissances à partir de Textes

MaIAGE UR 1404 / Bibliome
INRAE et Université Paris-Saclay
http://maiage.inrae.fr/

Claire NÉDELLEC
claire.nedellec@inrae.fr

Robert BOSSY
robert.bossy@inrae.fr

Louise DELÉGER
louise.deleger@inrae.fr

Arnaud FERRÉ
arnaud.ferre@inrae.fr

Domaine de recherche

L'équipe Bibliome développe des méthodes d'extraction et de formalisation d'information à partir de textes écrits. Ces méthodes identifient et formalisent des informations et connaissances précises dans de larges corpus de documents de genres divers et les mettent en relation, faisant appel à des méthodes de traitement automatique de la langue et d'apprentissage automatique. Les principaux travaux concernent trois sujets :

1. l'apprentissage automatique pour la reconnaissance et la formalisation d'entités et de relations ;
2. la conception de terminologies et d'ontologies ;
3. l'intégration et l'évaluation des méthodes dans une infrastructure partagée.

Nos recherches sont guidées par des besoins applicatifs qui permettent de valider nos méthodes et d'identifier les objectifs prioritaires dans des domaines variés de la biologie, microbiologie, génétique et phénotypes des plantes et des animaux d'élevage.

Méthodes développées

Les méthodes en intelligence artificielle développées par l'équipe Bibliome traitent deux étapes clés, l'extraction et l'annotation des entités du texte par des concepts d'ontologie et l'extraction de relations formelles entre ces entités. Pour étudier des phénomènes scientifiques en sciences du vivant dispersés dans une grande quantité de documents, nos travaux ont pour objectif de compenser le petit nombre d'occurrences par des approches dites *knowledge intensive*, combinant analyse linguistique computationnelle, connaissance du domaine sous forme de lexiques et d'ontologie et apprentissage automatique, facilitant la généralisation des méthodes et leur adaptation à de nouvelles questions.

Par exemple, l'équipe Bibliome développe la méthode HONOR [7] qui intègre deux méthodes complémentaires pour la détection et le rattachement de termes du texte à des concepts d'une ontologie. La méthode ToMap [13] exploite la structure syntaxique et les similarités de forme des termes. La méthode C-Norm [6] associe par apprentissage profond les représentations vectorielles (*embeddings*) et la structure hiérarchique des ontologies. Nos méthodes pour l'extraction de relation combinent analyse linguistique profonde (résolution d'anaphore et dépendances syntaxiques) et méthodes d'apprentissage à noyau (*shortest path dependency kernel*) [14].

Domaine d'application

Nos domaines d'application en science de la vie, agriculture et alimentation sont variés par exemple, microbiologie [4], biologie végétale [5] et animale [9] sur des thèmes divers tels que la régulation génétique [2], la biodiversité microbienne [10], les phénotypes [11], l'épidémiologie végétale, santé humaine [3] et l'analyse bibliométrique [1].

Nos projets applicatifs en extraction d'information suivent un schéma récurrent : définir un mo-

dèle pour la représentation formelle des informations, construire un corpus pertinent de documents scientifiques, adapter ou concevoir les nomenclatures, terminologies et ontologies nécessaires, annoter manuellement les corpus de référence, concevoir des *workflows* d'entraînement et de prédiction d'entités et de relations, puis lier les prédictions à des données de référence du domaine d'application.

Construction de ressources sémantiques partagées

Nous publions les ressources sous licence ouverte, principalement des corpus annotés (BioNLP-OST) et des ontologies (AgroPortal). Les corpus de référence annotés manuellement sont nécessaires pour entraîner et évaluer des méthodes d'extraction d'information dans les domaines spécialisés de l'INRAE où elles sont rares ou inexistantes.

Nous concevons également des modèles formels et des ontologies qui permettent de normaliser les informations extraites du texte et les rattacher ensuite à des données issues d'autres sources dans un cadre de *linked open data*.

Nos projets de construction de ressources, corpus et ontologies, sont mis en œuvre grâce aux outils logiciels collaboratifs que nous développons et qui favorisent les échanges entre les participants avec des compétences diverses : biologie, traitement automatique de la langue, information scientifique et technique et ingénierie de la connaissance. Nous valorisons les corpus annotés et ontologie dans l'organisation régulière de *shared tasks* internationaux (BioNLP Open Shared Task) [8].

Développement logiciel

L'équipe développe la suite logicielle Alvis de conception de *workflow* de *text mining* à partir d'outils et de contenus pour l'extraction d'information. Elle facilite la mise en place d'expériences, la reproductibilité, la mutualisation des résultats au sein de l'équipe et le transfert. Nous contribuons à l'infrastructure européenne OpenMinTeD de *text mining*, en particulier sur le volet interopérabilité avec l'apport d'une bibliothèque d'outils de traitement automatique de la langue (AlvisNLP/ML) et services pour les sciences de la vie. Les services associés d'annotation (AlvisAE [12]), de visualisation et de recherche d'information (AlvisIR) permettent de visualiser et de communiquer les résultats des traitements aux applications tierce comme l'application Florilege.

Projets

Le projet H2020 OpenMinTeD d'infrastructure de *text mining* fait suite aux projets FP6 Alvis et BPI Quaero pour le développement d'un environnement de développement d'outils et de service de *text mining* pour les spécialistes et non spécialistes. Notre participation au projet ANR D2KAB approfondit ce thème à travers l'adaptabilité des méthodes de *text mining* à différents besoins et domaines et l'intégration avec des données hétérogènes impliquant des alignements sémantiques pour l'implémentation des principes FAIR dans un contexte de science ouverte.

Science ouverte

L'équipe y participe activement à travers son implication dans les e-infrastructures ouvertes (projets H2020 OpenMinTeD et CoSO Visa TM) et à des groupes de travail nationaux sur l'ouverture des publications au *text mining*. Notre objectif est de faciliter l'appropriation des technologies de *text mining* pour la recherche scientifique dans une perspective de Science Ouverte permettant la mutualisation des ressources et la reproductibilité des résultats.

Références

[1] Pascale Avril, Emilie Bernard, Maryse Corvaisier, Agnès Girard, Wiktoria Golik, Claire Nédellec, Marie-Laure Touze, and Nathaële Wacrenier. Analyser la production scientifique d'un département de recherche : construction d'une ressource termino-ontologique par des documentalistes. *Cahier des Techniques de l'INRA*, (89) :1–12, 2016.

[2] Robert Bossy, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Érick Alphonse, Maarten van de Guchte, Philippe Bessières, and Claire Nedellec. BioNLP Shared Task - The Bacteria Track. *BMC Bioinformatics*, 13(S-11) :S3, 2012.

[3] Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. A french clinical corpus with comprehensive semantic annotations : development of the medical entity and relation LIMSI annotated text corpus (MERLOT). *Language Resources and Evaluation*, 52(2) :571–601, 2018.

[4] Estelle Chaix, Louise Deléger, Robert Bossy, and Claire Nédellec. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology*, 81 :63 – 75, 2019. Microbial Spoilers in Food 2017 Symposium.

[5] Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Deléger, Pierre Zweigenbaum, Philippe Bessières, Loïc Lepiniec, and Claire Nedellec. Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop, BioNLP 2016, Berlin, Germany, August 13, 2016*, pages 1–11, 2016.

[6] Arnaud Ferré, Louise Deléger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec. C-Norm : a neural approach to few-shot entity normalization. *BMC Bioinformatics*, 21(23) :1–19, 2020.

[7] Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum, and Claire Nédellec. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[8] Kim Jin-Dong, Nédellec Claire, Bossy Robert, and Deléger Louise, editors. *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, Hong Kong, China, November 2019. Association for Computational Linguistics.

[9] Pierre-Yves Le Bail, Jérôme Bugeon, Olivier Dameron, Alice Fatet, Wiktoria Golik, Jean-François Hocquette, Catherine Hurtaud, Isabelle Hue, Catherine Jondreville, Léa Joret, Marie-Christine Salaun, Jean Vernet, Claire Nédellec, Matthieu Reichstadt, and Philippe Chemineau. Un langage de référence pour le phénotypage des animaux d'élevage : l'ontologie ATOL. *Productions animales*, 27(3) :195–208, 2014.

[10] Claire Nédellec, Robert Bossy, Estelle Chaix, and Louise Deléger. Text-mining and ontologies : new approaches to knowledge discovery of microbial diversity. *CoRR*, abs/1805.04107, 2018.

[11] Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktoria Golik, and Pierre Sourdille. Information Extraction from Bibliography for Marker-Assisted Selection in Wheat. In Sissi Closs, Rudi Studer, Emmanouel Garoufallou, and Miguel-Angel Sicilia, editors, *Metadata and Semantics Research*, pages 301–313, Cham, 2014. Springer International Publishing.

[12] Frédéric Papazian, Robert Bossy, and Claire Nédellec. AlvisAE : a collaborative web text annotation editor for knowledge acquisition. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 149–152, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics.

[13] Zorana Ratkovic, Wiktoria Golik, and Pierre Warnier. Event extraction of bacteria biotopes : a knowledge-intensive NLP-based approach. *BMC Bioinformatics*, 13(S-11) :S8, 2012.

[14] Dialekti Valsamou. *Information Extraction for the Seed Development Regulatory Networks of Arabidopsis Thaliana. (Extraction d'Information pour les réseaux de régulation de la graine chez Arabidopsis Thaliana)*. PhD thesis, University of Paris-Saclay, France, 2017.

LIFO/CA : Contraintes et Apprentissage

LIFO EA 4022/ CA
Université d'Orléans
https://www.univ-orleans.fr/lifo/equipes/CA/

Anaïs LEFEUVRE-HALFTERMEYER
anais.halftermeyer@univ-orleans.fr

Membres

- Cherifa BEN KHELIL (docteure)
- Sylvie BILLOT (MCF)
- Guillaume CLEUZIOU (PR)
- Thi-Bich-Hanh DIEP-DAO (Responsable) (MCF HDR)
- Denys DUCHIER (PR)
- Matthieu EXBRAYAT (MCF)
- Olivier GRACIANNE (doctorant)
- Antoine GUILLAUME (doctorant)
- Houda HARBAOUI (docteure)
- Anaïs LEFEUVRE-HALFTERMEYER (MCF)
- Willy LESAINT (MCF)
- Vincent LEVORATO (chercheur associé)
- Frédéric MOAL (MCF)
- Viet Dung NGHIEM NGUYEN (doctorant)
- Yannick PARMENTIER (MCF associé)
- Marcílio PEREIRA DE SOUTO (PR)
- Marta SOARE (MCF)
- Christel VRAIN (PR)

Thématiques générales de l'équipe

L'équipe est structurée autour de trois axes complémentaires :

- Contraintes : les formalismes logiques permettant la description de problèmes complexes et les techniques associées en assurant la résolution efficace.
- Apprentissage automatique : apprentissage symbolique, numérique et statistique permettant d'extraire automatiquement de masses de données des descriptions utiles et exploitables.
- Traitement automatique des langues (TAL) : formalisation, création d'outils et exploitation de ressources langagières au service du traitement de phénomènes linguistiques.

Description des travaux en traitement du langage humain

Acquisition de taxonomies lexicales L'équipe s'intéresse à l'étude théorique des méthodes d'apprentissage automatique/fouille de données et leur adaptation à des fins d'applications, en particulier dans le contexte de la fouille de textes et de l'acquisition de connaissances à partir de textes bruts. Elle développe des méthodologies d'exploitation de corpus pour la construction automatique ou semi-automatique de structures de connaissances de types "ontologies lexicales". Nous avons défini une nouvelle classe d'espaces prétopologiques et développé des méthodologies d'apprentissage (multi-instances) de ces espaces dans des contextes supervisés et semi-supervisés. Ces travaux ont été appliqués à l'extraction de relations taxonomiques à partir de texte bruts [7, 6, 5].

Méta grammaires L'équipe CA est fortement impliquée dans le développement du formalisme eXtensible MetaGrammaire (XMG). Deux de ses membres avaient participé à la conception et à l'implantation d'une première version du formalisme au cours des années 2000. En 2014, une nouvelle version du formalisme, baptisée XMG2, a vu le jour [15]. Cette version étend le formalisme d'une couche d'abstraction supplémentaire au moyen de briques de langages formels.

L'utilisateur linguiste souhaitant vérifier sa théorie linguistique en implantant une grammaire électronique peut à présent choisir quel langage de description utiliser en puisant dans une bibliothèque de briques de langages qui peuvent être assemblées.

Ces travaux sur les métagrammaires ont notamment permis, en collaboration avec le LLL (Laboratoire Ligérien de Linguistique), la description de la morphologie des noms et des verbes en ikota, une langue bantoue parlée au Gabon, de la syntaxe du Saõ-Tomense, un créole du portugais, et la syntaxe

de l'arabe littéraire. Pour des développements récents autour de l'interface syntaxe-sémantique de l'arabe voir [2, 3].

Grammaires de propriété Nous avons développé une approche en programmation par contraintes pour l'analyse syntaxique des grammaires de propriétés à traits. Cette approche est fondée sur la théorie des modèles [9]. Ce formalisme est adapté au traitement des énoncés agrammaticaux et fournit un environnement permettant à la mise en perspective des contraintes violées en cas d'agrammaticalité.

Traitement de l'oral Nous nous attachons à la production de ressources utiles pour la création et/ou l'entraînement de systèmes d'apprentissage. Notre approche générale se situe dans un TAL "de terrain" afin de produire des ressources et des systèmes permettant de traiter la langue orale.

Nous citons ici nos propositions en collaboration avec le le LIFAT (Laboratoire d'Informatique Fondamentale et Appliquée de Tours) autour de la *syntaxe* à partir du Stanford Parser réentrainé et de son adaptation à l'oral ainsi que de la livraison d'ODIL_Syntaxe, une ressource corrigée issue de cette proposition [16, 17]. Pour ce faire la création de Contemplata, un outil dédié, a été nécessaire et a été publié afin de permettre à quiconque de gérer un projet de production de ressource prenant en entrée du texte et proposant une analyse syntaxique en constituants ainsi que l'annotation sur la structure syntaxique directement [18].

Cet outil a déjà été adapté pour les besoins d'annotation des *injonctions* dans le cadre du projet RAVIOLI en collaboration avec le LLL, le LIFAT et le laboratoire PRISME (Laboratoire Pluridisciplinaire de Recherche en Ingénierie des Systèmes, Mécanique, Énergétique) qui propose de traiter les énoncés à partir de leur analyse syntaxique ainsi que le signal sonore pouvant porter des traces du caractère injonctif.

Dans l'objectif d'extraire des évènements et leur *temporalité* ainsi que de les ordonner, un premier travail a consisté en la revue de la norme ISO-TIMEML afin de couvrir intelligemment les besoins en annotation de la temporalité [1]. Une première tentative de création de système sur la base du corpus TIMEBANK [1] n'a pas été fructueuse et nous travaillons aujourd'hui à la mise en place d'une ressource annotée dans ce but [16]. Nous continuons d'explorer d'autres approches telles que la modélisation fine dans le cadre de l'analyse compositionnelle sémantique des temps conjugués [11] afin d'extraire les évènements présents dans le discours. En lien avec cette dernière question, nous nous penchons depuis peu sur l'extraction d'évènements d'actualité dans des tweets.

Tout comme pour la temporalité nous nous attachons à produire des ressources utiles pour l'apprentissage ainsi que des systèmes pour la résolution de la *coréférence*. Nous avons contribué à la production du corpus ANCOR [14] et affinons actuellement un système opérationnel basé sur l'apprentissage [8], originellement testé avec un SVM et un arbre de décision, nous évaluons actuellement les approches de type réseaux de neurones.

Un travail sur les mécanismes de référenciation des objets du discours et de leur forme de surface est mené en collaboration avec des linguistes s'attachant à ces questions. Dans le cadre de l'ANR TALAD (TAL et Analyse du Discours) l'équipe de Praxiling et celle d'Agora ont mené la rédaction d'une ontologie des concepts utiles à l'étude des phénomènes de nomination, qui cachent un acte de référenciation fin et à laquelle nous avons participé [10]. Plus largement dans le cadre de ce projet ANR, nous travaillons en collaboration avec le LIFAT à l'utilisation de la résolution de coréférence pour guider l'analyste du discours à la découverte des *nominations* d'un même concept [13].

Evaluations et métriques pour le TAL Nous participons à une réflexion sur la méthodologie de production de ressources et d'évaluation des systèmes dans le cadre de la mise au point de systèmes basés sur l'apprentissage automatique. [12] interroge l'interprétabilité des métriques de mesure de la coréférence tandis que [4] interroge plus largement le concept d'accord interannotateur, seule proposition générique à tout corpus annoté en terme de fiabilité.

1. http ://www.usna.edu/Users/cs/nchamber/caevo/

Références

[1] Jean-Yves Antoine, Jakub Waszczuk, Anaïs Lefeuvre-Halftermeyer, Lotfi Abouda, Emmanuel Schang, and Agata Savary. Temporal@ODIL Project : Adapting ISO-TimeML to Syntactic Treebanks for the Temporal Annotation of Spoken Speech. In *Thirteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-13), 12th International Conference on Computational Semantics (IWCS'2017), Montpellier, France.*, Montpellier, France, September 2017.

[2] Chérifa Ben Khelil, Chiraz Zribi, Denys Duchier, and Yannick Parmentier. A semi-automatically generated TAG for Arabic : Dealing with linguistic phenomena. In *19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018)*, Hanoi, Vietnam, March 2018.

[3] Chérifa Ben Khelil, Chiraz Zribi, Denys Duchier, and Yannick Parmentier. Interface syntaxe-sémantique au moyen d'une grammaire d'arbres adjoints pour l'étiquetage sémantique de l'arabe. In *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Rennes, France, May 2018.

[4] Dany Bregeon, Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre-Halftermeyer. Redonner du sens à l'accord interannotateurs : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation. *Traitement Automatique des Langues*, 60(2) :23, September 2019.

[5] G. Cleuziou and J. G. Moreno. Qassit at semeval-2016 task 13 : On the integration of semantic vectors in pretopological spaces for lexical taxonomy acquisition. In *SemEval@NAACL-HLT*, 2016.

[6] Guillaume Cleuziou, Davide Buscaldi, Vincent Levorato, Gaël Dias, and Christine Largeron. QASSIT : A Pretopological Framework for the Automatic Construction of Lexical Taxonomies from Raw Texts. In *International Workshop on Semantic Evaluation (SEMEVAL 2015)*, Denver, United States, 2015.

[7] Guillaume Cleuziou and Gaël Dias. Learning Pretopological Spaces for Lexical Taxonomy Acquisition. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Porto, Portugal, September 2015.

[8] Adèle Désoyer, Frédéric Landragin, Isabelle Tellier, Anaïs Lefeuvre-Halftermeyer, Jean-Yves Antoine, and Marco Dinarelli. Coreference Resolution for French Oral Data : Machine Learning Experiments with ANCOR. In *Computational Linguistics and Intelligent Text Processing.*, volume n° 9623-9624 of *Lecture Notes in Computer Science*. Springer, March 2018.

[9] Denys Duchier, Thi-Bich-Hanh Dao, and Yannick Parmentier. Model-Theory and Implementation of Property Grammars with Features. *Journal of Logic and Computation*, 24(2) :491–509, March 2014. Special issue on "Grammar, Parsing and Recognition". Available at http ://logcom.oxfordjournals.org/content/24/2/491.

[10] Agata Jackiewicz, Nadia Bebeshina-Clairet, Manon Cassier, Francesca Frontini, Anaïs Lefeuvre-Halftermeyer, Longhi Julien, Giancarlo Luxardo, and Damien Nouvel. Vers une ontologie de la nomination et de la référence dédiée à l'annotation des textes. In *13rd Terminology & Ontology : Theories and applications (TOTh) International Conference*, Chambéry, France, June 2019.

[11] Anaïs Lefeuvre-Halftermeyer, Richard Moot, and Christian Retoré. A computational account of virtual travelers in the Montagovian generative lexicon. In Michel Aurnague and Dejan Stosic, editors, *The Semantics of Dynamic Space in French*, Part IV. Formal and computational aspects of motion-based narrations, pages 407–450. John Benjamins, 2019.

[12] Adam Lion-Bouton, Loïc Grobol, Jean-Yves Antoine, Sylvie Billot, and Anaïs Lefeuvre-Halftermeyer. Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ? In Gilles Adda, Maxime Amblard, and Karën Fort, editors, *6e*

conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 10–18, Nancy, France, June 2020. ATALA.

[13] Julien Longhi, Jean-Yves Antoine, Mehdi Mirzapour, Agata Jackiewicz, and Anaïs Lefeuvre-Halftermeyer. Le repérage de nominations dans les corpus textuels : de l'exploitation de l'analyse des données textuelles à l'exploration des chaînes de coréférence par le TAL. In *JADT 2020 : 15es Journées internationales d'Analyse statistique des Données Textuelles*, Toulouse, France, June 2020.

[14] Judith Muzerelle, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau, and Iris Eshkol. ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA, editor, *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, pages 555–563, Les Sable d'Olonne, France, June 2011.

[15] Simon Petitjean, Denys Duchier, and Yannick Parmentier. XMG2 : Describing Description Languages. In Maxime Amblard, Philippe de Groote, Sylvain Pogodalla, and Christian Rétoré, editors, *Logical Aspects of Computational Linguistics (LACL 2016)*, volume 10054 of *Lecture Notes in Computer Science*, pages 255–272, Nancy, France, December 2016. Springer-Verlag. http ://www.springer.com/fr/book/9783662538258.

[16] Ilaine Wang, Jean-Yves Antoine, Lotfi Abouda, Jakub Waszczuk, Aurore Pelletier, and Anaïs Halftermeyer. Annoter la parole spontanée en arbres de constituants pour les besoins de l'analyse temporelle : résultats et comparaison français parlé / français écrit. In *Congrès Mondial de Linguistique Française*, Montpellier, France, July 2020.

[17] Ilaine Wang, Aurore Pelletier, Jean-Yves Antoine, and Anaïs Halftermeyer. ODIL Syntax : a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees. In *Language Resources and Evaluation Conference, LREC*, Marseille, France, May 2020.

[18] Jakub Waszczuk, Ilaine Wang, Jean-Yves Antoine, and Anaïs Halftermeyer. Contemplata, a Free Platform for Constituency Treebank Annotation. In *Language Resources and Evaluation Conference, LREC*, Marseille, France, May 2020.

AfIA
Association française
pour l'Intelligence Artificielle

CLLE : Cognition, Langues, Langage, Ergonomie

CLLE UMR 5263
CNRS et Université de Toulouse
https://clle.univ-tlse2.fr

Ludovic TANGUY
ludovic.tanguy@univ-tlse2.fr

Membres Impliqués

- Cécile FABRE (PR)
- Bruno GAUME (CR)
- Nabil HATHOUT (DR)
- Lydia-Mai HO-DAC (MCF)
- Ludovic TANGUY (MCF HDR)
- Assaf URIELI (membre associé)

Présentation générale

CLLE (Cognition, Langues, Langage, Ergonomie, UMR 5263) est un laboratoire pluridisciplinaire relevant des Sciences Cognitives : les travaux qui y sont menés couvrent les champs, à périmètre plus ou moins large, de la linguistique, de la psychologie, de l'informatique, de la philosophie, de l'éducation et des neurosciences.

Il est actuellement composé de trois équipes :

- Langues et Langage
- Processus langagiers et cognitifs
- Cognition en situation complexe

Les recherches de plusieurs membres de l'équipe Processus langagiers et cognitifs se situent dans le domaine du traitement automatique des langues et de la linguistique outillée. Ces recherches s'intègrent dans une thématique pluridisciplinaire intitulée *Outils Numériques : aspects cognitifs et langagiers* qui regroupe des chercheurs conduisant des travaux en lien avec un dispositif numérique impliqué dans des activités cognitives complexes et intégrant une composante langagière. Ces dispositifs peuvent être des éléments centraux (outils de traitement automatique des langues, systèmes d'information) ou des supports (communication médiée, documents numériques) ; ils peuvent être considérés du point de vue de leur fonctionnement (modèle, efficacité du traitement) ou de leur utilisation (modes d'interaction, ergonomie).

Le recours à des traitements assistés ou automatisés permet aux membres de l'équipe d'aborder des volumes importants de données langagières à des fins d'analyse linguistique, de pouvoir aborder efficacement des données complexes et hétérogènes et aussi d'être des interlocuteurs privilégiés en tant que spécialistes du langage pour collaborer avec d'autres disciplines et répondre à des besoins plus appliqués. Les membres de CLLE participent au dialogue entre la linguistique et les nouvelles techniques de TAL à base d'apprentissage, en utilisant celles-ci tout en gardant un œil critique sur leur articulation avec les connaissances et les modèles théoriques des sciences du langage.

Les principales productions scientifiques des membres de l'équipe sont des méthodes et modèles computationnels dans différents domaines de la linguistique (syntaxe, morphologie, sémantique), des corpus et bases de données lexicales enrichis et annotés, des solutions concrètes pour analyser semi-automatiquement ou automatiquement des données langagières. Toutes ces productions sont rendues accessibles à la communauté *via* le site web REDAC (Ressources Développées à CLLE).

Principaux thèmes de recherche

Analyse distributionnelle

L'analyse distributionnelle regroupe les méthodes qui, à partir de l'observation de leur usage en corpus, permettent d'identifier des similarités sémantiques entre les unités lexicales. Les travaux de CLLE dans ce domaine remontent à plusieurs années, et s'appuient aussi bien sur des méthodes classiques fréquentielles (basées sur la cooccurrence ou l'analyse syntaxique automatique) que sur les méthodes neuronales plus récentes (plongements lexicaux ou *word embeddings*).

Les investigations dans cette thématiques visent à la fois des questionnements fondamentaux sur les principes et les techniques de l'analyse distributionnelle (impact des corpus, évaluation quali-

tative, compositionnalité sémantique [7]), la mise en regard avec des domaines de la linguistique peu confrontés jusqu'ici à ces méthodes (morphologie, sociolinguistique [11]), les conditions de leur utilisation (reproductibilité, petits corpus, domaines de spécialité [8]) et des applications directes (construction de ressources spécialisées, analyse de données issues de tests psycholinguistiques [3]).

Face à un engouement massif et accru pour ces méthodes dans toutes les zones d'activité du TAL, les membres impliqués dans la problématique de l'analyse distributionnelle gardent une point de vue avant tout linguistique sur ces méthodes, et entendent jouer un rôle de premier plan face aux nouvelles questions sur la reproductibilité et l'intelligibilité des modèles neuronaux massivement utilisés en IA pour aborder le langage.

Structuration du lexique

Cette deuxième thématique regroupe un ensemble de travaux autour du lexique, sur les plans sémantique et morphologique, avec une double visée de modélisation et de construction de ressources à large couverture. Sur le plan de la morphologie computationnelle l'équipe est un lieu important dans le champ de la morphologie paradigmatique flexionnelle et dérivationnelle. Les différents travaux de l'équipe ont permis à la fois de développer des modèles paradigmatiques et des base de données morphologiques sur le français (Verbaction, Morphonette, Demonette, etc.) [1, 4].

Le membres de l'équipe mènent des travaux de production de bases lexicales à large couverture en prenant appui sur les dictionnaires collaboratifs (comme GLAFF et GLAWI, construits à partir du Wiktionnaire) en plusieurs langues (français, anglais, italien, serbe) et en proposant des sous-lexiques enrichis et spécifiques (comme Foulophonie qui inventorie les variantes régionales du français ou PsychoGlaff qui ajoute des caractéristiques pertinentes pour la sélection de matériel psycholinguistique) mais aussi des outils et interfaces permettant la manipulation de ces données. Ces bases de données lexicales sont régulièrement utilisées dans la communauté scientifique et pourraient à terme devenir des ressources de référence [5].

Les méthodes à base de graphes lexicaux développées dans CLLE de longue date (travaux de Bruno GAUME sur les marches aléatoires dans les graphes petits mondes) sont, dans le prolongement de travaux plus théoriques, appliquées à des bases de données lexicales et des corpus. Ces réalisations sont accessibles sur le Web (Cillex, Spiderlex, portail lexical du CNRTL, site web Autour du mot). Ces méthodes génériques et robustes s'appliquent à tout type de relations structurantes entre lexèmes et constituent des solutions concrètes pour des besoins en recherche d'information, de classification de document ou d'évaluation à visée psycholinguistique [2]. Les membres de l'équipe produisent des bases de données annotées visant des phénomènes linguistiques spécifiques comme les structures syntaxiques et aspectuelles (Treelex et Treelex++), ou des relations sémantiques en contexte pour la substitution lexicale (jeu d'évaluation SemDis).

Caractérisation et classification linguistique de corpus

CLLE est également le lieu où se réalisent de nombreux travaux en linguistique de corpus dans des domaines et sur des types de textes variés. Le point commun de ces travaux est de proposer des méthodes innovantes en linguistique de corpus outillée, prenant appui sur des données annotées et mobilisant de plus en plus systématiquement des méthodes quantitatives complexes, qu'il s'agisse d'analyses statistiques ou à base d'apprentissage automatique. Ces travaux illustrent parfaitement l'ouverture de l'équipe aux différents niveaux de description linguistique, son rayonnement interdisciplinaire et sa capacité à répondre à des besoins des acteurs socio-économiques. Sans prétendre ici à l'exhaustivité, notons la diversité des données abordées et des approches déployées :

- Rapports d'incidents/accidents aériens : identification des signaux faibles, étude de l'évolution temporelle, classification automatique et interactive (collaborations industrielles avec la société Satefy Data) [9].
- Articles scientifiques : constitution de corpus annotés, caractérisation des contextes linguistiques des citations en lien avec les relations entre auteurs, étude de la structure des titres [6].

- Écrits scolaires : constitution et annotation de corpus, étude de la structure du discours (coréférence), orthographe.
- Commentaires sportifs : constitution et annotation de corpus, étude de la structure syntaxique et prosodique avec des contraintes contextuelles.
- Communications médiées par les réseaux : caractérisation et profilage des échanges sur les forums en ligne (discussions Wikipedia, forums médicaux), étude des marques de l'interaction, conflits et controverses.
- Rapports médicaux : repérage d'entités et extraction d'information.
- Corpus écrits et oraux du français : constitution et annotation, étude des noms sous-spécifiés.

Les membres de l'équipe ont développé un ensemble de compétences autour de l'annotation des données. Ces compétences recouvrent un savoir-faire méthodologique en terme d'annotation humaine ou assistée par ordinateur (notamment au niveau discursif), allant de la définition de guides d'annotation à l'organisation de campagnes avec plusieurs annotateurs. Par ailleurs, l'une des thématiques historiques de CLLE est le développement et l'amélioration d'outils génériques d'annotation automatique de corpus, notamment l'analyseur en dépendances Talismane [10]. Cet outil, développé initialement par Assaf URIELI lors de sa thèse dans le laboratoire, est régulièrement amélioré et étendu.

Références

[1] Gilles Boyé and Gauvain Schalchli. The Status of Paradigms. In Andrew Hippisley and Gregory T. Stump, editors, *The Cambridge Handbook of Morphology*, pages 206–234. Cambridge University Press., 2016.

[2] Bruno Gaume, Karine Duvignau, Emmanuel Navarro, Yann Desalle, Hintat Cheung, S.K. Hsieh, Pierre Magistry, and Laurent Prevot. Skillex : a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing actions. *Traitement Automatique des Langues*, 55(3), 2016.

[3] Bruno Gaume, Ludovic Tanguy, Cécile Fabre, Lydia-Mai Ho-Dac, Bénédicte Pierrejean, Nabil Hathout, Jérôme Farinas, Julien Pinquier, Lola Danet, Patrice Péran, Xavier De Boissezon, and Mélanie Jucla. Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures. In *13th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, Krakow, Poland, 2018.

[4] Nabil Hathout and Fiammetta Namer. Paradigms in word formation : what are we up to ? *Morphology*, 29(2) :153–165, 2019.

[5] Nabil Hathout, Franck Sajous, and Basilio Calderone. GLÀFF, a Large Versatile French Lexicon. In *Proceedings of LREC*, pages 1007–1012, Reykjavik, Iceland, 2014.

[6] Béatrice Milard and Ludovic Tanguy. Citations in scientific texts : do social relations matter ? *Journal of the Association for Information Science and Technology*, 69(11) :1380–1395, 2018.

[7] Bénédicte Pierrejean and Ludovic Tanguy. Towards qualitative word embeddings evaluation : measuring neighbors variation. In *Proceedings of NAACL : Student Research Workshop*, New Orleans, USA, 2018.

[8] L. Tanguy, F. Sajous, and N. Hathout. évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *Traitement Automatique des Langues*, 56(2) :105–129, 2015.

[9] Ludovic Tanguy, Nikola Tulechki, Assaf Urieli, Eric Hermann, and Céline Raynal. Natural language processing for aviation safety reports : from classification to interactive analysis. *Computers in Industry*, 78 :80–95, 2016.

[10] Assaf Urieli and Ludovic Tanguy. L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur talismane. In *Actes de TALN*, 2013.

[11] Marine Wauquier, Cécile Fabre, and Nabil Hathout. Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. In *Congrès Mondial de Linguistique Française (CMLF)*, volume 46 of *6e Congrès Mondial de Linguistique Française*, Mons, Belgium, July 2018. EDP Sciences.

AfIA
Association française
pour l'Intelligence Artificielle

LIX/DaSciM : *Deep Learning for NLP and French Linguistic Resources*

LIX / DaSciM
Ecole Polytechnique, Institut Polytechnique de Paris
http://www.lix.polytechnique.fr/dascim/

Hadi ABDINE
habdine@lix.polytechnique.fr

Yanzhu GUO
yanzhu.guo@lix.polytechnique.fr

Moussa KAMAL EDDINE
kamaleddine@lix.polytechnique.fr

Giannis NIKOLENTZOS
nikolentzos@lix.polytechnique.fr

Stamatis OUTSIOS
soutsios@aueb.gr

Guokan SHANG
gshang@linagora.com

Christos XYPOLOPOULOS
cxypolop@lix.polytechnique.fr

Michalis VAZIRGIANNIS
mvazirg@lix.polytechnique.fr

Introduction

DaSciM (Data Science and Mining) une entité du LIX de l'École Polytechnique, *etc.*) établi en 2013 et depuis lors produisant des résultats de recherche dans le domaine de l'analyse de données à grande échelle via des méthodes d'apprentissage automatique et profond. Le groupe a été spécifiquement actif dans le domaine du NLP et du text mining avec des résultats intéressants au niveau méthodologique et des ressources. Voici nos différentes contributions d'intérêt pour la communauté AFIA.

Représentations graphiques pour le traitement automatique des langues et l'extraction de texte

Ces dernières années, les graphes sont devenus un outil largement utilisé pour la modélisation des données structurées. Pour permettre l'application des approches basées sur les graphes aux données textuelles, les membres de l'équipe DaSciM ont développé l'approche du graphe de mots (graph of words) [15] qui fait correspondre le texte à un graphe où les sommets correspondent aux termes et les arêtes représentent les cooccurrences entre les termes connectés dans une fenêtre de taille fixe. Une fois qu'un document est représenté sous forme de graphe, les algorithmes traditionnels, mais aussi modernes, conçus pour les données structurées en graphe peuvent être appliqués aux textes en langage naturel. Les chercheurs de l'équipe DaSciM ont étudié comment les différentes tâches d'exploration de texte peuvent bénéficier des algorithmes basés sur les graphes. L'une de ces tâches est l'extraction de mots-clés. En capitalisant sur le concept de dégénérescence des graphes, l'algorithme k-core a été appliqué à la représentation graphique du texte pour identifier les sous-graphes cohésifs [16, 21]. Les sommets de ces sous-graphes peuvent être considérés comme les termes les plus importants (c'est-à-dire les mots-clés) d'un document textuel donné. Les membres de l'équipe DaSciM ont également utilisé des algorithmes d'apprentissage automatique qui fonctionnent sur des graphes pour traiter des tâches telles que la catégorisation de textes et l'analyse de sentiments. Les noyaux de graphes [10] et les réseaux neuronaux de graphes

[11], les deux méthodologies dominantes pour l'apprentissage automatique sur les graphes, ont été appliqués à ces problèmes avec beaucoup de succès. L'efficacité des représentations à base de graphes a également été évaluée dans la tâche de *détection de sous-événements* (*detecting sub-events*) à partir de données collectées sur Twitter [7, 8]. L'occurrence d'un sous-événement est généralement associée à un changement dans le contenu des messages postés récemment par les utilisateurs par rapport au contenu des messages postés dans le passé. Un tel changement significatif du contenu se reflète dans la structure de la représentation graphique des tweets et peut être capturé par des approches basées sur les graphes.

Sens des mots et désambiguïsation

Le nombre de sens d'un mot donné, ou polysémie, est une notion très subjective, qui varie considérablement selon les annotateurs et les ressources. La création d'inventaires de sens de mots cohérents et de haute qualité est une condition préalable essentielle à la réussite de la désambiguïsation des sens de mots. Dans [23], les chercheurs de DaSciM proposent une nouvelle approche, entièrement non supervisée et pilotée par les données, pour quantifier la polysémie, en se basant sur la géométrie de base dans l'espace d'intégration contextuelle. L'approche proposée est basée sur des grilles multirésolution dans l'espace d'intégration contextuelle. Ce classement des mots en fonction de leur polysémie, entièrement basé sur des données, peut aider à créer de nouveaux inventaires de sens, mais aussi à valider et à interpréter les inventaires existants. De plus, la nature non supervisée de la méthode la rend applicable à n'importe quelle langue.

Résumé abstractif pour les documents et les réunions

Le résumé abstractif est une tâche importante et difficile, qui nécessite des capacités de compréhension et de génération de langage naturel diverses et complexes. Un bon modèle de résumé doit être capable de lire, de comprendre et d'écrire correctement. Comme la plupart des tâches de NLP, l'état actuel de l'art est basé sur des transformateurs pré-entraînés [22].

Entraînés sur des quantités gigantesques de données brutes et avec des centaines de GPU, les modèles basés sur l'architecture Transformer [22], tels que GPT [14] et BERT [4], ont établi de nouvelles performances de pointe dans chaque tâche NLU. De plus, les utilisateurs du monde entier peuvent facilement bénéficier de ces améliorations en adaptant les modèles pré-entraînés disponibles publiquement à leurs applications spécifiques. Cela permet également de réaliser des économies considérables en termes de temps, de ressources et d'énergie, par rapport à la formation de modèles à partir de zéro.

BART [6] combine un encodeur bidirectionnel de type BERT avec un décodeur de type GPT, et pré-entraîne cette architecture seq2seq comme un autocodeur de débruitage avec une formulation plus générale des objectifs de modélisation du langage masqué de BERT. Puisque non seulement l'encodeur de BART mais aussi son décodeur sont pré-entraînés, BART excelle dans les tâches impliquant la génération de texte.

Ces efforts ont permis de réaliser de grandes avancées. Cependant, la plupart des recherches et des ressources ont été consacrées à la langue anglaise, malgré quelques exceptions notables. Nous remédions en partie à cette limitation en proposant BARThez [2], le premier modèle seq2seq pré-entraîné pour le français formé par l'équipe DaSciM. BARThez : [5], basé sur BART, a été pré-entraîné sur un très grand corpus monolingue français issu de recherches antérieures que nous avons adapté aux schémas de perturbation spécifiques de BART.

Pourtant, alors que le résumé de documents textuels traditionnels est un sujet largement étudié, le résumé de conversations multipartites [3, 9, 17] reste un domaine de recherche comparativement émergent et sous-développé, même s'il a récemment gagné en attention. Cette asymétrie est due en grande partie à la nature de la conversation multipartite, qui pose des défis non rencontrés avec le texte traditionnel, mais aussi à un manque de données et de métriques d'évaluation appropriées. De tels problèmes ont poussé les chercheurs de DaSciM

2. le nom d'un gardien de but français légendaire, Fabien Barthez : https://en.wikipedia.org/wiki/Fabien_Barthez

à développer de nouvelles méthodes qui vont bien au-delà de l'état de l'art pour la tâche de résumé abstrait de réunion ([18]), ainsi que des sous-tâches connexes dans le domaine de la compréhension du langage parlé, telles que la détection de communauté abstraite ([19]) et la classification des actes de dialogue ([20]), comme tremplins vers la génération de meilleurs résumés.

Applications sur le texte juridique

Une application de longue date du traitement automatique des langues aux documents juridiques est l'extraction et la récupération d'informations à partir de décisions judiciaires. L'intérêt pour l'extraction de données à partir de jugements s'explique par le rôle critique qu'ils jouent dans l'administration de la justice, tant dans les systèmes de common law que de droit civil. Dans [2], les membres de l'équipe DaSciM ont utilisé des méthodes NLP pour extraire des informations des arrêts de la Cour d'appel française. Ils ont construit des indicateurs sur la difficulté des performances des avocats et des affaires en utilisant des techniques d'analyse de réseau sur les réseaux des avocats et les réseaux des affaires. L'objectif de cette recherche est d'utiliser ces indicateurs pour guider les non experts lorsqu'ils sont confrontés aux systèmes juridiques et de contribuer à la diminution de l'écart d'accès à la justice en réduisant l'asymétrie d'information qui caractérise le marché juridique.

Ressources linguistiques à grande échelle

Ressources françaises : Les représentations distribuées des mots sont couramment utilisées dans de nombreuses tâches du traitement du langage naturel, en ajoutant que les vecteurs de mots pré-entraînés sur d'énormes corpus de textes ont atteint une haute performance dans de nombreuses tâches différentes du NLP. Dans [1] Les chercheurs de DaSciM ont produit plusieurs vecteurs de mots statiques de haute qualité pour la langue française en utilisant Word2vec CBOW. Deux d'entre eux sont entraînés sur d'énormes données françaises de 33 Go par l'équipe de DaSciM et les autres sont entraînés sur un corpus français déjà existant. Nous estimons également la qualité de nos vecteurs de mots proposés et des vecteurs de mots français existants sur la tâche d'analogie de mots français. En outre, nous effectuons l'évaluation sur plusieurs tâches réelles de NLP qui montrent l'amélioration importante des performances des vecteurs de mots pré-entraînés par rapport aux vecteurs existants et aléatoires.

En plus des vecteurs de mots statiques, nous publions le premier modèle seq2seq pré-entraîné à grande échelle dédié à la langue française, BARThez [5], comportant 165M paramètres, et entraîné sur 101 GB de texte pendant 60 heures avec 128 GPUs. Nous évaluons BARThez sur cinq tâches discriminatives et deux tâches génératives, avec une évaluation automatisée et humaine, et montrons que BARThez est très compétitif par rapport à l'état de l'art. TPour remédier au manque de tâches génératives dans le benchmark FLUE existant, nous avons créé un nouveau jeu de données pour le résumé en français, OrangeSum, que nous publions et analysons dans cet article. OrangeSum est plus abstrait que les jeux de données de résumé traditionnels, et peut être considéré comme l'équivalent français de XSum.

Enfin, nous avons créé une application web [3] pour tester et visionner la qualité de BARThez et les embeddings de mots statiques obtenus. Les embeddings de mots français produits sont disponibles au public, ainsi que le code de réglage fin.

Nous présentons également BERTweetFR, le premier modèle linguistique pré-entraîné à grande échelle pour les tweets français. En tant que ressource précieuse pour les données des médias sociaux, les tweets sont souvent écrits sur un ton informel et présentent un ensemble de caractéristiques propres par rapport aux sources conventionnelles. Il est prouvé que le pré-entraînement adaptatif au domaine offre des avantages significatifs en aidant les modèles à encoder la complexité de domaines textuels spécifiques. Bien que des efforts d'adaptation au domaine des modèles de langage

3. http://master2-bigdata.polytechnique.fr/

à grande échelle aux tweets aient été faits en anglais, il n'existe aucun travail similaire dans une autre langue. Notre modèle est initialisé à l'aide d'un modèle de langage français général CamemBERT qui suit l'architecture de base de BERT. Le pré-entraînement adaptatif est effectué sur 8 GPU V100 avec un jeu de données de 16 Go contenant 226 millions de tweets français, pendant environ 8 jours. Les expériences montrent que BERTweetFR surpasse tous les modèles de langue française du domaine général précédents dans trois tâches NLP Twitter : l'identification de l'offensivité, la reconnaissance des entités nommées et la détection du glissement sémantique. Le jeu de données utilisé dans la tâche de détection de l'offensivité a été créé et annoté par notre équipe, comblant ainsi le manque de tels jeux de données analytiques en français. Nous mettons notre modèle à la disposition du public dans la bibliothèque "Transformers" dans le but de promouvoir la recherche future dans les tâches analytiques pour les tweets en français.

Ressources grecques : Dans [13], le Web grec a été utilisé pour produire un corpus de texte clair à grande échelle et ensuite diverses ressources comme les vecteurs formés, les mots d'arrêt, le vocabulaire, les unigrammes, les bigrammes et les trigrammes. Dans [12] nous avons évalué nos vecteurs nouvellement formés en utilisant deux ensembles de données nouvellement produits : Un jeu de test d'analogie de mots et un jeu de données de similarité de mots (WordSim353). Toutes nos ressources produites sont accessibles au public.

Références

[1] Hadi Abdine, Christos Xypolopoulos, Moussa Kamal Eddine, and Michalis Vazirgiannis. Evaluation of word embeddings from large-scale french web content, 2021.

[2] Paul Boniol, George Panagopoulos, Christos Xypolopoulos, Rajaa El Hamdani, David Restrepo Amariles, and Michalis Vazirgiannis. Performance in the courtroom : Automated processing and visualization of appeal court decisions in france. 2020.

[3] Giuseppe Carenini, Gabriel Murray, and Raymond Ng. Methods for mining and summarizing text conversations. *Synthesis Lectures on Data Management*, 3(3) :1–130, 2011.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[5] Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. Barthez : a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv :2010.12321*, 2020.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv :1910.13461*, 2019.

[7] Polykarpos Meladianos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. Degeneracy-based real-time sub-event detection in twitter stream. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

[8] Polykarpos Meladianos, Christos Xypolopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. An optimization approach for sub-event detection and summarization in twitter. In *European Conference on Information Retrieval*, pages 481–493. Springer, 2018.

[9] Gabriel Murray. *Using speech-specific characteristics for automatic speech summarization*. PhD thesis, Citeseer, 2008.

[10] Giannis Nikolentzos, Polykarpos Meladianos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. Shortest-path graph kernels for document similarity. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1890–1900, 2017.

[11] Giannis Nikolentzos, Antoine Tixier, and Michalis Vazirgiannis. Message passing attention networks for document understanding. In

3. http://archive.aueb.gr:7000/resources/

Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8544–8551, 2020.

[12] Stamatis Outsios, Christos Karatsalos, Konstantinos Skianis, and Michalis Vazirgiannis. Evaluation of greek word embeddings. *arXiv preprint arXiv :1904.04032*, 2019.

[13] Stamatis Outsios, Konstantinos Skianis, Polykarpos Meladianos, Christos Xypolopoulos, and Michalis Vazirgiannis. Word embeddings from large-scale greek web content. *arXiv preprint arXiv :1810.06694*, 2018.

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *https ://openai.com/blog/language-unsupervised/*, 2018.

[15] François Rousseau and Michalis Vazirgiannis. Graph-of-word and tw-idf : new approach to ad hoc ir. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 59–68, 2013.

[16] François Rousseau and Michalis Vazirgiannis. Main core retention on graph-of-words for single-document keyword extraction. In *European Conference on Information Retrieval*, pages 382–393. Springer, Cham, 2015.

[17] Guokan Shang. *Spoken Language Understanding for Abstractive Meeting Summarization*. Theses, Institut Polytechnique de Paris, January 2021.

[18] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 664–674, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[19] Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. Energy-based self-attentive learning of abstractive communities for spoken language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 313–327, Suzhou, China, December 2020. Association for Computational Linguistics.

[20] Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. Speaker-change aware CRF for dialogue act classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 450–464, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[21] Antoine Tixier, Fragkiskos Malliaros, and Michalis Vazirgiannis. A graph degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1860–1870, 2016.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[23] Christos Xypolopoulos, Antoine J-P Tixier, and Michalis Vazirgiannis. Unsupervised word polysemy quantification with multiresolution grids of contextual embeddings. *arXiv preprint arXiv :2003.10224*, 2020.

AfIA
Association française
pour l'Intelligence Artificielle

ERIC : Informatique et mathématiques appliquées pour les humanités numériques

Laboratoire ERIC UR 3083
Université de Lyon
https://eric.msh-lse.fr/

Julien VELCIN
julien.velcin@univ-lyon2.fr

Fadila BENTAYEB
fadila.bentayeb@univ-lyon2.fr

Le laboratoire ERIC, créé en 1995, a été l'un des pionniers dans la fouille des données complexes (*data mining*), un thème phare que l'on retrouve aujourd'hui dans la science des données (*data science*). Il est composé de deux équipes : Data Mining & Decision (DMD) et Systèmes d'Information Décisionnels (SID). Ses chercheurs développent des systèmes, des modèles, des algorithmes qui permettent notamment de traiter (c'est-à-dire nettoyer, stocker, indexer, modéliser, analyser, *etc.*) les données textuelles, mais qui le font en prenant en compte les autres types d'information qui accompagnent le plus souvent le texte, tels que la structure du réseau qui relie ces textes (par ex. les citations), la présence de méta-données (par ex. l'auteur) et le caractère souvent dynamique de l'information (par ex. l'étiquette temporelle) car celle-ci évolue.

Outre le fait de traiter les données textuelles dans le cadre général des données complexes, le laboratoire se distingue par le caractère pluridisciplinaire de ses membres, alliant chercheurs en informatique et en statistique. ERIC se distingue également par l'application de ses travaux à des champs variés, en particulier dans ceux rattachés aux Sciences Humaines et Sociales via la MSH de Lyon St-Etienne.

On peut ainsi citer les récentes collaborations avec des laboratoires en géographie (EVS), en sociologie (Max Weber), en littérature (MARGE) ou en archéologie (ArAr et Archéorient).

Les travaux du laboratoire ne se limitent cependant pas à ce type de partenariats puisqu'on compte également de nombreuses collaborations industrielles (par ex. Orange, EDF, Total).

Modélisation thématique de corpus

L'analyse automatique d'un corpus volumineux peut s'avérer complexe si l'on ne sait pas bien ce que l'on y cherche. Une technique très employée pour résumer un tel corpus est appelée la modélisation thématique (*topic modeling*) qui consiste à structurer l'ensemble des textes à l'aide d'un nombre limité de thématiques, interprétées comme des axes sémantiques permettant d'indexer le corpus. Cette analyse est généralement réalisée de manière totalement non supervisée.

À la suite de travaux pionniers (modèles LSA, pLSA, NMF, LDA), nous avons travaillé sur des modèles permettant de combiner les thématiques avec la polarité de l'opinion (par ex. positive ou négative), et de pouvoir suivre leur évolution dans le temps [7], en collaboration avec l'entreprise AMI Software.

Un travail plus récent a consisté, en collaboration avec le LHC, le LIRMM et le CIRAD, à rendre ces thématiques plus lisibles et à fournir un outil original de navigation appelé Readitopics [14]. D'autres travaux, en collaboration avec EDF (projet DyNoFlu), cherchent à découvrir l'émergence de nouvelles tendances à partir de flux de textes (par ex. des emails). Dans cet axe, une première contribution a permis de mettre en évidence une dynamique particulière de ces tendances dans les espaces de plongement [4].

Par le passé, les thématiques extraites de bulletins d'information avaient été étudiées dans le cadre de l'amélioration d'algorithmes de prévision, par exemple sur le cas de données boursières [9]. À la suite, certains de nos travaux actuels portent sur l'utilisation de sources textuelles pour améliorer la prédiction dans les séries temporelles.

Apprentissage de représentations

La science des données requiert souvent de trouver la représentation la plus adéquate pour résoudre le problème visé, qu'il s'agisse de classification ou

de *clustering* par exemple. Une telle représentation peut être construite en trouvant une base qui reflète la manière dont sont distribuées les données dans l'espace initial, comme par exemple en utilisant une analyse factorielle, ou en cherchant un sous-espace qui déforme le moins les données, comme en apprentissage de variétés (*manifold learning*). Des travaux plus récents utilisent une tâche déterminée (par ex. de classification) pour guider l'apprentissage de ces espaces et que l'on appelle apprentissage de représentations (*representation learning*).

Dans ce contexte, nous avons cherché à développer des modèles d'apprentissage adaptés à des réseaux de documents, c'est-à-dire présentant des informations textuelles et des relations entre ces textes (par ex. données bibliographiques, réseaux sociaux). Nous avons ainsi proposé GVNR qui étend GloVe, modèle initialement prévu pour le plongement de mots, aux graphes et aux réseaux de documents [2]. Nous avons ensuite proposé d'utiliser un mécanisme d'attention, mis en lumière par le succès de l'architecture du Transformer, dans ce formalisme [3]. Une autre voie de recherche a consisté à introduire une notion d'incertitude dans les représentations apprises [8].

Les applications visées avec ces espaces de représentation, dans le cadre d'une collaboration avec l'entreprise DSRT, sont des méthodes automatiques pour recommander des relecteurs potentiels ou des mots-clefs à partir du texte d'un article scientifique.

D'autres travaux ont également été menés récemment sur des données issues des réseaux sociaux, en partenariat avec l'Université de Californie à Davis (USA). Il s'agissait de décrire automatiquement des groupes d'utilisateurs de Twitter à partir d'information textuelle [5, 6].

Entrepôts et lacs de données textuelles

Ces dernières années, l'avènement des mégadonnées (*big data*) et l'émergence de technologies sans modèle ou à modèle fluide, telles que les modèles NoSQL ou les lacs de données (*data lakes*), ont changé nos conceptions de modélisation des systèmes d'information d'aide à la décision. Cela nous a conduits à faire des propositions de recherche pour tenir compte du volume, de la vélocité et de la variété des données dans un entrepôt de données (*data warehouse*). En particulier, nous nous sommes intéressés à la prise en compte des données textuelles dans les systèmes d'aide à la décision.

Dans ce contexte et dans le cadre du projet Tassili en collaboration avec l'Université Saad Dahleb (Algérie), une extension de la notion de cube OLAP (On-Line Analytical Processing) au texte a été proposée en combinant des techniques issues de la recherche d'information, de la fouille de données et des graphes avec l'analyse en ligne. Les mesures (indicateurs) textuelles sont alors présentées sous forme de vecteurs de termes et des opérateurs d'agrégation de documents textuels basés sur la notion de propagation de pertinence ont été définis [11]. Nous avons également intégré le contexte dans les cubes de textes afin d'obtenir des analyses OLAP plus pertinentes [10]. Un autre travail a consisté à définir de nouvelles fonctions d'agrégation pour les données textuelles basées sur les motifs fréquents [1].

Plus récemment, nous avons investi le domaine des lacs de données, concept apparu au début des années 2010 pour répondre aux problèmes induits par l'hétérogénéité des mégadonnées. Un lac de données propose un stockage intégré des données sans schéma prédéfini, ce qui nécessite un système de métadonnées efficace pour les interroger.

Dans ce contexte, nous avons établi une typologie des métadonnées d'un lac en métadonnées intra-objets (propres à un objet en particulier), inter-objets (relations) et globales (sémantiques et d'indexation) [12]. Nous avons ensuite identifié un ensemble de fonctionnalités d'un système de métadonnées. Nous avons proposé ainsi un modèle de métadonnées plus générique et complet, comparé aux systèmes de métadonnées de la littérature : MEDAL (*MEtadata model for DAta Lakes*), qui s'appuie sur notre typologie et adopte une modélisation à base de graphes [13].

MEDAL se décline particulièrement bien pour les lacs de données textuelles. Dans le cadre des projets COREL (relation client) et AURA-PMI (digitalisation et servicisation des PMI de la Région AURA), menés en collaboration avec des chercheurs en sciences de gestion, nous avons adjoint au système de métadonnées une couche logicielle per-

mettant à des utilisateurs non-experts d'effectuer des analyses OLAP, ainsi que des regroupements de documents similaires [12] pour, par exemple, comparer les vocabulaires utilisés dans les rapports financiers d'entreprises.

Références

[1] M. Bouakkaz, Y. Ouinten, S. Loudcher, and P. Fournier Viger. Efficiently mining frequent itemsets applied for textual aggregation. *Appl. Intell*, 48(4) :1013–1019, 2018.

[2] R. Brochier, A. Guille, and J. Velcin. Global Vectors for Node Representations. In *Proceedings of the Web Conference (WWW)*, pages 2587–2593. ACM, 2019.

[3] R. Brochier, A. Guille, and J. Velcin. Inductive Document Network Embedding with Topic-Word Attention. In *Proceedings of the 42nd European Conference on IR Research*, 2020.

[4] C. Christophe, J. Velcin, J. Cugliari, P. Suignard, and Boumghar M. Monitoring geometrical properties of word embeddings for detecting the emergence of new topics. In *Proceedings of The 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[5] I. Davidson, A. Gourru, and S. Ravi. The cluster description problem-complexity results, formulations and approximations. In *Advances in Neural Information Processing Systems*, pages 6190–6200, 2018.

[6] I. Davidson, A. Gourru, J. Velcin, and Y. Wu. Behavioral differences : insights, explanations and comparisons of French and US Twitter usage during elections. *Social Network Analysis and Mining*, 10(6), 2020.

[7] M. Dermouche, J. Velcin, L. Khouas, and S. Loudcher. A joint model for topic-sentiment evolution over time. In *IEEE International Conference on Data Mining*, pages 773–778, 2014.

[8] A. Gourru, J. Velcin, and J. Jacques. Gaussian Embedding of Linked Documents from a Pretrained Semantic Space. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 2020.

[9] T. H. Nguyen, K. Shirai, and J. Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24) :9603–9611, 2015.

[10] L. Oukid, N. Benblidia, F. Bentayeb, O. Asfari, and O. Boussaid. Contextualized Text OLAP Based on Information Retrieval. *International Journal of Data Warehousing and Mining*, 11(2) :1–21, 2015.

[11] L. Oukid, O. Boussaid, N. Benblidia, and F. Bentayeb. A New OLAP Aggregation Operator in Text Cubes. *International Journal of Data Warehousing and Mining*, 12(4) :54–74, 2016.

[12] P. N. Sawadogo, T. Kibata, and J. Darmont. Metadata Management for Textual Documents in Data Lakes. In *International Conference on Enterprise Information Systems*, pages 72–83, 2019.

[13] P. N. Sawadogo, E. Scholly, C. Favre, E. Ferey, S. Loudcher, and J. Darmont. Metadata Systems for Data Lakes : Models and Features. In *1st International Workshop on BI and Big Data Applications*, pages 440–451. Communications in Computer and Information Science, Vol. 1064, Springer, 2019.

[14] J. Velcin, A. Gourru, E. Giry-Fouquet, C. Gravier, M. Roche, and P. Poncelet. Readitopics : make your topic models readable via labeling and browsing. In *27th International Joint Conference on Artificial Intelligence*, pages 5874–5876, Stockholm, Sweden, 2018.

AfIA
Association française
pour l'Intelligence Artificielle

Inalco/ERTIM : Équipe de Recherche Textes, Informatique, Multilinguisme

Inalco EA 2520
http://www.er-tim.fr

Damien NOUVEL
Directeur
damien.nouvel@inalco.fr

Mathieu VALETTE
Directeur adjoint
mathieu.valette@inalco.fr

Membres permanents de l'équipe

- Kata GABOR (MCF)
- Pierre MAGISTRY (MCF)
- Damien NOUVEL (MCF)
- Frédérique SEGOND (associée)
- Mathieu VALETTE (PR)
- Ilaine WANG (IGR)

Textes, Informatique, Multilinguisme

L'équipe ERTIM est l'équipe de recherche spécialisée en Traitement Automatique des Langues (TAL) au sein de l'Institut National des Langues et Civilisations Orientales (Inalco, anciennement Langues O'). Le projet scientifique de l'équipe s'articule autour des thèmes suivants :

- la recherche en sémantique des textes et en analyse du discours,
- le développement de méthodologies pour l'ingénierie des textes et des documents numériques multilingues et la production de ressources multilingues,
- l'acquisition de connaissances.

Les champs disciplinaires dans lesquels l'équipe évolue sont ceux du traitement automatique des langues, des statistiques textuelles, de la terminologie et de l'ingénierie des connaissances, de la didactique, de la linguistique générale (lexicologie textuelle, sémantique textuelle, morphologie lexicale, syntaxe) et du multilinguisme (outils et ressources, en particulier pour les langues peu dotées).

L'équipe est structurée selon les axes suivants.

- *Sémantique de corpus et applications*. Cet axe vise à approfondir les propositions théoriques de la sémantique textuelle, en l'appliquant à l'ingénierie multilingue. Il s'agit notamment d'élaborer des méthodologies de traitement de corpus, de modéliser et de développer des outils de fouille de textes, d'analyse et d'assitance à l'interprétation de textes. Les applications visées sont celles de la recherche d'information, la classification de documents et la fouille de textes.
- *Acquisition des connaissances*. Cet axe porte sur l'élaboration et la mise en œuvre de méthodes pour l'acquisition et le traitement de corpus multilingues et multi-écritures pour la reconnaissance et l'extraction d'informations linguistiques (structuration de lexiques, de terminologies, *etc.*) et de connaissances (ontologies, web sémantique).
- *Technologies éducatives et apprentissage des langues*. Cet axe vise la conception et le développement finalisé de méthodes et d'outils d'apprentissage des langues fondés sur la création de ressources intégrant des techniques à partir de ressources numériques (corpus, lexiques, grammaires) et de TAL (outils d'analyse).
- *Corpus et multilinguisme*. Les thèmes abordés dans cet axe relèvent d'enjeux théoriques et pratiques concernant les corpus monolingues et multilingues, la problématique du multilinguisme dans le traitement automatique du document numérique et la prise en compte technique des spécificités associées (écritures, encodages).

Activités et projets

Les activités de l'équipe s'articulent autour des thèmes et axes présentés et se matérialisent par des projets de recherche décrits ci-dessous. Les méthodes et outils utilisés évoluent rapidement selon les avancées en traitement automatique des langues.

Sémantique textuelle et analyse du discours

L'équipe a une activité historique en sémantique textuelle, que celle-ci soit théorique [21] ou liée à des problématiques TAL [22]. Ces travaux peuvent se décliner pour des tâches particulières, comme par exemple la fouille d'opinion [23, 5, 9].

Le projet ANR *TALAD (2018-2022)* s'inscrit dans cette thématique. Il porte sur l'adaptation de techniques issues du TAL pour apporter à l'analyse du discours des jeux de descripteurs plus complexes. Le projet se concentre sur l'étude des nominations à l'aide d'outils TAL de détection d'entités nommées et de chaînes de coréférence [2].

Nos compétences en TAL pour la sémantique et l'analyse du discours permettent également d'investir d'autres domaines, comme les réseaux et graphes dans le cadre de nos travaux sur l'influence [7, 6, 8].

Dans la continuité de ces travaux, l'équipe s'intéresse actuellement à l'extraction dde structures discursives, comme par exemple à l'occasion d'une collaboration avec EDF R&D sur la détection d'argumentations dans des textes issus des grands débats nationaux [16].

Multilinguisme et langues peu dotées

Par son rattachement au sein de l'Inalco, notre équipe est spécialisée en traitements appliqués à une grande variété de langues, quelque soit leur degré de numérisation et de dotation en resources (corpus, lexiques, logiciels d'analyse).

De nombreux travaux ont été menés, souvent en collaboration avec d'autres équipes et département de l'Inalco, sur l'enseignement et l'apprentissage des langues [19, 14]. Ces travaux ont donné lieu à la mise en place d'outils disponibles sur Internet, comme par exemple des générateurs d'exercices s'appuyant sur des corpus et des outils de traitement morphologiques ou syntaxiques des langues concernées.

D'autres travaux ont une visée essentiellement linguistique. Nous avons par exemple contribué au développement d'un analyseur morpho-syntaxique en bambara (langue mandingue) à partir d'un corpus partiellement désambiguisé et de techniques d'apprentissage automatique [12] dans le cadre du projet Inalco *MANTAL (2014-2017)*.

Considérant un plus large éventail de langues, l'équipe participe à la documentation de leurs outillages en logiciels TAL au travers de la plateforme *MultiTAL (2015-2016)*. Celle-ci recense les outils et ressources pour le traitement automatique des langues orientales et des langues peu dotées [18, 20] et documente leur installation et leur utilisation.

De nombreuses autres langues sont objets des recherches de l'équipe au travers de divers projets et activités, pour en citer quelques-unes : le chinois [23, 5], l'hindi [19], l'arabe [10], les langues africaines [15], le quechua et le guarani [3, 4]. Notre équipe réagit au gré des sollicitations en la matière et décline des méthodes génériques du TAL en tenant compte des spécificités de chaque langue considérée.

Au travers de ces projets, notre équipe élabore des méthodologies génériques permettant de faire avancer les recherches en TAL à destination des chercheurs en humanités numériques : numérisation et transcription de l'écrit et de l'oral, encodages, développement de lexiques et grammaires, mise en œuvre de moteurs de recherche, analyses linguistiques (morphologiques, syntaxiques, discursives, sémantiques). Ce dialogue entre disciplines permet à l'équipe de renforcer la fiabilité des méthodes de TAL génériques et d'enrichir ses compétences pour une grande diversité de langues.

Extraction d'information et de connaissances

Les processus d'acquisition des connaissances nécessitent de faire appel aux dernières méthodes en compréhension du langage, que celles-ci fonctionnent à un niveau lexical ou syntaxique [1], qu'elles reposent sur des méthodes symboliques ou d'apprentissage automatique et d'intelligence artificielle (CRF, BERT).

De nombreux travaux ont été menés sur ce sujet, dont en particulier ceux en rapport avec la détection, la reconnaissance et la liaison d'entités [11, 17].

Cette activité se concrétise aujourd'hui par la mise au point de systèmes cherchant à lier textes et connaissances [13]. Elle suppose de disposer d'outils de TAL robustes et de bases de connaissances correctement De nombreux travaux sont

menés par l'équipe dans ce domaine, en accordant une importance particulière aux méthodologies mises en œuvre (transcription ou annotation de données, évaluation des performances des outils TAL) quelque soit la langue considérées. Le projet DGA *VITAL (2021-2023)* apporte une impulsion à cette activité, en se focalisant sur la mise au point et l'évaluation de méthodes d'apprentissage actif (*active learning*).

Références

[1] Raphaël Bailly and Kata Gábor. Emergence of Syntax Needs Minimal Supervision. In *ACL 2020*, Seattle, United States, July 2020.

[2] Manon Cassier, Julien Longhi, Damien Nouvel, Agata Jackiewicz, Jean-Yves Antoine, and Anaïs Lefeuvre-Halftermeyer. Analysis and Automatic Processing of Discourse. Corpus Linguistics (CL2019), July 2019. Poster.

[3] Johanna Cordova, Capucine Boidin, César Itier, Marie-Anne Moreaux, and Damien Nouvel. Processing quechua and guarani historical texts query expansion at character and word level for information retrieval. In *International Conference on Information Management and Big Data (SIMBIG'18)*, 2018.

[4] Johanna Cordova and Damien Nouvel. Toward creation of Ancash lexical resources from OCR. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 163–167, Online, June 2021. Association for Computational Linguistics.

[5] Qinran Dang. *Brouillard de pollution en Chine. Analyse sémantique différentielle de corpus institutionnels, médiatiques et de microblogues*. Theses, Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O', June 2020.

[6] Kévin Deturck. Détection d'influenceurs dans des médias sociaux (influencer detection in social medias). In *Actes de la Conférence TALN. Volume 2-Démonstrations, articles des Rencontres Jeunes Chercheurs, ateliers DeFT*, pages 117–130, 2018.

[7] Kévin Deturck. *Guide d'annotation en discours pour la détection d'influenceurs*. PhD thesis, Institut National des Langues et Civilisations Orientales, 2021.

[8] Kévin Deturck, Damien Nouvel, and Frédérique Segond. Évaluation comparative d'algorithmes de centralité pour la détection d'influenceurs. In *18ème Conférence Internationale sur l'Extraction et la Gestion des Connaissances (EGC'18)*, 2018.

[9] Egle Eensoo and Mathieu Valette. Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 107–118, 2015.

[10] Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. Arabic natural language processing : an overview. *Journal of King Saud University, Computer and Information Sciences*, 2019.

[11] Ngoc Tan Le, Fatiha Sadat, and Damien Nouvel. Data adaptation for named entity recognition in twitter with features-rich crf. In *Widening Natural Language Processing (WiNLP'18)*, 2018.

[12] Luigi Liu and Damien Nouvel. A bambara tonalization system for word sense disambiguation using differential coding, segmentation and edit operation filtering. In *International Joint Conference on Natural Language Processing (IJCNLP'18)*, 2017.

[13] Cédric Lopez, Elena Cabrio, and Frédérique Segond. Extraction de relations pour le peuplement d'une base de connaissance à partir de tweets. In *EGC2017 - Conférence Extraction et Gestion des Connaissances* , Grenoble, France, January 2017.

[14] Nadia Makouar and Maryvonne Holzem. Enjeux d'une praxis textuelle en éducation : réflexion sur l'apport des sciences de la culture en enseignement-apprentissage des langues. *Questions Vives*, (28), 2017.

[15] Elvis MBONING TCHIAZE, Daniel Baleba, Jean Marc Bassahak, and Ornella Wandji. Building Collaboration-based Resources In Endowed African Languages : Case Of NTeALan Dictionaries Platform. *Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL)*, May 2020.

[16] Elvis MBONING TCHIAZE and Damien Nouvel. NLU-Co at SemEval-2020 Task 5 : NLU/SVM based model apply to characterise and extract counterfactual items on raw data. In *SemEval-2020 (International Workshop on Semantic Evaluation 2020)*, number 1.87 in Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 670–676, Barcelone, Spain, December 2020.

[17] Damien Nouvel, Maud Ehrman, and Sophie Rosset. *Les entités nommées pour le traitement automatique des langues*. ISTE Editions, 2015.

[18] Damien Nouvel, Driss Sadoun, Satenik Mkhitaryan, and Mathieu Valette. Multital : une plateforme numérique en ligne pour recenser les outils tal pour les langues peu dotées. DGA TIM 2017 (présentation), 2017.

[19] Benedicte Parvaz Ahmad. *Production de ressources multilingues pour l'aide à la traduction du droit pénal en hindi, ourdou et français*. Theses, Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O', November 2019.

[20] Driss Sadoun, Satenik Mkhitaryan, Damien Nouvel, and Mathieu Valette. Readme generation from an owl ontology describing nlp tools. In *Natural Language Generation and the Semantic Web (WebNLG'16)*, 2016.

[21] Monique Slodzian and Mathieu Valette. Connaissances prescrites ou connaissances décrites ? L' apport de la sémantique des textes. In *Colloque International sur le document électronique (CIDE. 12)*, pages 129–141, France, 2009. Europia Productions.

[22] Mathieu Valette. Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée. In D. Mayaffre, C. Poudat, L. Vanni, V. Magri, and P. Follette, editors, *International Conference on Statistical Analysis of Textual Data (JADT2016)*, volume 2 of *Proceedings of 13th International Conference on Statistical Analysis of Textual Data, 7-10 June 2016, Nice (France)*, pages 697–706, Nice, France, June 2016.

[23] Liyun YAN, E. Danni, Mei Gan, Cyril Grouin, and Mathieu Valette. Inference Annotation of a Chinese Corpus for Opinion Mining. In *LREC*, Marseille, France, May 2020.

AfIA
Association française
pour l'Intelligence Artificielle

IRISA/EXPRESSION : *Expressiveness in Human Centered Data/Media*

IRISA/ EXPRESSION
Université Bretagne Sud
https://www-expression.irisa.fr/

Pierre-François MARTEAU
pierre-francois.marteau@univ-ubs.fr

Nicolas BÉCHET
nicolas.bechet@irisa.fr

Sylvie GIBET
sylvie.gibet@univ-ubs.fr

Damien LOLIVE
damien.lolive@irisa.fr

Introduction

L'expressivité est un terme utilisé dans un certain nombre de domaines. En biologie, cela fait référence à la génétique et aux phénotypes, alors qu'en informatique, l'expressivité des langages de programmation renvoie à la capacité de formaliser un large éventail de concepts. Nous considérons ici l'expressivité humaine, avec la lecture suivante : l'expressivité est la manière dont un être humain transmet un contenu informationnel influencé par des facteurs tels que le style, la qualité émotionnelle ou l'intention. Compte tenu de cette définition, l'équipe EXPRESSION se concentre sur l'étude des données du langage humain véhiculées par différents médias : le geste, la parole et le texte. Ces données présentent une complexité intrinsèque caractérisée par l'intrication de caractéristiques multidimensionnelles et séquentielles. En outre, ces caractéristiques peuvent ne pas appartenir aux mêmes niveaux de représentation. Fondamentalement, certaines caractéristiques peuvent être symboliques (par exemple les mots, les phonèmes, etc.) tandis que d'autres sont numériques (par exemple des positions, des angles ou des échantillons sonores). Dans ce dernier cas, la séquentialité résulte de la temporalité (exemple des séries temporelles).

Au sein de cette complexité, les données du langage humain intègrent des modèles structurels latents sur lesquels le sens est construit et d'où émergent l'expressivité et la communication. La caractérisation de cette expressivité et plus généralement de la variabilité présente dans les séries temporelles multidimensionnelles, les données séquentielles et les structures linguistiques constituent ainsi le fondement des travaux entrepris par EXPRESSION. Le principal objectif sous-jacent est l'étude des problèmes de représentation et de caractérisation de l'hétérogénéité, de la variabilité et de l'expressivité, en particulier pour l'identification et la catégorisation de motifs ou de comportements ré-exploitables notamment pour la génération de contenus expressifs.

Nos travaux visent l'exploration et la caractérisation de modèles de traitement de données dans trois contextes :

1. Expression dans le texte et la langue,
2. Analyse, synthèse et reconnaissance des gestes expressifs,
3. Analyse et synthèse de la parole expressive.

Quelques travaux clés de l'équipe

Nous listons ci après quelques résultats représentatifs des travaux de l'équipe EXPRESSION.

Parmi les travaux sur l'expressivité dans les données textuelles, nous nous intéressons à la notion de compréhension d'un texte par un lecteur ou auditeur. Cette dernière est conditionnée par l'adéquation des caractéristiques du texte avec les capacités et les connaissances de la personne. Cette adéquation est essentielle dans le cas d'un enfant car ses compétences cognitives et linguistiques sont encore en cours de développement. Plus particulièrement, nous nous intéressons à la tâche qui consiste à prédire l'âge à partir duquel un texte peut être compris par quelqu'un [4, 3, 17]. Ces recherches ont inclus des caractéristiques dérivées du domaine psycholinguistique, ainsi que certaines provenant de tâches

PNL connexes, et des techniques habituelles d'incorporation de mots pour décrire des phrases et des textes. Ensuite, nous avons proposé un ensemble de modèles de réseaux de neurones et les comparons sur un ensemble de données de textes français dédiés à un public jeune ou adulte. Les expériences montrent de bons résultats avec des prévisions qui sont similaires ou meilleures que celles faites par les experts.

Un autre axe important d'étude portant sur le média texte vise le traitement automatique des registres de langues. La notion de registre de langue est une caractéristique fortement perceptible du discours. Cependant, les registres sont encore peu étudiés en traitement du langage naturel. Dans ces travaux, nous avons développé une approche semi-supervisée qui construit conjointement un corpus de textes étiquetés dans des registres et un classificateur associé. Cette approche repose sur une petite graine initiale de données annotées par des experts. Après avoir récupéré de nombreuses pages Web, le principe est d'alterner itérativement entre l'apprentissage d'un classificateur intermédiaire et l'annotation de nouveaux textes pour augmenter le corpus étiqueté. L'approche est appliquée aux registres courants, neutres et soutenus, conduisant à un corpus de 750 millions de mots et à un classificateur neuronal final avec une performance acceptable [10].

Nous avons par ailleurs produit d'autres ressources textuelles comme celle basées sur le "Tenders Electronic Daily" (TED) européen. Ce dernier est une importante source de données semi-structurées et multilingues qui sont très précieuses pour la communauté du traitement du langage naturel. Ces ensembles de données peuvent être utilisés efficacement pour traiter la traduction automatique complexe, l'extraction de terminologie multilingue, l'exploration de texte ou pour comparer les systèmes de recherche d'informations. Malgré les services offerts sur le site Web du TED qui est mis à la disposition du public pour accéder à la publication de l'appel d'offres de l'UE, la collecte et la gestion de ce type de données restent assez difficiles et chronophages. Cela pourrait expliquer pourquoi une telle ressource n'est peu ou pas exploitée par des informaticiens ou des ingénieurs en traitement automatique des langues. Nous avons fourni dans l'équipe deux corpus multilingues documentés et faciles à utiliser (l'un d'eux est un corpus parallèle), extraits de la source Web TED qui sont mis à disposition de la communauté TAL [1].

Dans la lignée des travaux sur les registres de langue, nous nous intéressons également au transfert de style textuel qui consiste à modifier le style d'un texte tout en préservant son contenu. Cela suppose qu'il est possible de séparer le style du contenu. Dans ces travaux, nous examinons si cette séparation est possible [9]. Nous utilisons le transfert de sentiment comme étude de cas pour l'analyse de transfert de style. Notre méthodologie expérimentale définit le transfert de style comme un problème multi-objectif, équilibrant le changement de style avec la préservation du contenu et la fluidité. En raison du manque de données parallèles pour le transfert de style, nous utilisons une variété de réseaux encodeurs-décodeurs contradictoires dans nos expériences. En outre, nous utilisons une méthodologie de sondage pour analyser la façon dont ces modèles codent les caractéristiques liées au style dans leurs espaces latents. Les résultats de nos expériences révèlent un compromis inhérent entre les multiples objectifs de transfert de style et indiquent que le style ne peut pas être utilement séparé du contenu dans ces systèmes de transfert de style.

La manipulation de données textuelles nécessite également une étape préalable afin de "vectoriser" les textes. Nous abordons ces aspects dans les travaux de l'équipe, notamment dans le cadre d'une étude de classification de documents. Nous étudions les méthodes basées sur des plongements de mots (word2vec) ou de documents (analyse sémantique latente, ou sac de mots associées à diverses pondérations) ainsi que certaines combinaisons de ces méthodes [2]. A cette fin, nous évaluons ces méthodes de vectorisation en utilisant trois modèles de classification (un perceptron multicouches, une machine linéaire à vecteurs supports optimisée par descente de gradient stochastique et un classifieur multinomial naïf de Bayes). Les résultats de cette étude ont montré que le modèle proposé pour associer les méthodes word2vec et LSA, qui conjugue les deux caractérisations complémentaires du contexte d'occurrence des mots (local pour word2vec et global pour LSA), permet de produire une vectorisation robuste, en général plus discriminante que les

autres approches testées.

La modélisation, l'analyse et la synthèse du geste humain constitue un autre axe de recherche de l'équipe. Nous mettons l'accent sur l'aspect langagier et expressif de cette modalité de communication, en considérant que le mouvement peut être caractérisé par l'enchaînement d'actions signifiantes et par la façon dont les éléments atomiques sont organisés en structures ayant du sens et véhiculant une certaine forme d'expressivité. Les principaux domaines d'étude concernent les gestes des langues des signes et les gestes impliqués dans les arts performatifs. Les données mouvement sont traitées comme des séries temporelles multidimensionnelles représentées par les trajectoires des positions et angles aux articulations du squelette sous-jacent. D'un point de vue théorique, nous développons des modèles d'inversion et d'optimisation pour la synthèse du mouvement. Ainsi, une nouvelle méthode d'inversion guidée par les distances a été développée pour contrôler des systèmes articulés complexes [15]. Ces modèles d'inversion sont couplés à des modèles d'optimisation à base de représentations différentielles Laplaciennes pour la reconstruction haute résolution de mouvements [16]. Nous nous intéressons également à la classification automatique des émotions véhiculées par les mouvements ou les expressions faciales et à l'annotation automatique à partir de méthodes d'apprentissage automatique.

L'un des principaux domaines applicatifs est celui de la Langue des Signe Française (LSF). Nous avons développé dans l'équipe un pipeline complet *texte-vers-signe* qui permet de composer des phrases en LSF en associant des contenus numériques de différentes natures : données mouvement capturées, données d'annotation traduisant la sémantique des gestes à différents niveaux (phonétique, phonologique, syntaxique, sémantique, expressif). Les mouvements sont ainsi recomposés, assemblés et concaténés en tenant compte des contraintes de co-articulation et de fluidité dans les mouvements. Lorsque les mouvements ne sont pas présents dans la base de données, des méthodes de synthèse s'appuyant sur des contrôleurs dédiés aux différents canaux corporels (tels que corps, mains, expressions faciales, direction du regard), permettent de générer les mouvements adaptés et synchronisés [13]. Les résultats de la synthèse sont comparés à une vérité terrain de données multimodales capturées [14]. Un autre domaine applicatif concerne les gestes musicaux. Nous proposons de contrôler des sons de synthèse par le suivi en temps réel de caractéristiques cinématiques ou d'activités musculaires vibratoires, ou, en reprenant la métaphore du chef d'orchestre, de guider par le geste expressif l'interprétation d'une oeuvre musicale [5].

Par ailleurs, EXPRESSION aborde des travaux qui concernent la détection d'anormalité dans le comportement humain depuis quelques années [8]. Les approches semi-supervisées récemment développées à base d'auto-encodeurs exploitent des fonctions objectifs modifiées pour inclure une composante temporelle [6]. Plus précisément, les auto-encodeurs développés sont entraînés de manière à reconstruire une sous-séquence d'entrée partielle par interpolation des codes latents. Les applications cibles à ce jour concernent la détection d'anomalie dans des séquences de vidéo-surveillance, ainsi que la détection d'intrusion dans des systèmes cyber-physiques impliquant des opérateurs humains. En complément, une approche basée sur une fonction de similarité originale pour la comparaison de chaînes de caractères, calculable efficacement grâce à l'exploitation d'arbres de suffixes, a été développée pour la détection d'anomalie dans des séquences symboliques [11]. L'application cible concerne également la détection d'intrusion par analyse des séquences d'actions engagées par des opérateurs humains. D'autres approches basées sur des forêts dites d'isolation sont explorées [12].

Pour finir, une partie importante des travaux d'EXPRESSION s'articule autour de la synthèse de la parole expressive. Plus particulièrement, afin d'avoir plus de contrôle sur la synthèse de la parole ou *Text-to-Speech* (TTS) et pour améliorer l'expressivité, il est nécessaire de démêler les informations prosodiques portées par l'identité vocale du locuteur de celles appartenant aux propriétés linguistiques. Dans ces travaux, nous proposons d'analyser comment les informations liées à l'identité vocale du locuteur affectent un modèle de synthèse vocale multi-locuteur basé sur un réseau neuronal profond. Pour ce faire, nous alimentons le réseau avec une vectorisation des informations du locuteur en plus d'un ensemble de fonctionnalités linguistiques

de base. Nous comparons ensuite trois principales configurations de codage du locuteur : a) un simple vecteur unique décrivant le sexe et l'identifiant du locuteur ; b) un vecteur d'incorporation extrait d'un modèle pré-entraîné de reconnaissance de locuteur ; c) un vecteur prosodique qui résume des informations telles que la mélodie, l'intensité et la durée. Pour mesurer l'impact du vecteur d'entités en entrée, nous étudions la représentation de l'espace latent à la sortie de la première couche du réseau. L'objectif est d'avoir une vue d'ensemble de la représentation de nos données et du comportement de nos modèles. De plus, nous avons effectué une évaluation subjective pour valider le résultat. Les résultats montrent que l'identité prosodique du locuteur est capturée par le modèle et permet donc à l'utilisateur de contrôler plus précisément la synthèse.

Outre les travaux mentionnés précédemment, et afin de mener à bien nos travaux autour de la synthèse de la parole, il est nécessaire d'utiliser des corpus. En effet, le corpus vocal joue un rôle crucial dans la qualité de la génération de la parole synthétique, spécialement sous une contrainte de longueur. La création d'une nouvelle voix est coûteuse et la sélection du script d'enregistrement pour une tâche de TTS expressive est généralement considérée comme un problème d'optimisation afin d'obtenir un corpus riche et parcimonieux. Afin de vocaliser un livre donné en utilisant un système TTS, nous étudions quatre approches de sélection de script. Sur la base d'observations préliminaires, nous proposons simplement de sélectionner les énoncés les plus courts du livre et de comparer les réalisations de cette méthode avec celles de l'état de l'art pour deux livres, avec des longueurs et des styles d'énoncés différents, en utilisant deux types de systèmes TTS basés sur la concaténation. L'étude des coûts TTS indique que la sélection des énoncés les plus courts pourrait se traduire par une meilleure qualité synthétique, ce qui est confirmé par un test perceptif [19, 18]. En examinant les critères habituels de conception de corpus dans la littérature, comme la couverture des unités ou la similitude de distribution des unités, il s'avère que ce ne sont pas des mesures pertinentes dans le cadre de cette étude. De plus, nous avons étudié l'idée de mélanger des signaux vocaux naturels et synthétiques pour contrôler le compromis entre la qualité globale du livre audio et son coût de production. Les signaux entièrement synthétiques et les signaux mixtes synthétiques et naturels sont tout d'abord comparés en utilisant différents niveaux de qualité synthétique. La perception des auditeurs montre que les signaux mixtes sont préférés. Ensuite, l'ordre et la configuration des signaux mixtes sont étudiés. Le test perceptif ne montre pas de différence significative entre les différentes configurations.

Comme évoqué précédemment, une partie de l'évaluation des travaux de l'équipe repose sur une des campagnes de tests perceptifs multimédia. En ce sens, nous avons développé au sein d'EXPRESSION l'application *FlexEval* qui permet de concevoir et de déployer des tests de perception multimédia sous la forme d'un site Web léger [7]. L'utilisation de technologies Web standard et ouvertes permet à *FlexEval* d'offrir une grande flexibilité de conception, des garanties de durabilité, ainsi qu'un support des communautés d'utilisateurs actives. L'application est open-source et disponible via le référentiel Git https://gitlab.inria.fr/expression/tools/flexeval.

Références

[1] Oussama Ahmia, Nicolas Béchet, and Pierre-François Marteau. Two Multilingual Corpora Extracted from the Tenders Electronic Daily for Machine Learning and Machine Translation Applications. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Myazaki, Japan, May 2018.

[2] Oussama Ahmia, Nicolas Béchet, Pierre-François Marteau, and Alexandre Garel. Utilité d'un couplage entre Word2Vec et une analyse sémantique latente : expérimentation en catégorisation de données textuelles. In Lydia Boudjeloud-Assala Marie-Christine Rousset, editor, *Extraction et Gestion des Connaissances (EGC 2019)*, number 35 in Revue des Nouvelles Technologies de l'Information, pages 129–140, Metz, France, January 2019.

[3] Alexis Blandin, Gwénolé Lecorvé, Delphine Battistelli, and Aline Étienne. Recommandation d'âge pour des textes. In Christophe

Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider, editors, *6e conférence conjointe JEP, TALN, RÉCITAL*, pages 164–171, Nancy, France, 2020. ATALA.

[4] Alexis Blandin, Gwénolé Lecorvé, Delphine Battistelli, and Aline Étienne. Age Recommendation for Texts. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France*, May 2020.

[5] Tiago Brizolara da Rosa, Sylvie Gibet, and Caroline Larboulette. Elemental : a gesturally controlled system to perform meteorological sounds. In *Proceedings of The International Conference on New Interfaces for Musical Expression, NIME 2020, Birmingham, United Kindom*, 2020.

[6] Valentin Durand De Gevigney, Pierre-François Marteau, Arnaud Delhay, and Damien Lolive. Video Latent Code Interpolation for Anomalous Behavior Detection. In *IEEE SMC 2020 - International Conference on Systems, Man, and Cybernetics*, Toronto / Virtual, Canada, October 2020.

[7] Cédric Fayet, Alexis Blond, Grégoire Coulombel, Claude Simon, Damien Lolive, Gwénolé Lecorvé, Jonathan Chevelu, and Sébastien Le Maguer. FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider, editors, *6e conférence conjointe JEP, TALN, RÉCITAL*, pages 22–25, Nancy, France, 2020. ATALA.

[8] Cédric Fayet, Arnaud Delhay, Damien Lolive, and Pierre-François Marteau. Big Five vs. Prosodic Features as Cues to Detect Abnormality in SSPNET-Personality Corpus. In *Interspeech*, Stockholm, Sweden, August 2017.

[9] Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John D Kelleher. Style versus Content : A distinction without a (learnable) difference ? In *International Conference on Computational Linguistics (COLING)*, Virtual, Spain, December 2020.

[10] Gwénolé Lecorvé, Hugo Ayats, Benoît Fournier, Jade Mekki, Jonathan Chevelu, Delphine Battistelli, and Nicolas Béchet. Towards the Automatic Processing of Language Registers : Semi-supervisedly Built Corpus and Classifier for French. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, La Rochelle, France, April 2019.

[11] Pierre-François Marteau. Sequence Covering for Efficient Host-Based Intrusion Detection. *IEEE Transactions on Information Forensics and Security*, 14(4) :994–1006, April 2019.

[12] Pierre-François Marteau. Random Partitioning Forest for Point-Wise and Collective Anomaly Detection - Application to Network Intrusion Detection. *IEEE Transactions on Information Forensics and Security*, 16 :2157–2172, January 2021.

[13] Lucie Naert, C. Larboulette, and S. Gibet. A survey on the animation of signing avatars : From sign representation to utterance synthesis. *Computers & Graphics Journal*, 92 :76–98, 2020.

[14] Lucie Naert, Caroline Larboulette, and Sylvie Gibet. LSF-ANIMAL : A motion capture corpus in french sign language designed for the animation of signing avatars. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France*, pages 6008–6017, May 2020.

[15] Thibaut Le Naour, Nicolas Courty, and Sylvie Gibet. Kinematics in the metric space. *Computer & Graphics Journal*, 84 :13–23, 2019.

[16] Thibaut Le Naour, Nicolas Courty, and Sylvie Gibet. Skeletal mesh animation driven by few positional constraints. *Journal of Computer Animation and Virtual Worlds*, 30(3-4), 2019.

[17] Rashedur Rahman, Gwénolé Lecorvé, Jonathan Chevelu, Nicolas Béchet, Aline Étienne, and Delphine Battistelli. Mama/Papa, Is this Text for Me ? In *International Conference on Computational Linguistics (COLING'2020)*, Virtuel, Spain, December 2020.

[18] Meysam Shamsi, Nelly Barbot, Damien Lolive, and Jonathan Chevelu. Mixing Synthetic and

Recorded Signals for Audio-Book Generation. In *Speech and Computer (SPECOM) 2020*, pages 479–489. September 2020.

[19] Meysam Shamsi, Jonathan Chevelu, Nelly Barbot, and Damien Lolive. Corpus design for expressive speech : impact of the utterance length. In *10th International Conference on Speech Prosody 2020*, pages 955–959, Tokyo, Japan, May 2020. ISCA.

AfIA
Association française
pour l'Intelligence Artificielle

LIG/GETALP : Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole

LIG UMR 5217 / GETALP
CNRS et Université Grenoble Alpes
lig-getalp.imag.fr

Didier SCHWAB
didier.schwab@univ-grenoble-alpes.fr

François PORTET
Responsable d'équipe
françois.portet@univ-grenoble-alpes.fr

Membres permanents de l'équipe

- Véronique AUBERGÉ (CR)
- Valérie BELLYNCK (MCF)
- Hervé BLANCHON (MCF)
- Francis BRUNET-MANQUAT (MCF)
- Maximin COAVOUX (CR)
- Marco DINARELLI (CR)
- Emmanuelle ESPERANÇA-RODIER (MCF)
- Jérôme GOULIAN (MCF)
- Benjamin LECOUTEUX (MCF)
- Mathieu MANGEOT-NAGATA (MCF)
- François PORTET (Pr)
- Fabien RINGEVAL (MCF)
- Solange ROSSATO (MCF)
- Didier SCHWAB (MCF)
- Gilles SÉRASSET (MCF)

Thématique générale de l'équipe

L'équipe GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) est née en 2007 lors de la création du Laboratoire d'Informatique de Grenoble.

Issue de l'union vertueuse de chercheurs en traitement de l'écrit et de la parole, le GETALP est une équipe pluridisciplinaire (informaticiens, linguistes, phonéticiens, traducteurs et traiteurs de signaux, *etc.*) dont l'objectif est d'aborder tous les aspects théoriques, méthodologiques et pratiques de la communication et du traitement de l'information multilingue (écrite ou orale).

La méthodologie de travail du GETALP s'appuie sur des allers-retours continus entre collectes de données, investigations fondamentales, développement de systèmes opérationnels, applications et évaluations expérimentales.

Thématiques de recherche

Les domaines de recherche de GETALP trouvent des applications directes dans divers domaines tels que l'accès à l'information, la robotique, les technologies d'assistance pour les personnes en situation de handicap ou celles qui subissent une perte d'autonomie.

Traduction assistée par ordinateur. Lointaine héritière du CETA (Centre d'Étude en Traduction Automatique) créé dès 1959 par le CNRS, l'équipe a su suivre les évolutions du domaine et s'est ouverte à d'autres thématiques [4]. Depuis 2014, le domaine est confronté à un changement méthodologique majeur avec l'essor des réseaux neuronaux profonds. Des progrès tangibles ont été réalisés ces dernières années [3, 34] et ont contribué à rendre la TA visible et utile pour un large éventail d'applications. Les modèles les plus courants sont composés d'un encodeur bidirectionnel utilisant des unités récurrentes (GRU ou LSTM), associé à un décodeur (également composé de GRU ou LSTM) et pourvu d'un mécanisme d'attention permettant de se concentrer sur une partie spécifique de l'entrée pour produire un mot en sortie [3]. Plus récemment, des modèles très efficaces sans unités récurrentes sont apparus comme le modèle Transformer [34]. L'équipe GETALP a donc pris ce virage méthodologique et a obtenu plusieurs résultats significatifs dans cette thématique.

Nous avons, par exemple, introduit une alternative aux approches actuelles qui s'appuient sur

4. Pour un historique de notre équipe, le lecteur pourra consulter [20]

un réseau neuronal convolutionnel 2D [9] ; contribué à la production, à l'extension et à l'amélioration de corpus multilingues par traduction automatique (TA) et post-édition contributive (PE) [37], et exercé une très forte activité autour de l'évaluation de la traduction automatique qui est un domaine de recherche en soi. Ainsi, nous avons présenté une approche combinant des ressources lexico-sémantiques et des plongements de mots (*word embeddings*) pour l'évaluation en traduction automatique [29].

Transcription et traduction automatique de la parole. GETALP est un acteur incontournable dans le domaine de la reconnaissance automatique de la parole (RAP) et de la traduction automatique de la parole (TAP). On peut citer par exemple des contributions dans de nouvelles directions telles que la prédiction de performance [10] ou la découverte non supervisée d'unités à partir de la parole [27].

L'estimation automatique de la qualité de la traduction orale ([18]) est une tâche relativement nouvelle, définie et formalisée comme un problème d'étiquetage de séquences où chaque mot de l'hypothèse est étiqueté comme bon ou mauvais selon un grand ensemble de caractéristiques. Nous avons proposé plusieurs estimateurs de confiance sur les mots fondés sur une évaluation automatique de la qualité de la transcription (ASR), de la traduction (MT) ou des deux (ASR et MT combinés).

GETALP a également été le premier groupe de recherche à proposer un système de traduction de l'oral de bout-en-bout qui n'utilise aucune transcription symbolique dans la langue source [5]. Une approche similaire a ensuite été proposée et évaluée par des chercheurs de Google [38] avant que nous prolongions notre travail initial en étudiant la traduction de bout en bout de la parole au texte sur un corpus de livres audio – *LibriSpeech* – spécifiquement augmenté pour cette tâche [4].

Traitement des langues sous-dotées. Ce thème a été initié par GETALP il y a 15 ans et reste un domaine d'excellence de l'équipe, en témoignent deux projets ANR récents.

Le projet ALFFA s'est concentré sur le développement des technologies de la parole (ASR et TTS) pour les langues d'afrique subsaharienne [12] tandis que le projet ANR-DFG (franco-allemand) BULB [2] a jeté les bases d'un nouveau domaine de recherche : la documentation des langues assistée par la machine. L'idée est de faire évoluer les méthodologies pour la documentation et la description des langues vers une recherche hautement interdisciplinaire où la linguistique de terrain fait appel à des modèles informatiques et à l'apprentissage automatique.

Traitement / analyse de la parole, des affects sociaux et des interactions dans l'environnement ambiant. GETALP est actif depuis 2000 sur ce thème qui place le traitement de la parole dans l'intelligence ambiante (maison intelligente, smartphones, et plus récemment robots compagnons).

Dans le cadre du projet CIRDO ANR-TECSAN, l'accent a été mis sur la mise au point de technologies vocales pour la détection de situation de détresse des personnes âgées isolées à leur domicile. L'équipe a recueilli des données sur la parole en français chez les personnes âgées et a identifié les facteurs (dépendance) autres que l'âge qui peuvent prédire la performance des systèmes de RAP pour cette population [32]. L'équipe a également développé une chaîne complète de traitement du son en temps réel pour cette tâche (Cirdox) et a mis à disposition un premier corpus audiovisuel [33]. Dans le cadre du projet VocADom (ANR en cours en collaboration avec l'équipe IIHM du LIG), nous abordons les commandes vocales dans un contexte domestique bruité (TV, ventilateur, fond sonore) et avec plusieurs résidents. Le projet est également axé sur l'intégration de la compréhension du langage naturel (NLU) dans le processus d'analyse [8]. Ces projets ont également été l'occasion d'étudier la robustesse de la reconnaissance automatique de la parole dans des conditions d'acquisition ou le ou les micros sont éloignés (spécifique des cas d'utilisation de la maison intelligente) [19].

En ce qui concerne la reconnaissance automatique des émotions [25, 14], l'équipe a proposé de nombreuses contributions originales pour exploiter efficacement les méthodes de l'apprentissage profond pour l'informatique affective. On peut notamment citer l'utilisation de GANs (*Generative Adver-*

sarial Networks) [7] ou des systèmes fondés sur une boucle reconstruction/prédiction [13].

Les comportements affectifs humains ont également été analysés dans le contexte du rire [16], et comme moyen d'effectuer un diagnostic automatique des troubles du spectre autistique [23].

Clarification automatique et interactive du sens. La clarification du sens qui inclut la désambiguïsation lexicale (DL) est une tâche centrale à plusieurs applications du TALN comme, par exemple, la traduction automatique ou la recherche d'information. L'équipe GETALP se concentre sur la désambiguïsation lexicale multilingue. Schématiquement, il s'agit de trouver quelle que soit la langue, quel sens particulier est utilisé pour chacun des mots d'un texte parmi un inventaire de sens prédéfinis. Par exemple, dans la phrase « la souris mange le fromage », il faudra préférer le sens d'animal plutôt que le sens de dispositif électronique. Dans ses recherches, GETALP met un accent particulier sur l'enrichissement et l'exploitation des ressources multilingues et sur l'accès multilingue avec un sens garanti.

Nous étudions comment il est possible de clarifier automatiquement un texte en fonction des ressources disponibles pour une langue donnée. Dans ce cadre, les ressources les plus importantes sont les bases lexicales et les corpus annotés en sens. Avant 2016, les corpus en anglais annotés manuellement se présentaient sous des formats hétérogènes et avec différentes versions de bases lexicales. Pour résoudre ce problème, notre équipe a unifié l'ensemble des corpus annotés en sens (UFSAC – 2 000 000 de mots annotés) [35].

Les autres langues ont très peu de corpus annotés manuellement en sens (au mieux, 10 000 mots). Nous tirons partie de notre corpus anglais unifié et utilisons la traduction automatique pour projeter des annotations dans les langues cibles. Nous avons ainsi publié UFSAC-ara (pour la langue arabe) et UFSAC-fra (pour le français). En ce qui concerne les méthodes, nous utilisons aujourd'hui intensivement des réseaux neuronaux profonds pour WSD et avons proposé plusieurs algorithmes dont [36] qui permettent d'obtenir des résultats état-de-l'art sur l'ensemble des langues testées (anglais, arabe et français). Nous avons également appliqué WSD à la traduction automatique, à la détection du plagiat multilingue et à l'augmentation des ressources lexicales.

Résumé automatique de données ambiantes. Depuis 2015, GETALP est impliqué dans le domaine de la génération automatique du langage naturel (NLG) et s'attaque en particulier à l'une des faiblesses des systèmes actuels, le manque de structures narratives. Pour progresser dans cette direction, l'équipe a proposé en collaboration avec d'autre équipes du laboratoire (IIHM et AMA) une méthode pour générer un *récit* à partir d'un ensemble de données de capteurs (acquises par exemple pendant une activité de ski de randonnée). La chaîne de traitement peut traiter les données des capteurs, extraire les activités pertinentes, les organiser dans un scénario et générer du texte [24]. GETALP s'est également lancé dans les approches de bout en bout pour la génération avec des systèmes qui apprennent conjointement la planification des phrases et la réalisation de surface. L'équipe a étudié plusieurs variantes des modèles de séquence à séquence pour la génération et le résumé à partir d'un large ensemble d'articles de Wikipedia décrivant des entreprises [26]. Une autre application est la synthèse des résultats des votes du Parlement européen (avec l'équipe SIGMA et le LIRIS) par génération de langage naturel et fouille de texte. Le système a remporté le prix de la meilleure démonstration à EGC 2019 [6].

Collecte et interopérabilité des ressources lexicales multilingues. Depuis le début des années 1990 et les travaux de Gilles SÉRASSET, GETALP est fortement impliqué dans la thématique de la structuration et l'interopérabilité des ressources lexicales multilingues. En témoignent, les travaux sur Dbnary qui est une extraction en RDF (Resource Description Framework) des données lexicales de 21 éditions de Wiktionary. Les données linguistiques comprennent les langues suivantes : allemand, anglais, bulgare, espagnol, finnois, français, indonésien, grec, italien, japonais, latin, lituanien, malgache, norvégien, néerlandais, polonais, portugais, russe, serbo-croate, suédois et turc.

DBnary est en partie une duplication des données lexicales disponibles dans de nombreuses éditions linguistiques du projet Wiktionary [28]. Sa valeur ajoutée la plus simple est l'explicitation de beaucoup d'informations lexicales qui ne sont qu'implicitement présentes dans le wiktionnaire original. Cette ressource a été adoptée pour plusieurs cas d'utilisation et applications. En outre, [31] a développé un ensemble d'outils associés à DBnary principalement pour fournir des mesures de similarité sémantique multilingue. La ressource a également été utilisée conjointement avec des plongements de mots pour l'évaluation des systèmes de traduction automatique [29], pour la détection du plagiat multilingue [11] et enfin dans le cadre des travaux de GETALP avec le GipsaLab sur la communication alternative et augmentée.

Enrichissement, amélioration et démocratisation des bases de données lexicales. Cet axe, périphérique à la traduction automatique, peut être divisé en trois sous-thèmes.

Tout d'abord, GETALP développe des environnements permettant une gestion automatisée des ressources lexicales depuis leur importation jusqu'à leur réutilisation par d'autres outils et par leur consultation et édition par les contributeurs. A ce sujet, nous soulignons *iPolex*, un entrepôt de données lexicales [21] et *Jibiki*, une plate-forme générique de gestion de bases de données lexicales à structures hétérogènes Deuxièmement, nous soutenons la création de ressources lexicales en réutilisant les données existantes. Par exemple, *DiLAF* et *iBaatukaay* [17] sont des projets qui visent à créer des bases de données lexicales multilingues pour les langues nationales des pays d'Afrique de l'Ouest : Bambara, Hausa, Kanouri, Serere, Tamajaq, Wolof, Zarma. Nous avons également produit un dictionnaire japonais-français bilingue (154 000 entrées) [22] à partir de versions imprimées de dictionnaires libres de droits, enrichi par des ressources plus récentes telles que Wikipédia et corrigé en ligne par des contributeurs bénévoles. Troisièmement, les bases de données lexicales peuvent également être utilisées pour faciliter l'accès aux textes grâce à des outils de lecture active. Chaque texte est analysé morphologiquement, puis le serveur lexical est consulté pour chaque lemme (voir par exemple le projet *Etymolo* [1]) Un système d'aide à la compréhension des tweets multilingues contenant de l'alternance de code (*code switching*) a également été développé par [30] pendant son doctorat.

Enfin, notre équipe s'est penchée sur l'extraction du sens des textes et des flux textuels produits au cours de processus collaboratifs (courriels et documents textuels d'entreprises) [15].

Références

[1] Slimane Abdellaoui, Valérie Bellynck, Mathieu Mangeot, and Christian Boitet. Outillage de l'accès aux textes par la lecture active étymologique multilingue pour apprenants berbérophones et arabophones. 2018.

[2] Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Breaking the Unwritten Language Barrier : The BULB Project. In *SLTU-2016, 5th Workshop on Spoken Language Technologies for Under-resourced languages, 9-12 May 2016, Yogyakarta, Indonesia*, pages 8–14, 2016.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. End-to-End Automatic Speech Translation of Audiobooks. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Alberta, Canada, April 2018.

[5] Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. Listen and

Translate : A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain, December 2016.

[6] Charles de Lacombe, Antoine Morel, Adnene BELFODIL, François Portet, Cyril Labbé, Sylvie Cazalens, Marc Plantevit, and Philippe Lamarre. Analyse de comportements relatifs exceptionnels expliquée par des textes : les votes du parlement européen. In *Extraction et Gestion des connaissances (EGC)*, volume E-35 of *RNTI*, pages 437–440, Metz, France, January 2019.

[7] Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, and Björn Schuller. Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations. In *7th International Digital Health Conference*, volume 5, pages 53–57, Londres, United Kingdom, July 2017.

[8] Thierry Desot, Stefania Raimondo, Anastasia. Mishakova, François Portet, and Michel Vacher. Towards a French Smart-Home Voice Command Corpus : Design and NLU Experiments. In Sojka P., Horák A., Kopeček I., and Pala K., editors, *21st International Conference on Text, Speech and Dialogue TSD 2018*, volume 11107 of *Lecture Notes in Computer Science, TSD 2018*, pages 509–517, Brno, Czech Republic, September 2018. Springer.

[9] Maha Elbayad, Laurent Besacier, and Jakob Verbeek. Pervasive Attention : 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction. In *CoNLL 2018 - Conference on Computational Natural Language Learning*, pages 97–107, Brussels, Belgium, October 2018. ACL.

[10] Zied Elloumi, Laurent Besacier, Olivier Galibert, Juliette Kahn, and Benjamin Lecouteux. ASR performance prediction on unseen broadcast programs using convolutional neurol networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, April 2018.

[11] Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, and Didier Schwab. Using Word Embedding for Cross-Language Plagiarism Detection. In *EACL 2017*, volume 2, pages 415 – 421, Valence, Spain, April 2017.

[12] Elodie Gauthier, Laurent Besacier, and Sylvie Voisin. Speed perturbation and vowel duration modeling for ASR in Hausa and Wolof languages. In *Interspeech 2016*, San-Francisco, United States, September 2016.

[13] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. Reconstruction-error-based learning for continuous emotion recognition in speech. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, La Nouvelle Orléans (LA), United States, 2017.

[14] Shaoling Jing, Xia Mao, Lijiang Chen, Maria Colomba Comes, Arianna Mencattini, Grazia Raguso, Fabien Ringeval, Björn Schuller, Corrado Di Natale, and Eugenio Martinelli. A closed-form solution to the graph total variation problem for continuous emotion profiling in noisy environment. *Speech Communication*, 104 :66–72, 2018.

[15] Ruslan Kalitvianski, Valérie Bellynck, and Christian Boitet. Un outil de segmentation de courriels imbriqués en courriels individuels et en phrases. In *Atelier Fouille des Données Complexes @ EGC-2017 (Extraction et Gestion des Connaissances)*, Grenoble, France, January 2017.

[16] Reshmashree B Kantharaju, Fabien Ringeval, and Laurent Besacier. Automatic Recognition of Affective Laughter in Spontaneous Dyadic Interactions from Audiovisual Signals. In *International Conference on Multimodal Interaction (ICMI 2018)*, Proceedings of the 20th ACM International Conference on Multimodal Interaction, pages 220–228, Boulder, CO, United States, October 2018. ACM.

[17] Mouhamadou Khoulé, Mathieu Mangeot, and Mamadou Nguer. Manipulation de dictionnaires d'origines diverses pour des langues peu dotées : la méthodologie iBaatukaay. In *Traitement Automatique des Langues Africaines 2018*, Grenoble, France, September 2018.

[18] Ngoc-Tien Le, Benjamin Lecouteux, and Laurent Besacier. Automatic quality estimation for speech translation using joint ASR and MT features. *Machine Translation*, June 2018.

[19] Benjamin Lecouteux, Michel Vacher, and François Portet. Distant Speech Processing for Smart Home Comparison of ASR approaches in distributed microphone network for voice command. *International Journal of Speech Technology*, 21 :601–618, September 2018.

[20] Jacqueline Léon. Le CNRS et les débuts de la traduction automatique en France. *La revue pour l'histoire du CNRS*, 6 :6–24, 2002.

[21] Mathieu Mangeot and Valérie Bellynck. A micro-structure guesser to import or normalize lexical resources. In *Lexicologie Terminologie Traduction LTT 2018*, Grenoble, France, September 2018.

[22] Mathieu Mangeot-Nagata. Collaborative Construction of a Good Quality, Broad Coverage and Copyright Free Japanese-French Dictionary. *International Journal of Lexicography*, 31(1) :78–112, September 2016.

[23] Arianna Mencattini, Francesco Mosciano, Maria Colomba Comes, Tania Di Gregorio, Grazia Raguso, Elena Daprati, Fabien Ringeval, Bjorn Schuller, Corrado Di Natale, and Eugenio Martinelli. An emotional modulation model as signature for the identification of children developmental disorders. *Scientific Reports*, 8(14487), 2018.

[24] Belen Baez Miranda. *Génération de récits à partir de données ambiantes. (Generating stories from ambient data)*. PhD thesis, Grenoble Alpes University, France, 2018.

[25] Francesco Mosciano, Arianna Mencattini, Fabien Ringeval, Björn Schuller, Eugenio Martinelli, and Corrado Di Natale. An array of physical sensors and an adaptive regression strategy for emotion recognition in a noisy scenario. *Sensors and Actuators A : Physical*, 267 :48–59, November 2017.

[26] Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. Generation of Company descriptions using concept-to-text and text-to-text deep models : dataset collection and systems evaluation. In *11th International Conference on Natural Language Generation*, Tilburg, Netherlands, November 2018.

[27] Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stuker, Pierre Godard, Markus Muller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Linguistic unit discovery from multi-modal inputs in unwritten languages : Summary of the "Speaking rosetta" JSALT 2017 workshop. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Alberta, Canada, April 2018.

[28] Gilles Sérasset. DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, 6(4) :355–361, 2015.

[29] Christophe Servan, Alexandre Berard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. Word2Vec vs DBnary : Augmenting METEOR using Vector Representations or Lexical Resources ? In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1159–1168. ACL, 2016.

[30] Ritesh Shah. *SUFT-1, a system for helping understand spontaneous multilingual and code-switching tweets in foreign languages : experimentation and evaluation on Indian and Japanese tweets*. Theses, Université Grenoble Alpes, October 2017.

[31] Andon Tchechmedjiev. *Semantic Interoperability of Multilingual Lexical Resources in Lexical Linked Data*. Theses, Université Grenoble Alpes, October 2016.

[32] Michel Vacher, Frederic Aman, Solange Rossato, François Portet, and B Lecouteux. Making emergency calls more accessible to older

adults through a hands-free speech interface in the house. *ACM Transactions on Accessible Computing* , 12(2) :8 :1–8 :25, June 2019.

[33] Michel Vacher, Saida Bouakaz, Marc-Eric Bobillier-Chaumon, F Aman, Rizwan Ahmed Khan, S Bekkadja, François Portet, Erwan Guillou, S Rossato, and Benjamin Lecouteux. The CIRDO Corpus : Comprehensive Audio/Video Database of Domestic Falls of Elderly People. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 1389–1396, Portoroz, Slovenia, 2016. ELRA, ELRA.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[35] Loïc Vial, Benjamin Lecouteux, and Didier Schwab. UFSAC : Unification of Sense Annotated Corpora and Tools. In *Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, May 2018.

[36] Loïc Vial, Benjamin Lecouteux, and Didier Schwab. Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL)*, Toulouse, France, 2019.

[37] Lingxiao Wang. *Outils et environnements pour l'amélioration incrémentale, la post-édition contributive et l'évaluation continue de systèmes de TA. Application à la TA français-chinois. (Tools and environments for incremental improvement, contributive post-editing and continuous evaluation of MT systems. Application to French-Chinese MT)*. PhD thesis, Grenoble Alpes University, France, 2015.

[38] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly transcribe foreign speech. *CoRR*, abs/1703.08581, 2017.

AfIA
Association française
pour l'Intelligence Artificielle

GIPSA-Lab : Grenoble Images Parole Signal Automatique – Pôle Parole et Cognition

GIPSA-lab/ Pôle Parole et Cognition
CNRS et Univ. Grenoble Alpes, UMR 5216
www.gipsa-lab.fr

Nicolas MARCHAND
Directeur GIPSA-lab
nicolas.marchand@grenoble-inp.fr

Laurent GIRIN
Responsable PPC
laurent.girin@grenoble-inp.fr

Objectif scientifique et organisation

Gipsa-lab est un laboratoire pluridisciplinaire, unité de recherche mixte du CNRS, de Grenoble-INP et de l'Université Grenoble Alpes, développant des recherches fondamentales et finalisées sur les signaux et systèmes complexes. Il est reconnu internationalement pour ses recherches en automatique, signal et images, parole et cognition.

Le Pôle Parole et Cognition (PPC) est l'un des quatre pôles de GIPSA-lab. L'objectif scientifique du pôle est d'étudier et modéliser les signaux de parole, les systèmes physiques et cognitifs qui les produisent et les perçoivent, les systèmes linguistiques qui les organisent et de partager ces connaissances dans le développement de nouvelles technologies vocales. Pour cela les chercheur.e.s du PPC s'appuient sur une approche interdisciplinaire impliquant Sciences Cognitives, Sciences du Langage et Traitement Automatique de la Parole. La démarche scientifique allie observation, recueil de données par expérimentation (comportementale, neurophysiologique) de laboratoire et de terrain, analyse, modélisation et évaluation.

Le PPC compte 22 chercheur.e.s permanent.e.s et est organisé en trois équipes :

- Perception, Contrôle, Multimodalité et Dynamique de la Parole (PCMD) qui s'intéresse à la production et la perception de la parole, ainsi qu'aux systèmes physiques et cognitifs sous-jacents, avec une forte coloration Sciences Cognitives ;
- Systèmes Linguistiques, Dialectologie, Oralité (SYLDO) qui s'intéresse aux relations entre phonétique et phonologie dans différentes modalités de parole, aux dialectes et à la variation, au patrimoine oral, avec une forte coloration Sciences du Langage ;
- Cognitive Robotics, Interactive Systems and Speech Processing (CRISSP) qui s'intéresse au traitement automatique de la parole, aux technologies vocales, à la robotique sociale, avec une forte coloration Sciences et Technologies de l'Information.

Thématiques de recherche

Modélisation de la production et perception de la parole

A la confluence entre Sciences du Langage et Sciences Cognitives, le pôle s'intéresse aux mécanismes de production et perception de la parole et à leurs interrelations, depuis l'étude de leurs fondements neurocognitifs [34, 57] jusqu'à la conception de dispositifs de remédiation des handicaps communicationnels [17]. D'une part, il étudie le contrôle sensorimoteur de la production de parole (EMG [14], capture du mouvement des articulateurs par articulographie électromagnétique [7, 21], imagerie ultrasonore [3], IRM ou vidéo rapide), son développement [27, 48], sa plasticité [15, 28, 33, 58] et l'influence de ce contrôle bouclé sur le décodage du flux sensoriel de la parole d'un autre locuteur [21, 30]. D'autre part, il cherche à mieux comprendre les processus à l'oeuvre dans la perception de la parole [73], ses aspects de fusion multisensorielle [1, 45, 53, 57], ses composantes "top-down" d'attention et de prédiction [70] et sa dynamique temporelle [46, 67]. Au-delà de la psychophysique de ces fonctions, le PPC cherche à en mieux cerner les fondements neuronaux en utilisant l'IRM fonctionnelle, l'électro- ou la magnéto-encéphalographie (EEG, MEG) ou l'électro-corticographie (ECoG, réalisée

au bloc opératoire) [11, 26, 72], et à les modéliser par des processus probabilistes [4, 38, 52, 59].

Une caractéristique de ces travaux est de mettre l'accent sur une cognition incarnée et située par une collecte des données sur le terrain [10, 61] ou dans des conditions les plus écologiques possibles, et en s'interrogeant sur l'implication du corps [18] et l'influence du contexte (social, environnemental) dans les processus de parole et la structuration des systèmes linguistiques. Cette approche facilite le transfert des connaissances fondamentales vers les applications cliniques, en l'occurrence le traitement des troubles de la parole et de la communication [5, 16, 27, 29, 47, 60, 65, 68]. Les recherches du PPC rejoignent ainsi les champs de l'orthophonie et des technologies de suppléance vocale. A titre d'exemple, le PPC est un des moteurs du projet ITN Comm4CHILD qui rassemble de nombreux partenaires européens et dont l'objectif est de caractériser, expliquer et traiter les difficultés de communication des enfants présentant des troubles de l'audition.

Ces travaux s'enracinent dans un contexte local favorable, avec l'essor et la structuration des recherches en sciences cognitives sur le site grenoblois (Pôle Grenoble Cognition, CDP-IDEX Neuro-CoG) dans lesquels le PPC est fortement impliqué. Au niveau national, le PPC est également partie prenante de l'Institut Carnot "Cognition", qui coordonne notamment les interactions avec la recherche privée.

Traitement automatique de la parole, systèmes de communication interactifs et robotique sociale

Ces axes de recherches sont essentiellement menés dans l'équipe CRISSP. Il s'agit d'abord de continuer à apporter des contributions fondamentales et appliquées dans les domaines classiques du traitement automatique de la parole tels que la synthèse vocale [23, 69], le débruitage, la localisation, le tracking et la séparation de sources sonores [2, 35, 36, 41, 42, 43] (en collaboration avec Inria Grenoble).

Ces contributions ouvrent la voie au développement de systèmes de communication interactifs, c'est-à-dire des systèmes réactifs de communication "augmentée" exploitant les caractéristiques multimodales de la parole (son, vision de l'interlocuteur et gestes produits lors de l'interaction) [66]. Une grande partie de ces travaux trouvent leurs applications dans la remédiation du handicap, en orthophonie et dans l'apprentissage des langues [20, 29, 44, 45, 46].

Finalement, une partie des recherches porte sur la robotique cognitive et consiste à développer les capacités socio-communicatives de robots humanoïdes communicants, tel que le robot iCube Nina (photo ci-dessous). Ces recherches impliquent la conception de protocoles expérimentaux avec acquisition simultanée de différents signaux verbaux et co-verbaux sur un ou plusieurs humains et un robot en situation d'interaction (motion capture, eye-tracking, etc.) [12, 51, 54, 55].

Les domaines d'expertises des chercheur.e.s du PPC travaillant sur ces thématiques sont le traitement du signal audio, en particulier le signal de parole (analyse, représentation, filtrage, débruitage, séparation de sources, transformation, reconnaissance automatique, synthèse à partir du texte, synthèse articulatoire, inversion acoustico-articulatoire, etc.) ainsi que la modélisation par apprentissage statistique (machine learning et deep learning).

Linguistique et Humanités Numériques

Un autre domaine d'activités dont les équipes du PPC se sont saisies est celui des Humanités Numériques pour la gestion numérique de données en SHS (stockage, préservation, traitement, développement, mise à disposition auprès de la communauté scientifique, transfert vers la société civile). Le PPC entretient un fond documentaire spécialisé et une atlantothèque contenant la collection d'atlas linguistiques la plus riche d'Europe. Il est coordonnateur de chantiers atlantographiques multilingues fondés sur un réseau très vaste de collaborations scientifiques internationales : Atlas Linguistique Roman (ALiR), Atlas Multimédia Prosodique de l'Espace Roman (AMPER), Atlas Linguistique Multimédia de la région Rhône-Alpes (ALMURA) (https://www.atlas-almura.net/) (e.g. [13, 63]). Il participe à l'élaboration de l'Atlas Linguistique de l'Europe (ALE) et contribue à développer une ap-

proche outillée des questions de traitement et cartographie automatique de données linguistiques et ethnographiques issues de ces atlas [22].

Les activités du PPC s'inscrivent aussi dans des opérations de sauvegarde et valorisation du patrimoine immatériel et matériel. Le pôle a réalisé la numérisation complète d'enregistrements d'enquêtes (lexiques, ethnotextes), dont plus de 450 heures d'enquêtes dialectales dans les domaines franco-provençal, occitan et sarde qui sont désormais disponibles en format numérique pour un large public de chercheurs ou de non-spécialistes. Ont aussi été numérisées 1 920 cartes de l'Atlas Linguistique de la France (ALF) aujourd'hui disponibles en ligne (http://lig-tdcge.imag.fr/cartodialect4/). Ces travaux permettent, à la fois, un traitement des données extensif et multifactoriel en collaboration avec géomaticiens et statisticiens, le partage de ces données et leur conservation pérenne et sécurisée. Ainsi, outre des actions de recherche en linguistique aréale, le pôle poursuit les actions de valorisation et transfert des données du patrimoine culturel vers la société, par exemple vers des programmes muséographiques ou de revitalisation de langues menacées d'extinction (e.g. [19]).

Une autre des spécificités du pôle est d'étudier une diversité de langues d'Afrique, des Amériques, d'Europe et d'Asie dans différentes modalités de production (parole modale, criée, chuchotée, sifflée, tambourinée) [49, 50, 67]. Le PPC contribue également à l'analyse typologique des systèmes et des structures linguistiques (e.g. [6, 63, 71]) à partir de la constitution de corpora et bases de données multilingues ciblant certains types d'unités et de structures (e.g. Grenoble-UCLA Lexical and Syllabic Inventory Database http://g-ulsid.univ-grenoble-alpes.fr).

Une partie des travaux du PPC s'inscrit au niveau national dans les programmes de l'UMS Huma-Num (TGIR CNRS, réseau de moyens techniques et humains) et au niveau local dans l'UMS GRICAD (infrastructure de calcul intensif et de données) et le Grenoble Data Institute qui comporte un pan SHS important.

l'IA au PPC

Ces dix dernières années ont vu une évolution radicale de la méthodologie en traitement automatique des données de parole avec la vague des réseaux de neurones profonds (DNNs pour Deep Neural Networks) et des algorithmes d'apprentissage automatique sur des données massives. Forts de leurs compétences en machine learning pour la modélisation de la parole, notamment par modèles de mélange (voir par exemple [20, 24, 31]), les chercheur.e.s du PPC ont investi ce champ du deep learning en visant à confronter et combiner les connaissances en IA et en cognition pour un enrichissement mutuel. Ainsi, il s'agit d'exploiter dans ces architectures à base de modèles profonds les connaissances, les structures et les modèles identifiés par les études théoriques et expérimentales sur la parole et donc issues des sciences de la parole, des sciences du langage, des sciences cognitives et des neurosciences. A titre d'exemple, on peut citer une étude en modélisation du gain de prédiction de la parole acoustique et audiovisuelle par DNNs [32], le développement d'un synthétiseur articulatoire basé sur des DNNs [8, 9], la détection automatique de cris [37], la synthèse de textures sonores guidée par la perception [62] ou l'auto-apprentissage de localisation de sources sonores par le robot Nina [56]. Une approche originale en traitement de parole et analyse de scènes conversationnelles est de combiner les approches deep avec l'approche par modèles probabilistes Bayésiens traditionnels ; voir par exemple les travaux récents en débruitage de parole avec des auto-encodeurs variationels (VAEs) [39, 40, 64] (en collaboration avec Inria Grenoble) et l'extension du modèle VAE pour le traitement de données séquentielles [25]. Enfin, les modèles profonds sont maintenant au coeur des techniques de modélisation des comportements humains et des interactions sociales portées sur robots humanoïdes [12, 51, 55].

Le PPC est fortement investi dans MIAI (Grenoble Multidisciplinary Institute in Artificial intelligence ; https://miai.univ-grenoble-alpes.fr), un des quatre Instituts Interdisciplinaires d'Intelligence Artificielle (3IA) du plan d'investissement d'avenir. MIAI compte deux chaires portées par des chercheurs du PPC : "Bayesian Cognition and Machine

Learning for Speech Communication" et "Collaborative Intelligent Systems." Le PPC collabore à deux autres chaires "Artificial Intelligence and Language" portée par le Laboratoire d'Informatique de Grenoble et "Audio-Visual Machine Perception and Interaction for Companion Robots" portée par Inria Grenoble. Le PPC bénéficie ainsi d'un soutien de MIAI pour une demi-douzaine de doctorant.e.s travaillant sur un sujet en lien avec l'intelligence artificielle pour la parole. A titre d'illustration, un des objectifs de la première chaire citée est de développer un modèle computationnel d'apprentissage de la perception et production de parole. Ce modèle comporte deux agents virtuels : un adulte et un "très jeune enfant" qui doit apprendre à reconnaître et produire des unités de parole élémentaires à partir d'une faible quantité d'exemples produits par l'adulte.

Références

[1] G. Attigodu, F. Berthommier, C. Vilain, M. Sato, and J.-L. Schwartz. A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. *Frontiers in Psychology*, 5, 2014.

[2] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud. Variational Bayesian inference for audio-visual tracking of multiple speakers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Accepted for publication, 2020.

[3] G. Barbier, P. Perrier, Y. Payan, M. Tiede, S. Gerber, J. Perkell, and L. Ménard. What anticipatory coarticulation in children tells us about speech motor control maturity. *PLoS ONE*, 15(4) :e0231484, 2020.

[4] M.-L. Barnaud, J.-L. Schwartz, P. Bessière, and J. Diard. Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. *PLoS ONE*, 14(1) :e0210302, 2019.

[5] C. Bayard, L. Machart, A. Strauß, S. Gerber, V. Aubanel, and J.-L. Schwartz. Cued speech enhances speech-in-noise perception. *Journal of Deaf Studies and Deaf Education*, 24(3) :223–233, 2019.

[6] E. Biteeva Lecocq, N. Vallée, and D. Faure-Vincent. La phonotaxe du russe dans la typologie des langues : focus sur la palatalisation. In *XXXIIIes Journées d'Études sur la Parole*, Nancy, 2020.

[7] E. Biteeva Lecocq, N. Vallée, S. Gerber, and C. Savariaux. Variabilité du geste linguo-palatal. le cas du russe. In *XXXIIes Journées d'Études sur la Parole*, Aix-en-Provence, 2018.

[8] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert. Robust articulatory speech synthesis using deep neural networks for BCI applications. In *Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014.

[9] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert. Real-time control of an articulatory-based speech synthesizer for brain-computer interfaces. *PLOS Computational Biology*, 12(11), 2016.

[10] J. Bucci, P. Perrier, S. Gerber, and J.-L. Schwartz. Vowel reduction in Coratino (south Italy) : Phonological and phonetic perspectives. *Phonetica*, 2018.

[11] J. Bucci, C. Vilain, J.-L. Schwartz, and S. Dufour. Mapping vowel sounds onto phonemic categories in two regional varieties of French : An ERP study. *Journal of Neurolinguistics*, 2020.

[12] R. Cambuzat, F. Elisei, G. Bailly, O. Simonin, and A. Spalanzani. Immersive teleoperation of the eye gaze of social robots. In *International Symposium on Robotics (ISR)*, 2018.

[13] E. Carpitelli and M. Contini. Atlas linguistique roman. In *Atlas Linguistique Roman (ALiR), volume II.c*, Edizioni dell'Orso, 2019.

[14] T. Cattelain, M. Garnier, C. Savariaux, S. Gerber, and P. Perrier. Analyse électromyographique de la production des plosives labiales : enjeux méthodologiques. In *Journées d'Etude de la Parole*, Aix en Provence, France, 2018.

[15] T. Caudrelier, J.-L. Schwartz, P. Perrier, S. Gerber, and A. Rochet-Capellan. Transfer of learning : What does it tell us about speech

production units ? *Journal of Speech, Language, and Hearing Research*, 61(7) :1613–1625, 2018.

[16] D. Caussade, F. Gaubert, M. Sérieux, N. Henrich Bernardoni, J.-M. Colletta, and N. Vallée. Hand gestures and speech impairments in spoken and sung modalities in people with Alzheimer's disease. In *Gestures and Speech in Interaction (GESPIN)*, pages 67–72, 2015.

[17] L. Chasseur, M. Dohen, B. Lecouteux, S. Riou, A. Rochet-Capellan, and D. Schwab. Evaluation of the acceptability and usability of augmentative and alternative communication (AAC) tools : the example of pictogram grid communication systems with voice output. In *ACM Conference on Computers and Accessibility (SIGACCESS)*, Athens, Greece, 2020.

[18] M. Cherdieu, O. Palombi, S. Gerber, J. Troccaz, and A. Rochet-Capellan. Make gestures to learn : Reproducing gestures improves the learning of anatomical knowledge more than just seeing gestures. *Frontiers in Psychology*, 8 :1689 :1–15, 2017.

[19] G. Depau. Diffusion and transmission of francoprovençal : A study of speakers' linguistic conscience. In *French language policies and the revitalisation of regional languages in the 21st century*, pages 129–148, 2019.

[20] D. Fabre, T. Hueber, L. Girin, X. Alameda-Pineda, and P. Badin. Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. *Speech Communication*, 93(9) :63–75, 2017.

[21] M. Garnier, L. Ménard, and B. Alexandre. Hyper-articulation in Lombard speech : An active communicative strategy to enhance visible speech cues ? *Journal of the Acoustical Society of America*, 144(2) :1059 – 1074, 2018.

[22] P. Genou, M. Seffar, P. Garat, C. Chagnaud, and C. Chauvin-Payan. Les désignations du pissenlit en domaine gallo-roman. *Studia Linguistica Romanica*, 4 :à paraitre, 2020.

[23] B. Gerazov, G. Bailly, and Y. Xu. A weighted superposition of functional contours model for modelling contextual prominence of elementary prosodic contours. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.

[24] L. Girin, T. Hueber, and X. Alameda-Pineda. Extending the cascaded Gaussian mixture regression framework for cross-speaker acoustic-articulatory mapping. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(3) :662–673, 2017.

[25] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda. Dynamical variational autoencoders : A comprehensive review. *arXiv preprint arXiv :2008.12595*, 2020.

[26] R. Grandchamp, L. Rapin, M. Perrone-Bertolotti, C. Pichat, C. Haldin, E. Cousin, J.-P. Lachaux, M. Dohen, P. Perrier, M. Garnier, M. Baciu, and H. Loevenbruck. The ConDialInt Model : Condensation, dialogality, and intentionality dimensions of inner speech within a hierarchical predictive control framework. *Frontiers in Psychology*, 10 :2019, 2019.

[27] B. Grandon, M.-J. Martinez, A. Samson, and A. Vilain. Long-term effects of cochlear implantation on the intelligibility of speech in French-speaking children. *Journal of Child Language*, 47(4) :881–892, 2020.

[28] C. Haldin, A. Acher, L. Kauffmann, T. Hueber, E. Cousin, P. Badin, P. Perrier, D. Fabre, D. Pérennou, O. Detante, A. Jaillard, H. Loevenbruck, and M. Baciu. Speech recovery and language plasticity can be facilitated by Sensori-Motor Fusion (SMF) training in chronic non-fluent aphasia. A case report study. *Clinical Linguistics & Phonetics*, 32(7) :1 – 27, 2017.

[29] C. Haldin, H. Loevenbruck, T. Hueber, V. Marcon, C. Piscicelli, P. Perrier, A. Chrispin, D. Pérennou, and M. Baciu. Speech rehabilitation in post-stroke aphasia using visual illustration of speech articulators : A case report study. *Clinical Linguistics & Phonetics*, pages 1–24, 2020.

[30] A. Hennequin, A. Rochet-Capellan, S. Gerber, and M. Dohen. Does the visual channel improve the perception of consonants produced by speakers of French with Down syndrome ?

Journal of Speech, Language, and Hearing Research, 61(4) :957–972, 2018.

[31] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly. Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12) :2246–2259, 2015.

[32] T. Hueber, L. Girin, and J.-L. Schwartz. Evaluating the potential gain of auditory and audiovisual speech predictive coding using deep learning. *Neural Computation*, 32(3) :596–625, 2020.

[33] T. Ito, J. Coppola, and D. Ostry. Speech motor learning changes the neural response to both auditory and somatosensory signals. *Scientific Reports*, 6 :25626, 2016.

[34] T. Ito, H. Ohashi, and V. Gracco. Changes of orofacial somatosensory attenuation during speech production. *Neuroscience Letters*, 2020.

[35] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud. A variational EM algorithm for the separation of moving sound sources. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, 2015.

[36] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud. Exploiting the intermittency of speech for joint separation and diarization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, 2017.

[37] P. Laffitte, Y. Wang, D. Sodoyer, and L. Girin. Assessing the performances of different neural networks architectures for the detection of screams and shouts in public transportation. *Expert Systems With Applications*, 117 :29–41, 2019.

[38] R. Laurent, M.-L. Barnaud, J.-L. Schwartz, P. Bessière, and J. Diard. The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. *Psychological Review*, 124(5) :572–602, 2017.

[39] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud. A recurrent variational autoencoder for speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, 2020.

[40] S. Leglaive, L. Girin, and R. Horaud. Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, 2019.

[41] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud. Online localization and tracking of multiple moving speakers in reverberant environments. *IEEE Journal of Selected Topics in Signal Processing*, 13(1) :88–103, 2019.

[42] X. Li, L. Girin, R. Horaud, and S. Gannot. Estimation of the direct-path relative transfer function for supervised sound source localization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11) :2171–2186, 2016.

[43] X. Li, L. Girin, R. Horaud, and S. Gannot. Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10) :1007–2012, 2017.

[44] L. Liu and G. Feng. A pilot study on mandarin chinese cued speech. *American Annals of the Deaf*, 164-4 :496–518, 2019.

[45] L. Liu, G. Feng, D. Beautemps, and X.P. Zhang. A new re-synchronization method based multi-modal fusion for automatic continuous French cued speech recognition. *IEEE Transaction on Multimedia*, à paraitre, 2020.

[46] L. Liu, G. J., Feng, and X.P. Zhang. Automatic detection of the temporal segmentation of hand movements in British English cued speech. In *Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, 2019.

[47] L. Machart, A. Vilain, H. Lœvenbruck, G. Meloni, and C. Puissant. Production de parole

chez l'enfant porteur d'implant cochléaire : apport de la Langue française Parlée Complétée. In *Journées d'Études sur la Parole*, pages 388–396, Nancy, France, 2020.

[48] L. Ménard, P. Perrier, and J. Aubin. Compensation for a lip-tube perturbation in 4-year-olds : Articulatory, acoustic, and perceptual data analyzed in comparison with adults. *Journal of the Acoustical Society of America*, 139(5) :2514–2531, 2016.

[49] J. Meyer. Coding human languages for long-range communication in natural ecological environments : Shouting, whistling, and drumming. *In Coding Strategies in Vertebrate Acoustic Communication. Animal Signals and Communication*, 2020.

[50] J. Meyer, L. Dentel, and F. Meunier. Categorization of natural whistled vowels by naïve listeners of different language background. *Frontiers in Psychology*, 8 :25, 2017.

[51] A. Mihoub, G. Bailly, C. Wolf, and F. Elisei. Graphical models for social behavior modeling in face-to-face interaction. *Pattern Recognition Letters*, 74 :82–89, 2016.

[52] C. Moulin-Frier, J. Diard, J.-L. Schwartz, and P. Bessière. COSMO : A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics*, 53 :5–41, 2015.

[53] O. Nahorna, F. Berthommier, and J.-L. Schwartz. Audio-visual speech scene analysis : Characterization of the dynamics of unbinding and rebinding the McGurk effect. *Journal of the Acoustical Society of America*, 137(1) :362–377, 2015.

[54] D.-C. Nguyen, G. Bailly, and F. Elisei. Learning off-line vs. on-line models of interactive multimodal behaviors with recurrent neural networks. *Pattern Recognition Letters*, 100 :29–36, 2017.

[55] D.-C. Nguyen, G. Bailly, and F. Elisei. Comparing cascaded LSTM architectures for generating head motion from speech in task-oriented dialogs. In *International Conference on Human-Computer Interaction*, 2018.

[56] Q. Nguyen, L. Girin, G. Bailly, F. Elisei, and D.-C. Nguyen. Autonomous sensorimotor learning for sound source localization by a humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) – Workshop on Crossmodal Learning for Intelligent Robotics*, Madrid, 2018.

[57] R. Ogane, J.-L. Schwartz, and T. Ito. Orofacial somatosensory inputs modulate word segmentation in lexical decision. *Cognition*, 197, 2020.

[58] H. Ohashi and T. Ito. Recalibration of speech perception due to orofacial somatosensory inputs during speech motor adaptation. *Journal of Neurophysiology*, 2019.

[59] J.-F. Patri, P. Perrier, J.-L. Schwartz, and J. Diard. What drives the perceptual change resulting from speech motor adaptation ? Evaluation of hypotheses in a Bayesian modeling framework. *PLoS Computational Biology*, 14(1) :e1005942, 2018.

[60] L. Rapin, M. Dohen, and H. Loevenbruck. Les hallucinations auditives verbales. In S. Pinto & M. Sato, editor, *Traité de Neurolinguistique - Du cerveau au langage*, pages 347–370. De Boeck Supérieur, 2016.

[61] A. Remacle, M. Garnier, S. Gerber, D. Claire, and C. Pétillon. Vocal change patterns during a teaching day : Inter- and intra-subject variability. *Journal of Voice*, 32(1) :57–63, 2018.

[62] F. Roche, T. Hueber, S. Limier, and L. Girin. Autoencoders for music sound modeling : a comparison of linear, shallow, deep, recurrent and variational models. In *Sound and Music Computing Conference (SMC)*, Malaga, 2019.

[63] A. Romano, P. Boula de Mareüil, and J.-P. Lai. Prosodie du corse. In *Manuel de linguistique corse*, S. Retali-Medori (Ed). De Gruyter, Berlin, 2020.

[64] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud. Audio-visual speech enhancement using conditional variational auto-encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :1788–1800, 2020.

[65] L. Scarbel, D. Beautemps, J.-L. Schwartz, and M. Sato. Sensory-motor relationships in speech production in post-lingually deaf cochlear-implanted adults and normal-hearing seniors : Evidence from phonetic convergence and speech imitation. *Neuropsychologia*, 101 :39 – 46, 2017.

[66] T. Schultz, M. Wand, T. Hueber, D. Krusienski, C. Herff, and J. Brumberg. Biosignal-based spoken communication : A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12) :2257–2271, 2017.

[67] F. Seifart, J. Meyer, S. Grawunder, and L. Dentel. Reducing language to rhythm : Amazonian bora drummed language exploits speech rhythm for long-distance communication. *Royal Society Open Science*, 5 :170–354, 2018.

[68] A. Slis, M. Garnier, A. Da Fonseca, and C. Savariaux. Glottal characteristics of people who stutter and the interaction with syllable complexity. In *International Congress of Phonetic Sciences (ICPhS)*. 2019.

[69] B. Stephenson, L. Besacier, L. Girin, and T. Hueber. What the future brings : Investigating the impact of lookahead for incremental neural TTS. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

[70] A. Strauß and J.-L. Schwartz. The syllable in the light of motor skills and neural oscillations. *Language, Cognition and Neuroscience*, 32(5) :562–569, 2017.

[71] T. T. H. Tran, N. Vallée, and L. Granjon. Effects of word position on the acoustic realization of vietnamese final consonants. *Phonetica*, 76 :1–30, 2019.

[72] A. Treille, C. Vilain, S. Kandel, and M. Sato. Electrophysiological evidence for a self-processing advantage during audiovisual speech integration. *Experimental Brain Research*, 235(9) :2867–2876, 2017.

[73] A. Vilain, M. Dole, H. Loevenbruck, O. Pascalis, and J.-L. Schwartz. The role of production abilities in the perception of consonant category in infants. *Developmental Science*, 22(6) :e12830, November 2019.

AfIA
Association française
pour l'Intelligence Artificielle

GREYC : Groupe Recherche en Informatique, Image, Automatique et Instrumentation Caen

GREYC UMR 6072
CNRS, Normandie Université, École Nationale Supérieure d'Ingénieurs de Caen et Université de Caen Normandie
https://www.greyc.fr

Gaël DIAS
Directeur adjoint
gael.dias@unicaen.fr

Membres

- Navneet AGARWAL (doctorant)
- Usama AHMED (doctorant)
- Houssam AKHMOUCH (doctorant)
- Céline ALEC (MCF)
- Judith Jeyafreeda ANDREW (ATER)
- Anaëlle BALEDENT (doctorante)
- Pierre BEUST (MCF HDR)
- Gaël DIAS (PR)
- Emmanuel GIGUET (CR HDR)
- Amit KUMAR (doctorant)
- Yann MATHET (MCF HDR)
- Kirill MILINTSEVICH (doctorant)
- Fabrice MAUREL (MCF)
- Justine REYNAUD (MCF)
- Marc SPANIOL (PR)
- Antoine WIDLÖCHER (MCF)

Stratégie scientifique

Le traitement automatique des langues (TAL) et la recherche d'information (RI) sont deux thématiques historiques du GREYC. En particulier, une approche différentielle du langage naturel tient de fil rouge, et consiste à définir des modèles qui soient (le plus possible) indépendants de la langue, du genre ou du domaine utilisés. Ainsi, ces modèles peuvent être développés dans le cadre d'applications réelles multilingues sans que de nouveaux paramétrages ou apprentissages spécifiques à la langue soient nécessaires. Comme média préférentiels, les chercheurs du GREYC traitent des données hétérogènes et multilingues du web, en portant un intérêt particulier aux dispositifs nomades et à l'accessibilité des contenus pour les déficients visuels. Ainsi, les techniques d'apprentissage supervisé, non-supervisé, semi-supervisé, par renforcement, profond (*deep learning*) sont au cœur des activités de recherche en TAL et RI du GREYC.

Traitement automatique des langues et sémantique

Comprendre l'interaction entre les éléments constitutifs du langage pour en dégager du sens est une étape fondamentale pour la réussite des applications du TAL. Dans ce cadre, les chercheurs du GREYC s'évertuent à proposer des modèles de représentation du sens du langage naturel. En particulier, trois axes principaux sont abordés : la sémantique dénotative, la sémantique connotative et la sémantique morpho-dispositionnelle.

La sémantique dénotative s'intéresse au sens fondamental et stable d'une unité lexicale ainsi que de ses relations avec les autres unités lexicales. Dans ce cadre, des travaux sur l'extraction d'unités polylexicales (ou multi-mots) [5] et sur l'identification de relations lexico-sémantiques entre unités de sens (par apprentissage profond [23, 13] et par approche statistique [8]) ont été proposés. Également, des modèles d'organisation des unités (poly-)lexicales en ressource sémantique (ou ontologie lexicale) ont été développés par couplage de la théorie de la prétopologie et de l'apprentissage semi-supervisé ou auto-supervisé [4].

Dans la sémantique connotative, l'intérêt ne porte pas sur le sens littéral d'une unité lexicale mais sur les éléments de sens qui peuvent s'ajouter à celle-ci. La temporalité est la connotation que le GREYC étudie en priorité. Ainsi, différentes méthodes de propagation par apprentissage semi-supervisé ont été proposées pour associer à chaque *synset* de WordNet sa connotation temporelle. Ces travaux ont donné lieu à la création d'une ressource

langagière appelée TempoWordNet [19], à partir de laquelle de nombreuses applications ont pu émerger [20, 21].

En ce qui concerne la sémantique morpho-dispositionnelle, l'idée sous-jacente tient du fait que la mise en page participe à l'organisation sémantique des énoncés et qu'elle inclut une dimension sémantique supplémentaire à la compréhension du langage. Dans ce cadre, un modèle de transposition à l'oral de la sémantique morpho-dispositionnelle a été proposé pour une intégration de la structure visuelle des textes dans les systèmes *Text-to-Speech* (TTS) [3].

Digestion de l'information

Face à l'augmentation exponentielle de l'information sur le web, il est crucial d'en comprendre l'essence pour n'en retranscrire que l'essentiel. Dans ce cadre, les chercheurs du GREYC s'intéressent particulièrement au résumé de textes, au partitionnement éphémère et à l'enrichissement sémantique, et contribuent ainsi à développer la thématique de la digestion d'information [6] en s'appuyant sur une compréhension sémantique des textes.

Dans le cadre du résumé de texte, l'objectif est de réduire la taille d'un document sans en perdre la capacité informationnelle. Les travaux les plus significatifs dans ce domaine regroupent la segmentation thématique [7] avec la définition d'une mesure de similarité distributionnelle informationnelle, et la réduction phrastique à partir de techniques d'apprentissage non supervisé (programmation logique inductive) pour la découverte de règles de réécriture simplifiée [18].

Le partitionnement éphémère consiste à regrouper selon un ou plusieurs critères donnés (par ex. thématique, temporel, émotionnel) les documents récupérés par un moteur de recherche en réponse à une requête. Il permet ainsi de comprendre la diversité d'une collection de textes. Dans ce cadre, les chercheurs du GREYC ont proposé l'algorithme Dual C-means dont l'originalité réside sur le calcul simultané des classes et des étiquettes de classe dans un cadre polythétique, et la définition d'un critère de partitionnement optimal [16]. Dans le cadre du partitionnement temporel, une mesure de similarité symétrique agrégative de troisième ordre a été proposée pour évaluer la similitude entre une unité de sens et une expression temporelle [28].

Pour hisser l'analyse des textes au niveau sémantique, et non plus seulement opérer au niveau des mots-clefs, des modèles ont été proposés pour relier les entités nommées à leurs entités canoniques [14]. Ainsi, les informations sur les entités peuvent ensuite être utilisées pour de nombreuses applications, comme par exemple la datation automatique de photographies [27], la représentation des contenus du web par *empreinte sémantique* [11] ou l'étude de la viralité de l'information [10].

Dynamique de l'information

Comprendre l'information du web selon une acception dynamique correspond à étudier les évolutions des unités informationnelles participant au texte selon un axe temporel. En effet, la conservation et l'organisation des données d'Internet ne permettent pas seulement d'écrire l'histoire des contenus numériques d'origine, mais aussi de capter l'air du temps de différentes périodes couvrant plus d'une décennie. L'étude de ces données longitudinales est communément appelée *web analytics*.

L'hypothèse de recherche est que les événements (par ex. des nouveautés ou des changements dans l'opinion publique) sont interdépendants et se manifestent par certaines cooccurrences. Donc, pour pouvoir comprendre les dépendances entre le contenu du web et la connaissance sociale correspondante, il faut tracer et exploiter systématiquement les contenus produits par des communautés d'utilisateurs (même dans plusieurs langues) [22]. Dans ce cadre, les chercheurs du GREYC ont développé plusieurs systèmes qui permettent (1) de prédire l'évolution des taxonomies [25], (2) d'aligner automatiquement des bases de connaissances structurées [26], ou hétérogènes dans différentes langues [24], (3) de prédire la diffusion d'un évènement dans des communautés parlant une langue étrangère [12], et (4) d'analyser des documents web en fonction des entités nommées qu'ils contiennent [9].

Évaluation en TAL et RI

L'évaluation est une discipline de recherche à part entière qui est trop souvent délaissée par ses

acteurs. Ainsi, les chercheurs du GREYC proposent de développer des méthodes d'évaluation pertinentes dans le domaine des technologies du langage humain.

D'une part, alors que beaucoup de données annotées sont produites pour l'apprentissage, leur mise à disposition ne devrait se faire que dans la mesure où leur consistance est etablie. Cela est souvent fait en procédant à l'annotation multiple de mêmes données, et en observant dans quelle mesure les différents annotateurs sont d'accord, grâce aux classiques "mesures d'accord inter-annotateurs". Cependant, l'annotation en TAL se fait sur des structures continues (texte, audio, vidéo) sur lesquelles les annotateurs doivent par eux-mêmes identifier et positionner des "unités". Il est donc nécessaire de disposer de mesures d'accord prenant en compte cette spécificité, les mesures standard de type Kappa étant dédiées à l'annotation d'items prédéfinis, et donnant lieu dans de tels cas à des résultats biaisés. Notre équipe a donc conçu et développé les mesures Gamma [35, 34] qui s'appuient sur un processus unifié (i.e. simultané) d'alignement des annotations des différents annotateurs et du calcul de l'accord qui en résulte, ce qui permet d'obtenir des valeurs d'accord plus pertinentes (ce qui a été établi via des expériences spécifiques [35]).

D'autre part, l'évaluation du partitionnement est une tâche complexe pour laquelle plusieurs mesures de performance existent mais qui représentent toutes un biais particulier. Ainsi, étudier les performances d'un algorithme de partitionnement ne peut se faire que sous le prisme d'un ensemble de métriques. Dans ce cadre, nous avons proposé une nouvelle métrique qui permet de prendre en compte le caractère non balancé des classes découvertes [15].

Handicap et santé mentale

Les applications du GREYC en TAL et RI se concentrent majoritairement autour du handicap et de la santé mentale.

Dans le cadre du handicap, les recherches se focalisent sur l'accès à l'information du web pour les déficients visuels. Ainsi, plusieurs dispositifs qui intègrent des modèles théoriques pour le balayage (*scanning*) ou le survol (*skimming*) d'une page web à partir de son partitionnement en clusters cohérents [17] ont été développés. En particulier, un dispositif haptique permet d'appréhender la structure d'un document à partir du toucher sur une tablette tactile [33] (voir figure ci-dessous). Parallèlement, une transposition orale de la structure visuelle des contenus textuels est possible grâce à une architecture TTS concurrente [2].

Dans le cadre de la santé mentale, plusieurs études ont été menées sur le diagnostic automatique de la dépression à partir de l'analyse d'entretiens patient-thérapeute. Ainsi, différents modèles de fusion précoce ont été proposés afin de mieux combiner les modalités visuelles, textuelles et acoustiques pour la régression du score PHQ-8 [31]. Ces modèles ont ensuite été améliorés par l'apport de différentes stratégies d'apprentissage multitâches, notamment par le couplage classification/régression [32] et le couplage régression/classification de la dépression/régression des émotions [30]. Plus récemment, nous nous sommes intéressés au transfert de style [1] et à la génération de dialogues multi-parties [29] avec des modèles encodeurs-décodeurs pour la création d'agents conversationnels incarnés, à terme applicables au diagnostic précoce des troubles mentaux.

Références

[1] Rane C., Dias G., Lechervy A., and Ekbal A. Improving neural text style transfer by introducing loss function sequentiality. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, 2021.

[2] Maurel F., Dias G., Ferrari S., Andrew J-J., and Giguet E. Concurrent speech synthesis to improve document first glance for the blind. In *Proceedings of the 2nd International Workshop on Human-Document Interaction (HDI*

2019) associated to 15th International Conference on Document Analysis (ICDAR 2019), 2019.

[3] Maurel F., Mojahid M., Vigouroux N., and Virbel J. Documents numériques et transmodalité. transposition automatique à l'oral des structures visuelles des textes. *Document Numérique*, 9(1) :25–42, 2006.

[4] Cleuziou G. and Dias G. Learning pretopological spaces for lexical taxonomy acquisition. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2015)*, 2015.

[5] Dias G. Multiword unit hybrid extraction. In *Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2003)*, pages 41–49, 2003.

[6] Dias G. Information digestion, 2010. HDR Thesis. Université d'Orléans.

[7] Dias G., Alves E., and Lopes J.G.P. Topic segmentation algorithms for text summarization and passage retrieval : An exhaustive evaluation. In *22nd Conference on Artificial Intelligence (AAAI 2007)*, pages 1334–1340, 2007.

[8] Dias G., Moraliyski R., Cordeiro J.P., Doucet A., and Ahonen-Myka H. Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. *Journal of Natural Language Engineering (JNLE 2010)*, 16(4) :439–467, 2010.

[9] Govind, Kumar A., Alec C., and Spaniol M. CALVADOS : A Tool for the Semantic Analysis and Digestion of Web Contents. In *Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019)*, pages 84–89, 2019.

[10] Govind, Alec C., and Spaniol M. ELEVATE-Live : Assessment and Visualization of Online News Virality via Entity-Level Analytics. In *Proceedings of 18th International Conference on Web Engineering (ICWE 2018)*, pages 482–486, 2018.

[11] Govind, Alec C., and Spaniol M. Fine-grained Web Content Classification via Entity-level Analytics : The Case of Semantic Fingerprinting. *Journal of Web Engineering*, 17(6&7) :449–482, 2019.

[12] Govind and Spaniol M. ELEVATE : A Framework for Entity-level Event Diffusion Prediction into Foreign Language Communities. In *Proceedings of the 9th International ACM Web Science Conference (WebSci 2017)*, pages 111–120, 2017.

[13] Akhmouch H., Dias G., and Moreno J. Understanding feature focus in multitask settings for lexico-semantic relation identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

[14] Hoffart J., Yosef M.A.and Bordino I., Fürstenau H., Pinkal M., Spaniol M., Thater S., and Weikum G. Robust disambiguation of named entities in text. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 782–792, 2011.

[15] Moreno J. and Dias G. Adapted b-cubed metrics to unbalanced datasets. In *Proceedings of the 38th Annual ACM SIGIR Conference (SIGIR 2015)*, 2015.

[16] Moreno J., Dias G., and Cleuziou G. Query log driven web search results clustering. In *Proceedings of the 37th Annual ACM SIGIR Conference (SIGIR 2014)*, pages 777–786, 2014.

[17] Andrew J-J., Ferrari S., Maurel F., Dias G., and Giguet E. Web page segmentation for non visual skimming. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2019)*, 2019.

[18] Cordeiro J.P., Dias G., and Brazdil P. Unsupervised induction of sentence compression rules. In *Proceedings of the Workshop on Language Generation and Summarisation of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian*

Federation of Natural Language Processing (ACL/IJCNLP 2009), pages 15–22, 2009.

[19] Hasanuzzaman M., Dias G., Ferrari S., and Mathet Y. Propagation strategies for building temporal ontologies. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 6–11, 2014.

[20] Hasanuzzaman M., Saha S., Dias G., and Ferrari S. Understanding temporal query intent. In *Proceedings of the 38th Annual ACM SIGIR Conference (SIGIR 2015)*, 2015.

[21] Hasanuzzaman M., Sze W.L., Salim M.P., and Dias G. Collective future orientation and stock markets. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI 2016)*, 2016.

[22] Spaniol M. *A Framework for Temporal Web Analytics*. habilitation, Université de Caen Basse-Normandie, 2014.

[23] Bannour N., Dias G., Chahir Y., and Akhmouch H. Learning lexical-semantic relations using intuitive cognitive links. In *Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020)*, 2020.

[24] Boldyrev N., Spaniol M., and Weikum G. Multi-Cultural Interlinking of Web Taxonomies with ACROSS. *The Journal of Web Science*, 3(1), 2017.

[25] Prytkova N., Spaniol M., and Weikum G. Predicting the Evolution of Taxonomy Restructuring in Collective Web Catalogues. In *Proceedings of the 15th International Workshop on the Web and Databases*, 2012.

[26] Prytkova N., Spaniol M., and Weikum G. Aligning multi-cultural knowledge taxonomies by combinatorial optimization. In *Proceedings of the 24th International Conference on World Wide Web (WWW 2015)*, pages 93–94, 2015.

[27] Martin P., Spaniol M., and Doucet A. Temporal Reconciliation for Dating Photographs Using Entity Information. In *Proceedings of the 8th Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 39–41, 2015.

[28] Campos R., Dias G., Jorge A., and Nunes C. Identifying top relevant dates for implicit time sensitive queries. *Information Retrieval Journal*, 2017. ISSN : 1573-7659.

[29] Kumar R., Chauhan D., Dias G., and Ekbal A. Modelling personalized dialogue generation in multi-party settings. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN 2021)*, 2021.

[30] Qureshi S.A., Dias G., Hasanuzzaman M., and Saha S. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 2020. ISSN : 1556-603X.

[31] Qureshi S.A., Hasanuzzaman M., Saha S., and Dias G. The verbal and non verbal signals of depression - combining acoustics, text and visuals for estimating depression level. *CoRR*, abs/1904.07656, 2019.

[32] Qureshi S.A., Saha S., Hasanuzzaman M., and Dias G. Multi-task representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 2019. ISSN : 1541-1672.

[33] Safi W., Maurel F., Routoure J-M., Beust P., Molina M., Sann C., and Guilbert J. Blind navigation of web pages through vibro-tactile feedbacks. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology (VRST 2019)*, 2019.

[34] Mathet Y. The agreement measure γ_{cat} a complement to γ focused on categorization of a continuum. *Computational Linguistics*, 43(3) :661–681, 2017.

[35] Mathet Y., Widlöcher A., and Métivier J-P. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3) :437–479, 2015.

AfIA
Association française
pour l'Intelligence Artificielle

INA : Analyse des médias à l'Institut National de l'Audiovisuel

Service de la Recherche / Département Recherche et Innovation
Institut National de l'Audiovisuel
https://institut.ina.fr/

Boris JAMET-FOUNIER
bjametfournier@ina.fr

Membres impliqués

- Abdelkrim BELOUED
- Olivier BUISSON
- Jean-Hugues CHENOT
- David DOUKHAN
- Moritz HENGEL (doctorant)
- Nicolas HERVÉ
- Louis LABORELLI
- Steffen LALANDE
- Quentin LEROY (doctorant)
- Pierre LETESSIER
- Jean-Etienne NOIRÉ
- Zeynep PEHLIVAN
- Thomas PETIT (doctorant)
- Agnès SAULNIER
- Paul TOMI
- Rémi URO (doctorant)
- Laurent VINET

Introduction

L'INA, Institut national de l'audiovisuel, créé en 1974, assume les missions d'archivage, de recherche et de création audiovisuelle, ainsi que de formation professionnelle. Riche d'une collection de plus de 22 millions d'heures de contenus audiovisuels, il assure le dépôt légal de la radio, de la télévision et du web média, rend accessible ses fonds à des fins de recherche dans les centres de consultation INAthèque, et commercialise un très important fonds d'archives.

La recherche à l'INA

Aux carrefours des mondes académique et industriel, des sciences du numérique et des sciences sociales, du passé et du futur, la recherche de l'INA marie données, intelligence artificielle, analyse et synthèse de l'image et du son, pour préserver, comprendre et valoriser le patrimoine média national. Atelier de l'audiovisuel et laboratoire des médias, la Recherche de l'INA construit des outils pratiques et fait avancer la connaissance sur les médias. Elle se définit à la fois par ses objets et ses thématiques. Les objets de recherche sont, d'une part, les données de l'INA (radio, télévision, web, documentation, métadonnées) enrichies de corpus extérieurs (presse, corpus scientifiques) et, d'autre part, les cadres d'usage actuels ou pressentis des usager·e·s (internes et externes) et des client·e·s de l'Institut. Il s'agit ainsi de concevoir des outils et des méthodologies permettant de renouveler la manière dont l'INA appréhende ses collections et ses missions, dans leur gestion interne (numérisation, documentation) et dans leur usage par les client·e·s (journalistes, producteur·rice·s) et les usager·e·s (chercheur·se·s, grand public). Les thématiques de recherche couvrent plusieurs champs disciplinaires : numérisation, traitement de signal, intelligence artificielle, apprentissage automatique, web sémantique, analyse, fouille et visualisation de données. Ces domaines sont d'autant plus variés que la Recherche s'inscrit dans une approche transdisciplinaire en collaborant avec des chercheur·se·s en sciences humaines et sociales dans le cadre fertile des humanités numériques.

Technologies du Langage Humain

Le traitement automatique des langues, la transcription automatique de la parole, la reconnaissance des personnes (faciale et vocale) et la reconnaissance du texte affiché sont indispensables pour mener à bien les travaux de recherche sur l'analyse et la compréhension des médias. Ces thématiques se trouvent ainsi au cœur de plusieurs des projets de recherche de l'Institut.

Le projet ANR Gender Equality Monitor (GEM) vise à décrire les différences de représentation existant entre les femmes et les hommes dans les médias. Ce projet pluridisciplinaire est coordonné par

l'INA et implique un consortium de sept partenaires : LIUM, LISN (anciennement LIMSI), LERASS, CARISM, Centre Max Weber et la société Deezer. Plusieurs outils développés dans le cadre de ce projet permettent d'automatiser le décompte du temps de parole (InaSpeechSegmenter) ou du temps d'apparition à l'image (InaFaceGender) des femmes et des hommes. D'autres approches semi-automatiques permettent d'accélérer la catégorisation des intervenant-e-s des journaux télévisés en se fondant sur une analyse OCR des incrustations de texte utilisées pour présenter les personnes. Ces stratégies permettent de traiter des volumes de données exhaustifs (supérieurs au million d'heures), limiter les biais d'échantillonnage, produire des descriptions diachroniques et contribuer à alimenter le débat citoyen. GEM est un projet à fort impact social : outre l'intérêt manifesté par le grand public pour les études associées à ce projet, les analyses réalisées sont intégrées depuis 2020 au rapport annuel du CSA sur la représentation des femmes à la télévision et à la radio, et ont également été utilisées pour contribuer au rapport de la députée Céline Calvez sur la place des femmes dans les médias en période de crise (Covid). Les travaux en cours visent désormais à mieux caractériser et détecter les interruptions, notamment pour mieux décrire les phénomènes de "mecterruption" ("manterrupting").

Le projet ANR ANTRACT « Analyse transdisciplinaire des Actualités filmées (1945-1969) », mené en partenariat avec le Centre d'histoire sociales de mondes contemporains (CHS), le LIUM, Eurecom et l'IHRIM, se consacre à l'analyse des images et des sons produits pendant près de vingt-cinq ans par les Actualités Françaises, société de presse filmée créée en 1945 grâce à des outils technologiques d'analyse des contenus audiovisuels et textuels : analyse de l'image et du son, transcription automatique de la parole et textométrie.

Issue du projet ANR OTMedia, la plateforme OTMedia+ a pour objectif de permettre l'analyse transmédia d'importants volumes de données hétérogènes provenant de sources audiovisuelles et textuelles diverses (radio, télévision, presse, et Twitter en particulier) au plus près possible du temps réel. Intégrant des résultats de systèmes de transcription automatique, la plateforme permet l'étude de différents phénomènes de propagation de l'information dans les médias.

La plateforme Okapi (pour « Open Knowledge-based Annotation and Publishing Interface ») est le résultat de plusieurs projets de recherche, et en particulier du projet ANR Campus AAR. Il s'agit d'une plateforme client-serveur intégrant des fonctionnalités de documentation, de recherche d'information et de publication hypermédia au sein d'un même environnement. Ce système est entièrement fondé sur les techniques et standards du web sémantique. Il peut notamment gérer des modèles de description définis par les utilisateur-rice·s ainsi que des portails web entièrement paramétrables.

Références

[1] Eléonore Alquier, Jean Carrive, and Steffen Lalande. Production documentaire et usages. L'automatisation dans les outils de consultation et de documentation de l'Institut national de l'audiovisuel (Ina). *Document Numérique*, 20(2-3) :115–136, 2017.

[2] Pierre-Alexandre Broux, Florent Desnous, Anthony Larcher, Simon Petitrenaud, Jean Carrive, and Sylvain Meignier. S4D : Speaker Diarization Toolkit in Python. In *Interspeech 2018*, Hyderabad, India, 2018.

[3] Pierre-Alexandre Broux, David Doukhan, Simon Petitrenaud, Sylvain Meignier, and Jean Carrive. Computer-assisted Speaker Diarization : How to Evaluate Human Corrections. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 2018.

[4] Pierre-Alexandre Broux, David Doukhan, Simon Petitrenaud, Sylvain Meignier, and Jean Carrive. Segmentation et Regroupement en Locuteurs : comment évaluer les corrections humaines. In *Journées d'Études sur la Parole (JEP)*, Aix-en-Provence, France, 2018.

[5] Jean Carrive. Using artificial intelligence to preserve audiovisual archives : New horizons, more questions. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1–2. ACM, 2019.

[6] David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. An

open-source speaker gender detection framework for monitoring gender equality. In *IEEE International Conference on Acoustic Speech and Signal Processing*, Calgary, Canada, 2018.

[7] David Doukhan, Eliott Lechapt, Marc Evrard, and Jean Carrive. Ina's mirex 2018 music and speech detection system. In *Music Information Retrieval Evaluation eXchange*, 2018.

[8] David Doukhan, Cécile Méadel, and Marlène Coulomb-Gully. En période de coronavirus, la parole d'autorité dans l'info télé reste largement masculine, 2020. La revue des médias.

[9] David Doukhan, Géraldine Poels, Zohra Rezgui, and Jean Carrive. Describing gender equality in french audiovisual streams with a deep learning approach. *Journal of European Television History and Culture*, 2018.

[10] David Doukhan, Zohra Rezgui, Géraldine Poels, and Jean Carrive. Estimer automatiquement les différences de représentation existant entre les femmes et les hommes dans les médias. In *journée DAHLIA : "Informatique et Humanités numériques : quelles problématiques pour quels domaines ?"*, Nantes, France, 2019.

[11] Anthony Larcher, Ambuj Mehrish, Marie Tahon, Sylvain Meignier, Jean Carrive, David Doukhan, Olivier Galibert, and Nicholas Evans. Speaker embeddings for diarization of broadcast data in the allies challenge. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5799–5803, 2021.

[12] Béatrice Mazoyer, Julia Cage, Céline Hudelot, and Marie-Luce Viaud. Real-time collection of reliable and representative tweets datasets related to news events. In *First International Workshop on Analysis of Broad Dynamic Topics over Social Media (BroDyn 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018)*, Grenoble, France, 2018.

[13] Béatrice Mazoyer, Nicolas Hervé, and Céline Hudelot. Réduire les biais dans la collecte de tweets. In *journée DAHLIA : "Informatique et Humanités numériques : quelles problématiques pour quels domaines ?"*, Nantes, France, 2019.

[14] Haolin Ren, Benjamin Renoust, Guy Melançon, Marie-Luce Viaud, and Shin'ichi Satoh. Exploring temporal communities in mass media archives. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, 2018.

[15] Haolin Ren, Benjamin Renoust, Marie-Luce Viaud, Guy Melançon, and Shin'ichi Satoh. Generating "visual clouds" from multiplex networks for tv news archive query visualization. In *CBMI*, 2018.

[16] Haolin Ren, Marie-Luce Viaud, and Guy Melançon. Mainmise sur les médias et suivi de communautés dans les graphes dynamiques. In *Extraction et Gestion des Connaissances*, 2018.

[17] Rémi Uro, Marc Evrard, Nicolas Hervé, and Béatrice Mazoyer. Création d'un corpus de tweets en français pour la détection automatique de position-nement (stance). In *journée DAHLIA : "Informatique et Humanités numériques : quelles problématiques pour quels domaines ?"*, Nantes, France, 2019.

[18] Rémi Uro, Marc Evrard, Nicolas Hervé, and Béatrice Mazoyer. The constitution of a French tweet corpus for automatic stance detection. In *International Conference on Statistical Language and Speech Processing*, Ljubljana, Slovenia, 2019.

[19] Marie-Luce Viaud, Agnès Saulnier, Nicolas Hervé, Benjamin Renoust, and Jérôme Thièvre. Otmedia : outils de fouille multimodales transmedia de l'actualité. *Médias et Humanités*, 2018.

AfIA
Association française
pour l'Intelligence Artificielle

IRIT/IRIS : *Information Retrieval & Information Synthesis*

IRIT UMR 5505 / IRIS
CNRS et Université de Toulouse
https://www.irit.fr/departement/gestion-des-donnees/iris/

Gilles HUBERT
gilles.hubert@irit.fr

Membres permanents impliqués

- Mohand BOUGHANEM (PR)
- Guillaume CABANAC (MCF HDR)
- Taoufiq DKAKI (MCF)
- Gilles HUBERT (MCF HDR)
- Lynda LECHANI-TAMINE (PR)
- José MORENO (MCF)
- Karen PINEL-SAUVAGNAT (MCF HDR)
- Yoann PITARCH (MCF)

Thématiques de l'équipe

Les activités de recherche de l'équipe IRIS (Information Retrieval and Information Synthesis) sont axées sur la conception de modèles de recherche d'information (par ex. fondés sur un apprentissage profond) et sur l'élaboration de méthodes d'exploration de données, d'agrégation d'information et de scientométrie. Elles s'inscrivent principalement dans le domaine de la *recherche d'information* (RI) avec de nombreuses connexions avec les domaines du *traitement automatique des Langues* (TAL) et de l'*intelligence artificielle* (IA).

Recherche d'information

Les sujets de recherche de l'équipe IRIS comprennent la recherche d'information, y compris la conception de modèles, la recherche collaborative et l'apprentissage pour la recherche d'information.

Modèles de recherche d'information : Le principal défi abordé dans le cadre de ce thème est la modélisation de la pertinence, qui a toujours été un défi central dans la recherche d'information. Ce sujet a été étudié sous différents angles, en fonction de facteurs d'impact spécifiques et dans une perspective d'estimation de la pertinence. Ces facteurs comprennent, entre autres, la multiplicité des dimensions de pertinence, la temporalité des documents, le contexte de recherche comme les signaux des médias sociaux. Les travaux menés autour de la question de la multidimensionnalité de la pertinence ont par exemple porté sur (1) la définition d'un opérateur d'agrégation multicritères basé sur l'intégrale de Choquet [9], (2) des représentations multifacettes des documents et leur comparaison originale sous forme de tournoi pour classer les documents répondant à un besoin utilisateur [7], (3) des représentations conceptuelles [11] afin de réduire l'écart sémantique dans la correspondance requête-document, ou encore (4) l'exploitation de signaux sociaux comme *a priori* pour réviser les modèles linguistiques [1]. Autour de ce thème, nous avons coordonné le projet ANR CAIR (2014-2018).

Recherche collaborative : La recherche collaborative est une forme de recherche dynamique impliquant un groupe d'utilisateurs engagés dans une tâche de recherche exploratoire complexe et partagée. La recherche collaborative englobe la recherche et les applications de l'interaction homme-machine, des sciences de l'information et, plus récemment, des domaines de la recherche d'information. [13] fournit une vue d'ensemble des différentes formes de soutien à la collaboration qui s'y rapportent, principalement basées sur des approches algorithmiques, y compris des approches axées sur l'utilisateur et des approches systémiques. Dans la perspective de la recherche d'information, les travaux menés ont abordé l'apprentissage dynamique des interactions utilisateurs-système et utilisateurs-utilisateurs passées pour prédire l'avenir en termes d'estimation de la copertinence [12]. Une analyse plus approfondie des différences de comportement des utilisateurs [15] et des pratiques de recherche au sein des plateformes de médias sociaux [16] nous a permis d'ouvrir des opportunités de recherche sur les systèmes de questions-réponses sociaux coopératifs [14]. Autour de ce thème, nous avons coordonné un projet de recherche pluridisciplinaire

(PEPS CNRS EXPAC 2014-2015), donné deux tutoriels lors de grandes conférences de RI (ECIR'16, ICTIR'17) et animé deux éditions d'ateliers internationaux (ECol'17, ECol'15).

Recherche d'information et apprentissage : Ce domaine de recherche porte sur l'utilisation d'approches d'apprentissage automatique pour résoudre des problèmes de recherche d'information de base comme la représentation de l'information (document, requête) et le classement des documents. L'équipe IRIS s'est concentrée sur un nouvel axe de recherche (depuis 2016) lié à la conception de modèles d'apprentissage de la représentation des documents et de leurs constituants pour faire face à la question bien connue du fossé sémantique qui sous-tend les tâches de recherche telles que le classement des documents et l'annotation sémantique. Les travaux menés ont porté notamment (1) sur la combinaison de la sémantique distributive (basée sur la prédiction du contexte) et de la sémantique relationnelle dans un cadre d'apprentissage hybride unifié [10] ainsi que (2) sur la construction conjointe de plongements de mots et d'entités en utilisant un texte d'ancrage existant dans plusieurs corpus, tel que Wikipedia [8]. Autour de ce thème, nous coordonnons le projet ANR CoST (2019-2022) portant spécifiquement sur les modèles de séquences pour la recherche interactive complexe, participons au projet ANR MEERQAT (2020-2023) et avons initié des collaborations de recherche avec les sociétés ATOS et RENAULT au travers de thèses de doctorat CIFRE.

Synthèse d'information

L'objectif principal visé dans cet axe de recherche est la conception de solutions efficaces et efficientes pour révéler des connaissances exploitables à partir de données complexes et volumineuses (graphiques, flux, semi-structurés, *etc.*). Nous étudions en particulier les axes de recherche portant sur la détection de points de vue et d'expertises ainsi que la scientométrie.

Détection d'opinion et de point de vue, détection d'expertise : La détection des opinions et des points de vue vise à identifier les points de vue ou les opinions exprimés dans les textes ou à déterminer l'adhésion des auteurs à certains points de vue (par ex. les utilisateurs des médias sociaux). Une première série de contributions a consisté (1) à définir des modèles thématiques probabilistes pour la découverte de points de vue et d'opinions dans les textes de réseaux sociaux [17, 18], (2) à proposer un modèle de propagation des points de vue basé sur différentes proximités définies entre les nœuds d'un réseau social [5]. La détection de l'expertise consiste à déterminer les domaines et les niveaux d'expertise des personnes à partir de leur production en termes de documents, de messages échangés, de réponses aux questions. À partir d'une représentation graphique des réseaux sociaux et des plateformes collaboratives, les travaux menés ont permis de définir un modèle d'autorité fondé sur le renforcement mutuel et l'influence cumulative dans un graphique hétérogène. Les travaux autour de cette thématique ont bénéficié d'une participation à divers projets (par ex. FUI ACOVAS, ANR LISTIC) et d'une collaboration initiée avec le CEA.

Scientométrie : La scientométrie est un domaine de recherche interdisciplinaire se référant à l'étude quantitative de la science et de l'innovation par une combinaison de méthodes algorithmiques et statistiques. La recherche dans ce domaine exploite les matériaux produits par les chercheurs, c'est-à-dire les 1,3 millions de publications publiées chaque année qui constituent une ressource clé à explorer car elles transmettent des informations riches et hétérogènes : métadonnées sur les auteurs et les résultats de la recherche, texte intégral, réseaux de scientifiques/affiliations/pays et citations. Les questions que nous abordons concernent la nature même de l'information bibliométrique : hétérogénéité, accessibilité, temporalité et multidimensionnalité des données. Notre recherche s'efforce de concevoir les schémas de traitement de données appropriés pour extraire les textes scientifiques et les réseaux latents (concernant le lexique, les références, les auteurs, les affiliations, *etc.*) afin de tester les théories et hypothèses issues des sciences sociales, découvrir de nouvelles connaissances et révéler les forces motrices de la créativité scientifique et de la diffusion du savoir scientifique. Les contributions de l'équipe dans ce domaine concernent par exemple les processus de création de textes scien-

tifiques [2, 6, 3] ou la mise en place et l'évolution des réseaux de collaboration [4].

Nous collaborons avec des scientifiques de différents domaines de recherche, tels que la sociologie de la science (ANR RésoCit), l'agroéconomie (projets Labex SMS HERA et INRA BILAG), la géographie (Labex SMS NetScience), la pharmacologie (projet Pharmakon). Nous nous efforçons de promouvoir et de renforcer les liens entre la scientométrie et la recherche d'information via l'implication dans les cinq éditions de l'atelier BIR (Bibliometric-enhanced Information Retrieval, tenu entre 2016 et 2019 à la conférence ECIR) et BIRNDL (Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries tenu en 2016 à la conférence JCDL), ainsi que la coédition de deux numéros spéciaux de l'International Journal on Digital Libraries and Scientometrics.

Références

[1] Ismail Badache and Mohand Boughanem. Emotional social signals for search ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*. ACM Press, 2017.

[2] Guillaume Cabanac. Extracting and quantifying eponyms in full-text articles. *Scientometrics*, 98(3) :1631–1645, 2014.

[3] Guillaume Cabanac, Gilles Hubert, and James Hartley. Solo versus collaborative writing : Discrepancies in the use of tables and graphs in academic articles. *Journal of the Association for Information Science and Technology*, 65(4) :812–820, 2014.

[4] Guillaume Cabanac, Gilles Hubert, and Béatrice Milard. Academic careers in computer science : continuance and transience of lifetime co-authorships. *Scientometrics*, 102(1) :135–150, 2015.

[5] Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. Stance classification through proximity-based community detection. In *Proceedings of the 29th on Hypertext and Social Media - HT '18*. ACM Press, 2018.

[6] James Hartley and Guillaume Cabanac. Do men and women differ in their use of tables and graphs in academic publications ? *Scientometrics*, 98(2) :1161–1172, 2014.

[7] Gilles Hubert, Yoann Pitarch, Karen Pinel-Sauvagnat, Ronan Tournier, and Léa Laporte. Tournarank : When retrieval becomes document competition. *Information Processing & Management*, 54(2) :252–272, 2018.

[8] Jose G. Moreno, Romaric Besançon, Romain Beaumont, Eva D'hondt, Anne-Laure Ligozat, Sophie Rosset, Xavier Tannier, and Brigitte Grau. Combining word and entity embeddings for entity linking. In *Proceedings of European Semantic Web Conference ESWC 2017 : The Semantic Web*, Lecture Notes in Computer Science, page 337–352. Springer, 2017.

[9] Bilel Moulahi, Lynda Tamine, and Sadok Ben Yahia. iaggregator : Multidimensional relevance aggregation based on a fuzzy operator. *Journal of the Association for Information Science and Technology*, 65(10) :2062–2083, 2014.

[10] Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, and Nathalie Souf. A tri-partite neural document language model for semantic information retrieval. In *Proceedings of European Semantic Web Conference ESWC 2018 : The Semantic Web*, Lecture Notes in Computer Science, page 445–461. Springer, 2018.

[11] Lynda Said Lhadj, Mohand Boughanem, and Karima Amrouche. Enhancing information retrieval through concept-based language modeling and semantic smoothing. *Journal of the Association for Information Science and Technology*, 67(12) :2909–2927, 2015.

[12] Laure Soulier, Chirag Shah, and Lynda Tamine. User-driven system-mediated collaborative information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 485–494, New York, NY, USA, 2014. ACM.

[13] Laure Soulier and Lynda Tamine. On the collaboration support in information retrieval. *ACM Computing Surveys*, 50(4) :1–34, 2017.

[14] Laure Soulier, Lynda Tamine, and Gia-Hung Nguyen. Answering twitter questions. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*. ACM Press, 2016.

[15] Lynda Tamine and Laure Soulier. Understanding the impact of the role factor in collaborative information retrieval. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*. ACM Press, 2015.

[16] Lynda Tamine, Laure Soulier, Lamjed Ben Jabeur, Frederic Amblard, Chihab Hanachi, Gilles Hubert, and Camille Roth. Social media-based collaborative information access. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media - HT '16*. ACM Press, 2016.

[17] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. Vodum : A topic model unifying viewpoint, topic and opinion discovery. In *Proceedings of the European Conference on Information Retrieval ECIR 2016 : Advances in Information Retrieval*, Lecture Notes in Computer Science, page 533–545. Springer, 2016.

[18] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. Users are known by the company they keep. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*. ACM Press, 2017.

AfIA
Association française
pour l'Intelligence Artificielle

L3i/IC : Images et Contenus

L3i EA 2118/IC
La Rochelle Université
https://l3i.univ-larochelle.fr/

Antoine DOUCET
antoine.doucet@univ-lr.fr

Fort d'un historique et d'une visibilité importante dans le domaine de l'analyse de document, l'équipe de recherche « Images et Contenus » (IC) du laboratoire L3i a développé des travaux autour des technologies du langage humain afin de compléter sa chaîne de traitement, qui va désormais de la numérisation au traitement des langues (TAL) et à la recherche d'information (RI).

Ces travaux s'appuient sur un contexte partenarial important, avec, en cours au moment d'écrire ces lignes : 2 projets H2020 (*NewsEye - a digital investigator for historical newspapers*, en coordination de 2018 à 2022 [9] et *Embeddia : crossingual embeddings for less represented languages in European news media*, en tant que partenaire de 2019 à 2022) et 5 projets région Nouvelle-Aquitaine dont 3 en coordination. Une activité de transfert est assurée par l'intermédiaire de 4 thèses CIFRE en cours de déroulement, d'un laboratoire commun (ANR LabCom IDEAS 2019-2022), et plusieurs contrats de collaboration.

Thématiques de recherche

Nos travaux sont à l'intersection du TAL, de la RI, de la fouille de données textuelles et de l'intelligence artificielle (IA). Nous étudions le texte sous toutes ses formes (écrit/transcrit, numérisé/natif, soutenu/parlé, etc.), avec pour ligne directrice d'opter pour des méthodes qui soient le plus générique possible, c'est à dire qui fonctionnent de la même façon pour toutes les langues. Cela requière de s'appuyer principalement sur des éléments statistiques plutôt que sur des ressources linguistiques spécifiques aux langues telles que des dictionnaires ou des outils d'analyse linguistique automatique (syntaxique, morphologique, etc.), dont la qualité est variable en fonction des langues et qui n'existent d'ailleurs pas pour toutes.

L'intérêt de ce choix est que tout type de texte peut être analysé, ce qui est particulièrement utile dans les nombreux cas où les ressources sont insuffisantes voire inexistantes : de nombreuses langues, dites « peu dotées », ne disposent en effet pas d'outils d'analyse linguistique ou de ressources dédiées de qualité permettant d'entraîner des modèles d'IA. C'est également le cas de certaines formes de langages comme celles utilisées sur les réseaux sociaux ou dans les SMS, avec de nombreux raccourcis, abréviations, hashtags et autres émojis.

Les approches génériques permettent aussi d'analyser des textes dit « bruités », comme ceux qui sont issus d'une reconnaissance automatique de la parole ou d'un processus de numérisation de documents : le texte est alors imparfaitement extrait, par exemple à cause d'une tâche d'encre, d'une pliure, d'un coup de tampon, ou tout simplement de la dégradation naturelle du support au fil du temps.

Approches contrastives du texte

S'appuyant historiquement sur notre expertise des approches indépendantes des langues, faisant autant que possible abstraction de bases de connaissances externes, nous avons dans plusieurs contextes travaillé à l'intersection du TAL, de la fouille de données textuelles et de la RI.

Nos travaux contrastifs visant à comparer de façon non supervisée un ou des documments à un flux de documents passés ont trouvé des applications dans le résumé multidocument mutilingue [10], la veille multilingue [15] et la génération d'humour par estimation de différents facteurs déclenchant comme la surprise [23]. Les travaux autour de la veille ont plus récemment été étendus vers des approches supervisées [18] et des approches plus génériques à grain fin [4], comprenant la détection de tendances et la détéction de signaux faibles [17].

Crosslingual embeddings

Nos approches indépendantes des langues ont par ailleurs trouvé un terreau favorable dans le développement des ressources statistiques que sont les plongements de mots (word embeddings) et leurs déclinaisons. Nous avons développé des approches pour tirer profit des word embeddings de façon cross-lingue, c'est-à-dire, par exemple, en utilisant des ressources en français pour améliorer l'analyse de textes en croate [7]. Ce type d'approche est particulièrement utile pour les langues peu dotées en ressources linguistiques.

Documents numérisés

Une grande partie de nos travaux a porté sur l'analyse de documents anciens numérisés, trouvant des utilisations notamment dans le cadre des humanités numériques. Nous avons réalisé des travaux couvrant la totalité de la chaîne de traitement sémantique des documents [14].

Post-correction d'OCR : Tout d'abord, afin d'améliorer la qualité des contenus textuels extraits, nous avons proposé des méthodes d'analyse détaillée des erreurs d'OCR et des approches pour la correction de ces erreurs [20, 21]. Nous avons également rendu publics des jeux de données d'évaluation dans le cadre de 2 compétitions internationales en 2017 et 2019, couvrant des documents imprimés et manuscrits, dans un total de 10 langues [8, 22].

Robustesse au bruit : En nous appuyant sur des méthodes de synthèse de la dégradation de documents, nous avons par ailleurs étudié en détail l'impact de cette dégradation sur l'analyse sémantique des contenus textuels résultants (incluant notamment une reconnaissance optique de caractères imparfaite), détaillant notamment nos travaux sur les cas de la reconnaissance [13] et de la désambiguïsation [16] des entités nommées, de la polarité d'opinion relative à chaque mention d'une entité nommée, et de la détection d'évènements [19]. Ces travaux ont permis le développement de méthodes adaptées, fonctionnant de façon indépendante des langues. Leur efficacité a notamment été démontrée lors de la compétition d'évaluation comparative CLEF HIPE 2020 visant à la reconnaissance et à la désambiguïsation d'entités nommées dans des corpus de presse ancienne en anglais, français et allemand, où nos approches sont arrivées en tête de 50 évaluations sur 52 [5, 6].

Documents administratifs : D'autres travaux sur les documents numérisés concernent l'analyse de documents administratifs récents pour les classer et en extraire de l'information. La classification de documents s'appuie sur des approches multimodales mêlant des contenus textuels et visuels. Une fois ces documents classés, des approches d'extraction d'information adaptée et personnalisées sont proposées, en réemployant les approches issues des documents patrimoniaux (concept d'entités nommées pour des documents administratifs) d'une part [11], et de nouvelles sont en cours de développement en s'appuyant sur la structure des patterns (textuels ou graphiques) dans les documents [12].

Des images de document vers leurs contenus textuels

Comme dans le cas des documents administratifs, le contexte et les travaux de l'équipe l'ont conduite vers une combinaison des approches combinant l'analyse des images de documents et leurs contenus textuels, créant un continuum entre l'analyse d'image et celle des contenus textuels.

Détection de fraude : Sur le sujet de la détection de fraude, nous avons en efffet développé des méthodes s'appuyant sur les indices textuels pour le cas d'utilisation des tickets de caisse, visant à interpréter et modéliser leur contenu textuel et à estimer la cohérence des informations qu'ils contiennent [1]. Des travaux récents visent à la combinaison multimodale de ces approches par une approche hybride de hachage basé sur le contenu (authentification du modèle de document) et de vérification de la cohérence des contenus (authentification du texte) et la création du tout premier corpus public de documents fraudés [2].

Recherche d'information cross-modale unifiée : Nous avons enfin développé des travaux multimodaux, incluant en particulier un système de RI multimodale qui est au niveau de l'état de l'art avec pour particularité qu'il fonctionne de façon identique, quelles que soient les modalités des questions

et des réponses (texte, image, ou les deux), là où l'existant est spécialisé selon les modalités d'entrée et de sortie [3].

Références

[1] Chloé Artaud, Antoine Doucet, Vincent Poulain D
apos ;andecy, and Jean-Marc Ogier. Automatic Matching and Expansion of Abbreviated Phrases without Context. In *19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing2018)*, Hanoi, Vietnam, March 2018.

[2] Chloé Artaud, Nicolas Sidère, Antoine Doucet, Jean-Marc Ogier, and Vincent Poulain D
apos ;andecy. Find it ! Fraud Detection Contest Report. In *24th International Conference on Pattern Recognition (ICPR 2018)*, pages pp. 13–18, Beijing, China, August 2018.

[3] Viviana Beltrán, Juan C. Caicedo, Nicholas Journet, Mickaël Coustaty, François Lecellier, and Antoine Doucet. Deep multimodal learning for cross-modal retrieval : One model for all tasks. *Pattern Recognition Letters*, 146 :38–45, 2021.

[4] Imen Bizid, Nibal Nayef, Patrice Boursier, and Antoine Doucet. Detecting prominent microblog users over crisis events phases. *Information Systems*, 78 :173 – 188, November 2018.

[5] Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)*, pages 431–441, Online, November 2020. Association for Computational Linguistics.

[6] Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José G. Moreno, Nicolas Sidère, and Antoine Doucet. Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696 of *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, pages 1–17, Thessaloniki, Greece, September 2020. CEUR-WS Working Notes.

[7] Luis Adrián Cabrera-Diego, Jose G. Moreno, and Antoine Doucet. Simple ways to improve NER in every language using markup. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), Ljubljana, Slovenia, April 12, 2021 (online)*, volume 2829 of *CEUR Workshop Proceedings*, pages 17–31, 2021.

[8] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. ICDAR2017 Competition on Post-OCR Text Correction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 1423–1428, Kyoto, France, November 2017. IEEE.

[9] Antoine Doucet, Martin Gasteiner, Mark Granroth-Wilding, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, Guenter Muehlberger, Eva Pfanzelter, Marie-Eve Therenty, Hannu Toivonen, and Mikko Tolonen. NewsEye : A digital investigator for historical newspapers. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020*, Ottawa, Canada, July 2020.

[10] Oskar Gross, Antoine Doucet, and Hannu Toivonen. Language-Independent Multi-Document Text Summarization with Document-Specific Word Associations. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016*, page 8, Pisa, Italy, 2016. ACM.

[11] Ahmed Hamdi, Elodie Carel, Aurélie Joseph, Mickaël Coustaty, and Antoine Doucet. Information extraction from invoices. In *2021 International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021*.

[12] Ahmed Hamdi, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, Antoine Doucet, and Jean-Marc Ogier. Feature Selection

for Document Flow Segmentation. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 245–250, Vienna, Austria, April 2018. IEEE.

[13] Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. An Analysis of the Performance of Named Entity Recognition over OCRed Documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, volume 24, pages 333–334, Champaign, United States, June 2019. IEEE.

[14] Axel Jean-Caurant and Antoine Doucet. Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform. In *JCDL '20 : The ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 531–532, Virtual Event, China, August 2020. ACM.

[15] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. Multilingual Event Extraction for Epidemic Detection. *Artificial Intelligence in Medicine*, 65(2) :131–143, October 2015.

[16] Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, and Antoine Doucet. Impact of OCR Quality on Named Entity Linking. In *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia, November 2019.

[17] Julien Maitre, Michel Ménard, Guillaume Chiron, Alain Bouju, and Nicolas Sidère. A meaningful information extraction system for interactive analysis of documents. In *International Conference on Document Analysis and Recognition (ICDAR 2019)*, 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 92–99, Sydney, Australia, September 2019.

[18] Stephen Mutuvi, Emanuela Boros, Antoine Doucet, Gaël Lejeune, Adam Jatowt, and Moses Odeo. Multilingual Epidemiological Text Classification : A Comparative Study. In *COLING, International Conference on Computational Linguistics*, pages 6172–6183, Barcelone, Spain, December 2020.

[19] Nhu Khoa Nguyen, Emanuela Boros, Gaël Lejeune, and Antoine Doucet. Impact Analysis of Document Digitization on Event Extraction. In *4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020) co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2020)*, volume 2735 of *Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020)*, pages 17–28, Virtual, Italy, November 2020.

[20] Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickaël Coustaty, Nhu-Van Nguyen, and Antoine Doucet. Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 29–38, Champaign, France, June 2019. IEEE.

[21] Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickaël Coustaty, and Antoine Doucet. Neural Machine Translation with BERT for Post-OCR Error Detection and Correction. In *JCDL '20 : The ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 333–336, Virtual Event, China, August 2020. ACM.

[22] Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. ICDAR 2019 Competition on Post-OCR Text Correction. In *15th International Conference on Document Analysis and Recognition*, pages 1588–1593, Sydney, Australia, September 2019.

[23] Alessandro Valitutti, Antoine Doucet, Jukka M. Toivanen, and Hannu Toivonen. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5) :727–749, 2016.

AfIA
Association française
pour l'Intelligence Artificielle

LSCP/LAAC : Acquisition du Langage à travers Différentes Cultures

LSCP/LAAC
Département d'études cognitives, ENS, EHESS, CNRS, PSL University
https://lscp.dec.ens.fr/fr/recherche/equipes-du-lscp/lacquisition-du-langage-travers-differentes-cultures

Alejandrina CRISTIA
alecristia@gmail.com

Marvin LAVECHIN
marvinlavechin@gmail.com

Membres de l'équipe

- Alejandrina CRISTIA (DR)
- Lucas GAUTHERON (manageur de données)
- Cécile ISSARD (post-doctorante)
- Marvin LAVECHIN (doctorant)
- Georgia LOUKATOU (post-doctorante)
- Nicolas ROCAHT (analyste de données)
- Camila SCAFF (post-doctorante)
- Valentin THOUZEAU (post-doctorant)
- Catherine URBAN (manageur de projet)

Thématique générale de l'équipe

Notre objectif est de mettre en lumière les mécanismes et les processus impliqués dans l'acquisition précoce des langues dans une variété de cultures et de communautés linguistiques. À cette fin, nous utilisons une approche interdisciplinaire (allant de la modélisation informatique aux expériences en laboratoire et aux analyses de données avancées) dans le cadre d'une science ouverte, collaborative et engagée publiquement. Pour plus d'informations sur notre démarche scientifique, voir la déclaration de mission, accessible sur notre site.

Démarche scientifique

Nous travaillons sur l'acquisition précoce des langues. Les expériences acquises au cours des premières années de la vie sont cruciales pour le développement des compétences cognitives, et notamment de la/des langue(s) maternelle(s) : nos capacités de perception et de production s'adaptent aux langues humaines spécifiques auxquelles nous sommes exposés au cours de cette période. Nous les assimilons avec une facilité encore inégalée par les adultes, les animaux ou encore les apprenants artificiels. C'est d'autant plus étonnant que les expériences des jeunes enfants peuvent varier considérablement, au moins sur deux plans. Premièrement, les caractéristiques typologiques des différentes langues sont très variables. Par exemple, certaines langues n'utilisent que 10 à 15 sons différents pour coder tous leurs mots, alors que d'autres utilisent dix fois plus de sons. Nous constatons également de grandes différences dans d'autres niveaux d'organisation linguistique, y compris par exemple le nombre de formes différentes qu'un mot peut avoir : en anglais, un verbe peut avoir une poignée de formes (go, went, gone, etc.), alors qu'en chintang, il peut y avoir mille formes différentes. Deuxièmement, certains travaux suggèrent que la quantité et la qualité du discours adressé aux enfants varient considérablement d'une population à l'autre.

Nous nous intéressons à la manière dont les enfants en viennent à apprendre leur(s) langue(s) maternelle(s), aux types d'expériences dont ils ont besoin pour le faire, et aux mécanismes d'apprentissage sous-jacents, ayant en vue cette variabilité linguistique et de population. À cette fin, nous utilisons tous les outils disponibles : expériences en laboratoire et modélisation informatique pour vérifier les mécanismes d'apprentissage spécifiques, analyses des enregistrements d'une journée et d'autres corpus pour décrire les apports et les résultats, analyses de cohortes pour étudier les variations individuelles et collectives, méta-analyses pour synthétiser les données, etc. Par conséquent, notre travail se situe à l'intersection de plusieurs disciplines : la psychologie développementale, la linguistique, l'anthropologie, les sciences du comportement et l'intelligence artificielle.

Concernant les problématiques interdisciplinaires que nous abordons, trois thèmes ayant une forte composante en intelligence artificielle sont actuellement étudiés par notre équipe :

L'étude de l'acquisition du langage à l'ère du big data : Une des méthodes d'observation utilisées par les linguistes de terrain consiste à mettre un microphone directement dans la veste d'un enfant. Ces enregistrements audios peuvent durer plusieurs jours et être collectés simultanément sur plusieurs enfants et à travers différentes cultures. Cet outil, bien que porteur de nombreuses promesses (facilité d'acquisition, invasivité minimale, etc.) amène aussi de nombreux challenges à la frontière de la linguistique et du traitement du signal. Etant donné de tels enregistrements, il nous faut extraire, de façon automatique, diverses informations d'intérêt comme : Qui parle quand ? Qui parle à qui ? Quelle est la quantité de mots prononcés ? Quelle est la complexité des vocalisations produites par l'enfant ? Notre équipe est leader dans le développement de routines d'analyse open source en ce domaine [1].

Les propriétés d'apprentissage en fonction des langues et cultures : Les études précédentes sur l'apprentissage des langues se sont principalement concentrées sur l'anglais, une langue assez unique du point de vue typologique, et sur la parole adressée à l'enfant chez les anglophones, qui est différente de la parole adressée aux adultes de façon très marquée (le "parentais"). A quel point est-ce que les différences entre langues et entre registres impactent l'apprentissage linguistique ? Pour répondre à cette question, nous créons des modèles d'apprentissage non-supervisé, dont la plupart utilisent des stratégies qui ont été avérés chez des enfants. Nous n'utilisons pas seulement un modèle mais plusieurs modèles alternatifs car il est possible qu'aucun ne représente exactement ce que l'enfant fait, mais ensemble ils peuvent montrer ce que des appreneurs peuvent extraire de cette expérience. L'objectif de ces travaux est ainsi d'etudier l'impact des différences linguistiques [2] et de registre [3].

Ingénierie inverse du langage : Comment les modèles d'intelligence artificielle peuvent-ils nous éclairer sur les mecanismes en jeu dans l'acquisition du langage ? En entraînant des modèles d'apprentissage de façon non-supervisée, directement à partir des données collectées chez l'enfant. Nous étudions si ces systèmes in silico sont de plausibles modèles de l'acquisition du langage telle qu'elle pourrait se jouer chez les enfants. Si oui, peut-on utiliser les prédictions de ces modèles pour guider les théories de l'acquisition du langage ? Si non, comment modifier ces modèles afin d'améliorer leur plausibilité biologique et cognitive ?

Références

[1] M Lavechin, R Bousbib, H Bredin, E Dupoux, and A Cristia. An open-source voice type classifier for child-centered daylong recordings. *Interspeech*, 2020.

[2] G R Loukatou, S Moran, D Blasi, S Stoll, and A Cristia. Is word segmentation child's play in all languages ? In *ACL*, pages 3931–3937, 2019.

[3] B Ludusan, R Mazuka, M Bernard, A Cristia, and E Dupoux. The role of prosody and speech register in word segmentation : A computational modelling perspective. In *ACL*, pages 178–183, 2017.

AfIA
Association française
pour l'Intelligence Artificielle

LabHC : Laboratoire Hubert Curien

Laboratoire Hubert Curien UMR 5516
CNRS et Université de Lyon - UJM Saint-Etienne
https://laboratoirehubertcurien.univ-st-etienne.fr

François JACQUENET
Francois.Jacquenet@univ-st-etienne.fr

Membres impliqués

- Marc BERNARD (MCF)
- Mathias GERY (MCF)
- Christophe GRAVIER (PR)
- Amaury HABRARD (PR)
- François JACQUENET (PR)
- Charlotte LACLAU (MCF)
- Christine LARGERON (PR)
- Pierre MARET (PR)
- Fabrice MUHLENBACH (MCF)

Contexte

Le laboratoire Hubert Curien est un laboratoire pluridisciplinaire qui développe une activité de recherche en informatique au sein de l'équipe Data Intelligence. Cette équipe est spécialisée d'une part dans le domaine du machine learning, dont l'objectif est d'apprendre automatiquement des modèles par optimisation mathématique à partir d'exemples, et d'autre part dans le domaine de l'analyse de données visant à extraire de la connaissance pertinente à partir de grands volumes d'informations, potentiellement complexes. À partir d'une recherche située à la frontière de l'informatique, des mathématiques appliquées et des statistiques, l'équipe a su développer une activité reconnue dans les domaines du *representation learning*, *metric learning*, *transfer learning*, *optimal transport*, *statistical learning theory* et *complex data analysis*.

Plusieurs membres de l'équipe Data Intelligence, cités ci-dessus, s'intéressent entre autre au traitement automatique des langues. Nous travaillons d'une part autour d'approches numériques (statistiques, neuronales) et d'autre part autour d'approches symboliques (fondées sur la logique du premier ordre).

Travaux développés

Nous nous intéressons à la prise en compte de la sémantique dans l'analyse et la génération de données textuelles. Dans un premier temps, nos travaux ont eu pour objectif de développer de nouvelles mesures efficaces de similarité sémantique entre textes [24]. Cela nous a conduit à nous intéresser aux approches à base d'apprentissage profond et notamment à la construction de nouveaux *word embeddings* à partir de dictionnaires en ligne [26] ainsi qu'à leur représentation par vecteurs de bits [27]. Les approches Few-Shot learning sont également au coeur de nos travaux actuels dans ce contexte [13, 12]. Nous avons également développé des techniques à base d'apprentissage profond pour la production automatisée de résumés de textes [28, 19, 17] ou la génération automatique de questions [15].

Nous intégrons des techniques du web sémantique dans des travaux en *question answering* [10, 23, 8, 9], nous permettant d'obtenir des performances aussi bonnes que les meilleures approches, tout en améliorant les critères de rapidité, multilinguisme, multigraphes, passage à l'échelle [11]. Nous travaillons également sur l'analyse de textes guidée par des ressources sémantiques pour augmenter des graphes de connaissances.

Dans le domaine de la recherche d'information, de l'extraction d'information et des systèmes de recommandation, nous cherchons à prendre en compte divers types d'informations (structure des documents textuels [7], images contenues dans les textes [22], modèles de langages personnalisés [2], combinaison de données de type texte, de variables descriptives et d'informations issues de traitement du signal [21]) à développer des systèmes plus efficaces [25, 4], permettant de découvrir des relations avec un rappel élevé [14] ou de favoriser des recommandations musicales cherchant à étendre l'univers culturel de l'utilisateur.

Nous avons développé des techniques de *text mining* dans le cadre de la veille technologique et économique afin de découvrir de l'information inattendue dans des corpus de textes [18], d'enrichir en méta-données, par une approche sémantique, les mots clés décrivant un article scientifique afin de favoriser l'accès à des documents intéressants dans une bibliothèque numérique [1], de classer des documents [20] ou d'identifier automatiquement des auteurs d'articles [16].

Parallèlement à ces approches numériques, nous avons mené un certain nombre de travaux basés sur des approches symboliques (logique du premier ordre) pour la prise en compte de la sémantique dans le cadre de l'apprentissage de modèles de langages. Nous avons ainsi développé d'une part des techniques fondées sur un modèle élève/professeur où l'élève apprend un langage en produisant des phrases qui sont corrigées ensuite par le professeur [3] et d'autre part des techniques de programmation logique inductive pour apprendre la sémantique des mots d'un langage à partir de paires d'images et de textes décrivant ces images [5, 6].

Références

[1] Hussein T. Al-Natsheh, Lucie Martinet, Fabrice Muhlenbach, et al. Metadata enrichment of multi-disciplinary digital library : A semantic-based approach. In *Proc. of TPDL*, pages 32–43, 2018.

[2] Nawal Ould Amer, Philippe Mulhem, and Mathias Géry. Personalized parsimonious language models for user modeling in social bookmaking systems. In *Proc. of ECIR*, pages 582–588, 2017.

[3] Dana Angluin and Leonor Becerra-Bonache. A model of language learning with semantics and meaning-preserving corrections. *Artificial Intelligence*, 242 :23–51, 2017.

[4] Georgios Balikas, Charlotte Laclau, Ievgen Redko, and Massih-Reza Amini. Cross-lingual document retrieval using regularized wasserstein distance. In *Proc. of ECIR*, pages 398–410, 2018.

[5] Leonor Becerra-Bonache, Hendrik Blockeel, María Galván, and François Jacquenet. A first-order-logic based model for grounded language learning. In *Proc. of IDA*, LNCS 9385, pages 49–60, 2015.

[6] Leonor Becerra-Bonache, Hendrik Blockeel, María Galván, and François Jacquenet. Learning language models from images with regll. In *Proc. of ECML/PKDD*, LNCS 9853, pages 55–58, 2016.

[7] Michel Beigbeder, Mathias Géry, and Christine Largeron. Using proximity and tag weights for focused retrieval in structured documents. *Knowledge and Information Systems*, 44(1) :51–76, 2015.

[8] Dennis Diefenbach, Andreas Both, Kamal Singh, and Pierre Maret. Towards a question answering system over the semantic web. *Semantic Web*, 11(3) :421–439, 2020.

[9] Dennis Diefenbach, José M. Giménez-García, Andreas Both, Kamal Singh, and Pierre Maret. QAnswer KG : designing a portable question answering system over RDF data. In *Proc. of ESWC*, pages 429–445, 2020.

[10] Dennis Diefenbach, Pedro Henrique Migliatti, Omar Qawasmeh, Vincent Lully, Kamal Singh, and Pierre Maret. Qanswer : A question answering prototype bridging the gap between a considerable part of the LOD cloud and end-users. In *Proc. of WWW*, pages 3507–3510, 2019.

[11] Dennis Diefenbach, Kamal Singh, and Pierre Maret. On the scalability of the QA system wdaqua-core1. In *Proc. of the SemWebEval Challenge at ESWC*, pages 76–81, 2018.

[12] Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. A neural few-shot text classification reality check. In *Proceedings of EACL*, pages 935–943, 2021.

[13] Thomas Dopierre, Christophe Gravier, Julien Subercaze, and Wilfried Logerais. Few-shot pseudo-labeling for intent detection. In *Proc. of COLING*, pages 4993–5003, 2020.

[14] Hady ElSahar, Christophe Gravier, and Frédérique Laforest. High recall open IE for relation discovery. In *Proc. of IJCNLP*, pages 228–233, 2017.

[15] Hady ElSahar, Christophe Gravier, and Frédérique Laforest. Zero-shot question generation

from knowledge graphs for unseen predicates and entity types. In *Proc. of NAACL-HLT*, pages 218–228, 2018.

[16] Jordan Fréry, Christine Largeron, and Mihaela Juganaru-Mathieu. Author identification by automatic learning. In *Proc. of ICDAR*, pages 181–185, 2015.

[17] François Jacquenet, Marc Bernard, and Christine Largeron. Meeting summarization, A challenge for deep learning. In *Proc. of IWANN*, LNCS 11506, pages 644–655, 2019.

[18] François Jacquenet and Christine Largeron. Discovering unexpected documents in corpora. *Knowledge-Based Systems*, 22(6) :421–429, 2009.

[19] Lucie-Aimée Kaffee, Hady ElSahar, Pavlos Vougiouklis, Christophe Gravier, et al. Learning to generate wikipedia summaries for underserved languages from wikidata. In *Proc. of NAACL-HLT*, pages 640–645, 2018.

[20] Christine Largeron, Christophe Moulin, and Mathias Géry. Entropy based feature selection for text categorization. In *Proc. of SAC*, pages 924–928, 2011.

[21] Pierre-René Lhérisson, Fabrice Muhlenbach, and Pierre Maret. Fair recommendations through diversity promotion. In *Proc. of ADMA*, pages 89–103, 2017.

[22] Christophe Moulin, Christine Largeron, and Mathias Géry. Impact of visual information on text and content based image retrieval. In *Proc. of SSPR&SPR*, pages 159–169, 2010.

[23] Koichi Shimoda, Dennis Diefenbach, Kamal Singh, Akihito Taya, Yoshito Tobe, and Pierre Maret. RW-QAnswer : an assisting system for intelligent environments using semantic technology. *Journal on Reliable Intelligent Environments*, 6(4) :215–231, 2020.

[24] Julien Subercaze, Christophe Gravier, and Frédérique Laforest. On metric embedding for boosting semantic similarity computations. In *Proc. of ACL*, pages 8–14. The Association for Computer Linguistics, 2015.

[25] Julien Subercaze, Christophe Gravier, and Frédérique Laforest. Real-time, scalable, content-based twitter users recommendation. *Web Intelligence*, 14(1) :17–29, 2016.

[26] Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proc. of EMNLP*, pages 254–263, 2017.

[27] Julien Tissier, Christophe Gravier, and Amaury Habrard. Near-lossless binarization of word embeddings. In *Proc. of AAAI*, pages 7104–7111, 2019.

[28] Pavlos Vougiouklis, Hady ElSahar, Lucie-Aimée Kaffee, Christophe Gravier, Frédérique Laforest, Jonathon S. Hare, and Elena Simperl. Neural wikipedian : Generating textual summaries from knowledge base triples. *Journal of Web Semantics*, 52-53 :1–15, 2018.

AfIA
Association française
pour l'Intelligence Artificielle

CEA/LASTI : Laboratoire Analyse Sémantique Texte Image

CEA LIST / LASTI
http://www.kalisteo.fr

Bertrand DELEZOIDE
bertrand.delezoide@cea.fr

Membres permanents

- Bertrand DELEZOIDE, responsable
- Romaric BESANÇON
- Gaël DE CHALENDAR
- Anne-Laure DAQUO
- Olivier FERRET
- Benjamin LABBÉ
- Meriama LAIB
- Hervé LE BORGNE
- Olivier MESNARD
- Adrian POPESCU
- Nasredine SEMMAR
- Julien TOURILLE

Thématique du laboratoire

Au sein de l'institut LIST du CEA, le Laboratoire d'Analyse Sémantique des Textes et des Images (LASTI) est une équipe de 25 personnes (chercheurs, ingénieurs, doctorants) menant des travaux sur les technologies de description et de compréhension des contenus multimédia (image, texte, parole) et multilingues, en particulier à grande échelle. Ses enjeux scientifiques sont :

- développer des algorithmes efficaces et robustes pour l'analyse et l'extraction de contenu multimédia, leur classification et leur analyse sémantique ;
- la reconstitution ou la fusion de données hétérogènes pour l'interprétation de scènes ou de documents ;
- développer des méthodes et des outils pour la construction, la formalisation et l'organisation des ressources et connaissances nécessaires au fonctionnement de ces algorithmes ;
- intégrer les méthodes d'analyse des contenus développées afin d'accéder à l'information et répondre à un besoin utilisateur spécifique (moteurs de recherche, agents conversationnels, rapports synthétiques de veille, *etc.*).

Description de la thématique TAL

Le CEA LIST développe depuis le début des années 2000 des travaux dans le domaine du traitement automatique des langues (TAL) dans une perspective d'accès au contenu des documents textuels en mettant l'accent sur le multilinguisme et le multimédia. Ces recherches, dans lesquelles le LASTI s'inscrit, ont été menées en tenant compte des contraintes posées par une perspective industrielle : d'une part, le développement de travaux permettant d'explorer l'intérêt des approches fondées sur l'apprentissage automatique ; d'autre part, la poursuite de travaux concernant des approches hors apprentissage dans les contextes où celles-ci ne sont pas adaptées, notamment en l'absence de volumes conséquents de données annotées. Dans ce cadre, l'activité du LASTI peut se décliner au travers des trois grandes thématiques suivantes.

Analyse linguistique multilingue. En dépit du développement récent des approches de bout en bout, le traitement du texte reste dépendant d'une analyse linguistique capable d'intégrer les spécificités propres aux différentes langues existantes. Pour gérer la problématique du multilinguisme qui en résulte, le LASTI développe depuis plusieurs années la plateforme d'analyse linguistique LIMA (LIbre Multilingual Analyzer) [1], qui offre la modularité nécessaire à la prise en compte la plus générique possible d'un large ensemble de langues tant du point de vue des traitements que de leurs ressources. Cette plateforme, sous licence libre AGPL pour l'anglais, le français et le portugais, prend en charge à des degrés divers l'analyse linguistique principalement au niveau phrastique en allant de la segmentation en mots jusqu'à l'analyse en rôles sémantiques pour un ensemble de 11 langues allant du français à l'arabe en passant par l'allemand et l'espagnol.

Le développement d'une plateforme généraliste d'analyse linguistique s'accompagne également de travaux visant son application à des contextes plus

spécifiques, en particulier au niveau sémantique. Dans le cadre du projet DECODER, le LASTI s'intéresse ainsi à l'application de l'analyse en rôles sémantiques aux commentaires associés à du code informatique et à sa documentation pour faire le lien avec des spécifications formelles tandis que le projet LabForSIMS 2 a permis de considérer la problématique de l'analyse linguistique de résultats de transcriptions de parole pour le développement d'agents conversationnels de formation des médecins [6].

Extraction et synthèse d'information. Au-delà de l'analyse linguistique au niveau phrastique, une part importante des recherches menées par le LASTI se focalisent sur les problématiques complémentaires d'extraction et de synthèse d'information, avec des applications en lien avec la veille. Concernant l'extraction d'information, cette focalisation touche deux extrêmes. D'un côté, elle s'intéresse au niveau des entités au travers de la tâche de désambiguïsation d'entité (*entity linking*) [3], avec le souci d'un passage à l'échelle et l'intégration nouvelle de la modalité visuelle. De l'autre, elle intervient au niveau plus macroscopique des événements en considérant les tâches de détection supervisée de ces événements et de leurs arguments. Les travaux menés sur cette thématique [5] mettent particulièrement l'accent sur la prise en compte du niveau discursif en dépassant le cadre souvent privilégié de la phrase. Ils s'enrichissent en outre de la mise en évidence des relations entre événements, que ce soient des relations temporelles [11] ou de coréférence, explorées toutes deux dans le domaine médical.

Les travaux sur la synthèse d'information s'inscrivent quant à eux principalement dans le cadre du résumé multi-document par extraction en y intégrant dans leur déclinaison la plus récente une dimension de mise à jour temporelle exploitant, dans un même cadre d'optimisation linéaire en nombres entiers, la similarité sémantique des phrases fondée sur des plongements lexicaux [8] et la structure discursive des textes selon le paradigme de la *rhetorical structure theory* (RST).

Adaptation à de nouveaux contextes. De par son positionnement à l'interface entre la recherche académique et les besoins industriels, le LASTI est confronté à la nécessité d'adapter les outils qu'il développe à des contextes applicatifs divers, ce qui représente à la fois une difficulté du point de vue de la réalisation d'applications industrielles mais aussi une problématique de recherche de plus en plus prégnante. Le LASTI développe ainsi différentes stratégies pour minimiser l'effort d'adaptation à un nouveau contexte applicatif. L'une d'elles consiste à s'appuyer sur des processus non supervisés. Les travaux menés dans le cadre de l'extraction d'information ouverte (*open information extraction*) [12] ont ainsi montré la possibilité d'extraire des relations de façon générique à partir d'un corpus et de caractériser leur type a posteriori par le biais de processus de regroupement. Cette capacité a été appliquée en particulier au domaine de la sécurité dans le cadre du projet ePOOLICE et au domaine médical. Le même type de problématique a été étendu aux schémas d'événements grâce à des approches bayésiennes hiérarchiques [9] et se décline pour ces mêmes schémas d'événements au travers du projet ASRAEL pour les contenus journalistiques.

Une autre voie pour satisfaire les besoins d'adaptation est de considérer la construction la plus automatisée possible de ressources linguistiques en s'appuyant notamment sur les méthodes de l'analyse distributionnelle. Le LASTI est ainsi impliqué dans le projet ADDICTE, qui s'attache aux difficultés rencontrées par ce type de méthodes en domaine de spécialité et poursuit par ailleurs des travaux sur la constitution de ressources concernant la similarité sémantique et les thésaurus distributionnels [4]. Il s'est aussi intéressé à l'acquisition de cadres sémantiques en soutien à l'analyse en rôles sémantiques [10] avec le projet ASFALDA. Enfin, il n'a pas négligé la problématique du multilinguisme dans ce champ de recherche avec des travaux sur l'acquisition automatique de lexiques bilingues à partir de corpus parallèles et comparables [2].

La dernière stratégie expérimentée par le LASTI pour s'adapter à de nouveaux contextes se focalise sur la capacité, grâce à des approches neuronales, à transposer les annotations réalisées pour un type de tâche donné d'un corpus à un autre. Cette voie a d'abord été explorée par le biais des méthodes de projection d'annotations entre corpus [13] et se trouve étendue à présent au travers de méthodes

d'apprentissage par transfert [7]. En particulier, ces méthodes ont été expérimentées dans le cadre du projet ASGARD pour construire automatiquement des outils d'analyse de textes des réseaux sociaux en exploitant les similarités entre les textes d'une langue bien dotée (forme standard d'une langue) et les textes d'une langue peu dotée (*tweets*).

Références

[1] Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Faiza Gara, Olivier Mesnard, Meriama Laïb, and Nasredine Semmar. LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. In *7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 3697–3704, Valletta, Malta, may 2010.

[2] Dhouha Bouamor, Adrian Popescu, Nasredine Semmar, and Pierre Zweigenbaum. Building specialized bilingual lexicons using large scale background knowledge. In *2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 479–489, Seattle, Washington, USA, 2013.

[3] Hani Daher, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, Anne-Laure Daquo, and Youssef Tamaazousti. Supervised learning of entity disambiguation models by negative sample selection. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, Budapest, Hungary, April 2017.

[4] Olivier Ferret. Using pseudo-senses for improving the extraction of synonyms from word embeddings. In *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), short paper session*, pages 351–357, Melbourne, Australia, 2018. Association for Computational Linguistics.

[5] Dorian Kodelja, Romaric Besançon, and Olivier Ferret. Exploiting a more global context for event detection through bootstrapping. In *41st European Conference on Information Retrieval (ECIR 2019) : Advances in Information Retrieval, short article session*, pages 763–770, Cologne, Germany, 2019. Springer International Publishing.

[6] Fréjus A. A. Laleye, Antonia Blanié, Antoine Brouquet, Dan Benhamou, and Gaël de Chalendar. Hybridation d'un agent conversationnel avec des plongements lexicaux pour la formation au diagnostic médical. In *23ème Conférence sur Le Traitement Automatique Des Langues Naturelles (TALN 2019)*, pages 313–321, Toulouse, France, 2019.

[7] Sara Meftah, Youssef Tamaazousti, Nasredine Semmar, Hassane Essafi, and Fatiha Sadat. Joint learning of pre-trained and random units for domain adaptation in part-of-speech tagging. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2019)*, pages 4107–4112, Minneapolis, Minnesota, 2019.

[8] Maâli Mnasri, Gaël de Chalendar, and Olivier Ferret. Taking into account inter-sentence similarity for update summarization. In *Eighth International Joint Conference on Natural Language Processing (IJCNLP 2017), short paper session*, pages 204–209, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[9] Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. Generative event schema induction with entity disambiguation. In *53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 188–197, Beijing, China, July 2015.

[10] Quentin Pradet, Laurence Danlos, and Gaël de Chalendar. Adapting verbnet to french using existing resources. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1122–1126, Reykjavik, Iceland, 2014.

[11] Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. Neural architecture for temporal relation extraction : A bi-lstm approach for detecting narrative containers. In *55th Annual Meeting of the Association for*

Computational Linguistics (ACL 2017), short paper session, pages 224–230, Vancouver, Canada, July 2017.

[12] Wei Wang, Romaric Besançon, Olivier Ferret, and Brigitte Grau. Semantic clustering of relations between named entities. In *9th International Conference on Natural Language Processing (PolTAL 2014)*, pages 358–370, Warsaw, Poland, september 2014. Springer International Publishing.

[13] Othman Zennaki, Nasredine Semmar, and Laurent Besacier. Inducing multilingual text analysis tools using bidirectional recurrent neural networks. In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 450–460, Osaka, Japan, December 2016.

AfIA
Association française
pour l'Intelligence Artificielle

LATTICE : Langues, Textes, Traitements Informatiques, Cognition

LATTICE UMR 8094
CNRS, École normale supérieure/PSL et
Université Sorbonne Nouvelle
http://www.lattice.cnrs.fr

Thierry POIBEAU
thierry.poibeau@psl.ens.eu

Membres (au 01/07/2021)

- Pascal AMSILI (PU Sorbonne nouvelle)
- Mathieu DEHOUCK (CR CNRS)
- Salomé DO (doctorante)
- Frédéric LANDRAGIN (DR CNRS)
- Karim LASRI (doctorant)
- Yuanfeng LU (doctorant)
- Mylène MAIGNANT (doctorante)
- Frédérique MÉLANIE-BECQUET (IE CNRS)
- Thierry POIBEAU (DR CNRS)
- Olga SEMINCK (post-doctorante)

Introduction

Le LATTICE (Langues, Textes, Traitements informatiques, Cognition) est un laboratoire CNRS hébergé dans les locaux de l'École normale supérieure. Depuis sa création, le laboratoire a développé une approche résolument pluridisciplinaire, à l'interface des différents domaines évoqués dans son acronyme. Les recherches en traitement automatique des langues (TAL) se déclinent selon trois axes de recherche au sein du laboratoire : i) la mise au point de corpus annotés pour le français, ii) le développement de techniques fondamentales pour le TAL, notamment dans le domaine de l'apprentissage artificiel, iii) la mise en application de ces techniques, pour les humanités numériques notamment.

Aperçu des recherches

Annotation de corpus. On dispose encore de peu de corpus représentatifs pour le français, surtout quand on s'intéresse à des tâches particulières, comme l'analyse de la coréférence ou de l'évolution de la langue. Le Lattice a développé ces dernières années plusieurs corpus utiles et mis à disposition de la communauté, comme le corpus Democrat pour l'analyse de la coréférence.

Le projet ANR Democrat, dirigé par Frédéric Landragin, a fait l'objet d'une présentation dans le bulletin de l'AFIA n° 92 [3]. Cinq ans plus tard, le projet est terminé et l'ensemble des objectifs a été atteint. Les membres du projet ont notamment publié un corpus dans lequel 198.000 expressions référentielles ont été identifiées et annotées manuellement, ce qui représente environ 20.000 chaînes de coréférences (9.000 si on ne compte que celles comportant au moins trois maillons). On atteint ici le même ordre de grandeur que le seul corpus comparable existant pour la langue française, à savoir le corpus ANCOR, qui s'intéresse aux anaphores plutôt qu'aux coréférences. Ce seuil permet d'exploiter ces deux corpus en tant que corpus d'apprentissage. De nombreuses expérimentations se sont ainsi déroulées de 2018 à 2020, et ont permis de mettre au point deux systèmes de détection automatique de chaînes de référence : premièrement le système DeCOFR [1], entraîné sur ANCOR (thèse de Loïc Grobol au Lattice, cf. [2]), puis le système COFR, entraîné sur ANCOR et Democrat [13]. Ces systèmes implémentent tous les deux les dernières avancées (disponibles en leur temps) du deep learning appliqué à la tâche. Au final, DeCOFR et COFR sont les premiers systèmes à être capables de détecter automatiquement les coréférences dans des textes tout-venant en français, avec des performances comparables à celles obtenues par les systèmes conçus pour la langue anglaise.

Outre Democrat, le Lattice a récemment développé d'autres corpus, comme **le corpus Cidre** (the Corpus for Idiolectal Research) [11] . Cidre est composé d'œuvres de fiction de 11 auteurs français prolifiques du XIXe siècle (4 femmes, 7 hommes ; 22-62 œuvres/auteur ; total de 37 millions de mots). Ce coprus doit permettre d'aborder l'étude de l'évolution de la langue d'un individu (évolution de l'idiolecte ou. stylochronométrie, quand c'est l'évolution du tyle qui est en jeu). Des modèles d'évolution du style ont ensuite été développés, à base de régression logistique.

Mise au point de modèles d'analyse efficaces. Le laboratoire a suivi ces dernières années le mouvement général vers l'apprentissage profond, qui a permis des progrès remarquables pour la plupart des tâches traditionnelles du TAL. Le recours à des modèles relativement génériques et la disponibilités de données d'entraînement dans des formats standardisés ont permis le développement en temps extrêmement réduit de modèles capables d'analyser une grande variété de langues de manière efficace. Les enjeux aujourd'hui restent importants et sont aussi bien théoriques que pratiques. Sur le plan théorique, le laboratoire travaille à une meilleure connaissance du fonctionnement interne de ces modèles, de leurs limites et de la façon dont l'information y est encodée (thèse de Karim Lasri). Sur le plan pratique, les enjeux concernent notamment la mise au point de modèles efficaces avec peu de données d'entraînement, pouvant être appliqués au traitement de langues en danger.

L'analyse en dépendances est une composante essentielle de nombreuses applications de TAL, dans la mesure où il s'agit de fournir une analyse des relations entre les principaux éléments de la phrase. La plupart des systèmes d'analyse en dépendances sont issus de techniques d'apprentissage supervisées, à partir de grands corpus annotés. Ce type d'analyse est dès lors limité à quelques langues seulement, qui disposent des ressources adéquates. Pour résoudre ce problème, KyungTae Lim, dans sa thèse [5], a examiné trois méthodes d'amorçage : l'apprentissage par transfert multilingue, les plongements vectoriels contextualisés profonds et le co-entrainement [6]. L'analyseur syntaxique correspondant a été évalué sur une soixantaine de langues à travers la participation aux campagnes d'évaluation proposées dans le cadre de la conférence CoNLL. Le système du Lattice a obtenu des résultats très compétitifs lors de campagnes d'évaluation officielles, notamment lors des campagnes CoNLL 2017 et 2018. Ce travail offre donc des perspectives intéressantes pour le traitement automatique des langues peu dotées, un enjeu majeur pour le TAL dans les années à venir [7].

L'analyse d'erreur est aussi un point clé pour mieux comprendre et améliorer les systèmes d'analyse. Par exemple, dans le cadre du projet ANR Profiterole dont un des objectifs est l'annotation d'un corpus de français médiéval d'un million de mots, Mathieu Dehouck développe des modèles pour la détection automatique d'erreur en post-traitement pour un parseur syntaxique en dépendances. L'idée est de parser automatiquement des données annotées à la main et pour créer un jeu d'erreurs réalistes puis de l'utiliser pour entraîner un modèle à détecter les erreurs d'analyse. On aimerait ensuite combiner les prédictions de plusieurs modèles (statistiques ou à règles) pour avoir différents niveaux de supervision : une supervision venant directement des données annotées à la main ainsi qu'une supervision plus distante venant de l'accord ou du désaccord des différents modèles d'analyse.

Enjeux applicatifs : l'apport du TAL à l'analyse linguistique. Des travaux en cours (menés par M. Dehouck, à partir de données fournies notamment par A. François) portent sur la découverte de groupes de langues à partir de données d'isoglosses dans des modèles non plus arborescents (phylogénétiques) mais diffusionnistes. L'idée ici est d'essayer de réconcilier les méthodes de la dialectologie avec les méthodes de la linguistique comparatives. En posant que la diffusion entre les langues continue même après la séparation de fait de leurs locuteurs et que ces contacts sont mal représentés par les modèles arborescents, on essaye alors de reconstruire des groupes d'affinité entre les langues qui peuvent se chevaucher.

D'autres recherches s'intéressent à l'interface linguistique / TAL / sciences cognitives. Des travaux en cours portent par exemple sur la résolution et la modélisation des schémas de Winograd (cas particuliers du problème de résolution anaphorique, qui ne peuvent être résolus qu'en faisant appel à un raisonnement sur des connaissances du monde, recherches menées par P. Amsili et O. Seminck notamment [12]). On s'intéresse enfin à l'apport des méthodes neuronales pour la modélisation de la portée de la négation en français, ou aux liens possibles entre TAL, universaux linguistique et typologie, des domaines qui s'ignorent assez largement [8].

Application aux Humanités numériques. Les Humanités numériques constituent un domaine

d'investigation particulièrement prometteur pour les techniques de TAL, car les questions abordées sont liées à des données complexes pouvant porter sur des états de langue anciens et/ou comporter une dimension diachronique. Jusqu'ici le laboratoire a surtout travaillé (à travers notamment la thèse de Pablo Ruiz Fabo [10]) sur des cas liés aux sciences sociales (négociations climatiques [9], corpus PoliInformatics) et au domaine philosophie (projet Mapping Bentham). L'exploration de l'interface entre TAL et sciences sociales se poursuit actuellement à travers la thèse de Salomé Do, en partenariat avec le médialab de Sciences Po.

D'autres recherches portent sur l'analyse semi-automatique du discours littéraire. On s'intéresse par exemple à la détection de configurations lexico-grammaticales — appelées « motifs » — caractéristiques d'un genre textuel. Ces motifs expriment une des dimensions phraséologiques des romans, en particulier des romans sentimentaux, à savoir le cliché [4]. Deux thèses sont actuellement en cours dans ce domaine : la thèse de Yuanfeng Lu, sur l'analyse stylistique à base de motifs syntaxiques, et la thèse de Mylène Maignant, sur l'analyse contrastive de critique théâtrale en anglais (corpus journalistitique vs corpus issu de blogs en ligne)

Le laboratoire développe enfin **le versant français du projet Multilingual BookNLP**, BookNLP est à l'origine un projet de l'Université de Berkeley, visant à développer des outils de TAL adaptés à la littérature : les outils sont en effet conçus autour de tâches génériques (analyse morphosyntaxique, syntaxique, etc.), utiles mais peu adaptées au genre romanesque (la notion d'entité nommée est par exemple très insuffisante face à la notion de personnage). Or, l'expérience a montré qu'il était possible d'obtenir des outils efficaces pour l'étude de grands corpus littéraires, au moins pour l'anglais. Le projet vise à concevoir des outils d'annotation similaires pour le français. Actuellement, une vingtaine d'extraits de romans (du 19e et début 20e, libres de droit, environ 170.000 mots) ont été annotés au Lattice suivant le modèle BookNLP révisé (et en reprenant l'annotation en chaîne de coréférence du projet ANR Democrat, qui a constitué une base solide pour l'annotation). La mise au point des modèles d'annotation, l'évaluation de ces modèles et leur amélioration éventuelle avec des techniques d'apprentissage avancées (co-amorçage, apprentissage supervisée, etc.) devrait permettre d'obtenir à terme des performances identiques à celles de l'anglais.

Références

[1] Loïc Grobol. Neural coreference resolution with limited lexical context and explicit mention detection for oral french. In *Second Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC19-NAACL 2019)*, Minneapolis, 2019.

[2] Loïc Grobol. *Reconnaissance automatique de chaînes de coréférences en français par combinaison d'apprentissage automatique et de connaissances linguistiques*. Thèse, Univ. Sorbonne nouvelle, Paris, 2020.

[3] Frédéric Landragin. Description, modélisation et traitement automatique des chaînes de référence (Democrat). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, 92, 2016.

[4] Dominique Legallois, Thierry Charnois, and Thierry Poibeau. Identifying clichés in romance novels using the "motifs" method. *LIDIL - Revue de linguistique et de didactique des langues*, 2016.

[5] KyungTae Lim. *Méthodes d'amorçage pour l'analyse en dépendances de langues peu dotées*. Thèse, Univ. PSL, Paris, 2020.

[6] KyungTae Lim, Jay Yoon Lee, Jaime Carbonell, and Thierry Poibeau. Semi-supervised learning on meta structure : Multi-task tagging and parsing in low-resource scenarios. In *Proceedings of the AAAI Conference*, 2020.

[7] KyungTae Lim, Niko Partanen, and Thierry Poibeau. Multilingual dependency parsing for low-resource languages : Case studies on north saami and Komi-Zyrian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

[8] Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling Language Variation and Universals :

A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3) :559–601, 2019.

[9] Pablo Ruiz, Clément Plancq, and Thierry Poibeau. More than word cooccurrence : Exploring support and opposition in international climate negotiations with semantic parsing. In *Language Resources and Evaluation Conf. (LREC)*, 2016.

[10] Pablo Ruiz Fabo. *Concept-based and relation-based corpus navigation : applications of natural language processing in digital humanities*. Thèse, Univ. PSL, Paris, 2016.

[11] Olga Seminck, Philippe Gambette, Dominique Legallois, and Thierry Poibeau. The Corpus for Idiolectal Research (CIDRE). *Journal of Open Humanities Data*, 7, 2021.

[12] Olga Seminck, Vincent Segonne, and Pascal Amsili. Modèles de langue appliqués aux schémas Winograd français (language models applied to French Winograd schemas). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 343–350, Toulouse, France, 2019.

[13] Rodrigo Wilkens, Bruno Oberlé, Frédéric Landragin, and Amalia Todirascu. French coreference for spoken and written language. In *Language Resources and Evaluation Conference (LREC 2020)*, Marseille, 2020.

AfIA
Association française
pour l'Intelligence Artificielle

LIA : Laboratoire Informatique d'Avignon

LIA EA 4128
Avignon Université
https://lia.univ-avignon.fr

Corinne FREDOUILLE
Coordinatrice du bulletin au LIA
corinne.fredouille@univ-avignon.fr

Jean-François BONASTRE
Directeur du LIA (2016-2020)
jean-francois.bonastre@univ-avignon.fr

Mots-clés

Langage oral et écrit, locuteur, troubles de la parole et de la voix, interactions vocales, réseaux complexes, réseaux sociaux, partenariats industriels.

Le LIA, les TLH et l'IA

Le Laboratoire Informatique d'Avignon (LIA) a été fondé en 1987 par le Professeur Henri MÉLONI, anciennement chercheur du groupe d'intelligence artificielle de Luminy, à Marseille, et l'un des pères fondateurs du langage Prolog. À cette époque, les activités de recherche du laboratoire étaient dédiées exclusivement au traitement de la parole, impliquant des approches à base de règles, de la modélisation logique et symbolique, et des méthodes d'apprentissage automatique (avec, déjà, des réseaux de neurones). Les activités du LIA se sont progressivement étendues, au cours des années, vers le traitement automatique du langage (TAL) au sens large, mixant oral, écrit, et traces dans les réseaux sociaux. Elles constituent le thème phare du laboratoire.

Outre le thème du langage, le laboratoire en compte aujourd'hui deux autres principaux : les réseaux et la recherche opérationnelle. Le LIA se compose d'un peu plus d'une trentaine d'enseignants-chercheurs permanents et chercheurs associés, et d'autant de doctorants et post-doctorants. Il s'appuie sur trois ingénieurs permanents pour les questions techniques et d'une ingénieure à mi-temps dédiée à l'accompagnement de la recherche contractuelle.

La thématique Langage

La thématique Langage du LIA couvre un large faisceau d'activités dont quelques exemples de champs d'application sont donnés ci-après. Au travers de cette thématique, le LIA est membre de l'Institut Carnot Cognition, du LABEX BLRI et de l'IC ILCB.

Transcription-Traduction de la parole et détection d'événements

La thématique de la reconnaissance automatique de la parole existe au LIA depuis sa création. Cette activité de recherche y est toujours présente et le LIA maîtrise depuis longtemps les approches classiques markoviennes, y compris combinées à des réseaux de neurones profonds. La tâche de reconnaissance de la parole est aujourd'hui souvent couplée à d'autres tâches. Le LIA s'est investi depuis quelques années dans les approches neuronales profondes de bout-en-bout. Elles offrent la possibilité d'une optimisation jointe de l'ensemble des sous-modules impliqués dans des tâches complexes, et offrent l'opportunité de réduire la propagation des erreurs typiques des approches séquentielles.

Ainsi, le LIA coordonne le projet ANR ON-TRAC sur les approches neuronales profondes de bout-en-bout pour la traduction de la parole. Cela consiste à traduire directement, sans passer par une transcription intermédiaire, le signal de parole d'une langue source en texte d'une langue cible. Le LIA est en pointe dans ce domaine, comme le montrent les résultats de sa participation [36] via le consortium ON-TRAC à la campagne d'évaluation internationale IWSLT 2019.

Toujours dans l'exploration des approches neuronales profondes de type bout-en-bout à partir de la parole, le LIA aborde des problèmes de reconnaissance d'entités nommées [16] ou d'extraction de concepts sémantiques [46]. Il a très récemment proposé une approche d'apprentissage par transfert

s'appuyant sur une stratégie de curriculum qui permet de pallier le manque récurrent de données spécialisées annotées manuellement pour ce type de tâches [8].

Le LIA travaille depuis quelques années sur les problématiques de recherche de mots-clés dans des documents audios et sur les problématiques de *wake word*. Un *wake word* est un mot ou une phrase énoncée par une personne qui permet d'activer un appareil dormant. Ces approches sont notamment utilisées dans les assistants vocaux. Le LIA a travaillé sur des méthodes d'augmentation de données et d'apprentissage par transfert pour les réseaux de neurones profonds afin de rendre le réseau plus robuste et minimiser les fausses alertes d'activation.

Interaction vocale et agents conversationnels

L'interaction vocale entre l'humain et la machine est un défi de grande importance pour un large nombre de systèmes d'accès à l'information (web, serveurs vocaux, applications mobiles ...) ou en robotique (robots compagnons, *etc.*). Nous allons non seulement vers des systèmes plus performants, mais aussi plus naturels et plus interactifs qui prennent mieux en compte l'utilisateur [29]. Les questions que nous abordons dans cette activité sont par exemple : l'apprentissage en ligne d'agents conversationnels, le dialogue situé ou encore le développement d'interfaces emphatiques qui s'adaptent aux états émotionnels des utilisateurs ; en particulier nous questionnons le rôle possible de l'humour.

Les contributions du LIA s'inscrivent dans trois axes principaux :

1. La *compréhension de la parole* : pour une compréhension sans données d'apprentissage, nous avons proposé une approche basée sur des représentations vectorielles de mots complétée par un processus d'apprentissage en ligne permettant d'obtenir des utilisateurs les informations manquantes, tout en maîtrisant le coût engendré par cette étape [13, 42].
2. La *gestion du dialogue situé* : la mise en situation des systèmes interactifs, *i.e.* la machine partage le même environnement physique que l'humain, améliore leur performance. Dans le cadre du projet ANR MaRDi, le développement d'un système de dialogue basé sur des POMDP a permis une prise en compte améliorée de l'incertitude et l'optimisation automatique de la politique d'interaction grâce à l'apprentissage par renforcement en interaction directe avec les utilisateurs [12]. La méthode étendue permet la prise en compte d'informations situées, obtenues par d'autres modalités [14], ou une gestion plus naturelle des tours de parole [22].
3. La *génération de parole* : nous avons proposé une approche à base d'apprentissage automatique pour faciliter la conception d'un système de génération en langue naturelle, ainsi que plusieurs approches pour étendre les corpus de génération afin d'intégrer plus de variabilité dans les productions [43]. Un autre objectif est de rendre les systèmes plus naturels en automatisant la production de traits humoristiques dans un dialogue humain-machine. Un tel effet décalé devrait augmenter la dimension de sympathie envers le système d'interaction dans la perception de l'utilisateur. Plusieurs types de production automatique d'humour ont été élaborés, associés à un mécanisme d'apprentissage par renforcement permettant au système de dialogue d'apprendre une politique de gestion de production humoristique à partir des satisfactions des usagers [41].

Voix et identité : authentification, comparaison de voix en criminalistique, détection des fraudes, anonymisation de voix

Le LIA est un acteur reconnu en reconnaissance du locuteur et propose plusieurs outils pour l'authentification par la voix. Il a construit et maintient la plateforme « libre » ALIZE qui facilite la mise en place d'applications sur diverses architectures dont Android et s'est largement spécialisé sur la question de l'adaptation des systèmes à de nouveaux domaines d'utilisation [2, 28]. Les approches développées au LIA sont systématiquement évaluées dans le cadre de campagnes d'évaluation internationales et font l'objet de plusieurs collaborations académiques et industrielles (dont le projet ANR ROBOVOX). Le LIA travaille notamment sur l'adaptation au domaine [5]. Le LIA est un acteur actif dans le cadre des contremesures contre les attaques des systèmes de reconnaissance du locuteur [27], ainsi

que dans le domaine de l'anonymisation de la voix. Il est membre du projet bilatéral France (ANR) Japon (JST) « VoicePersonaë », dédié à ces sujets.

Par ailleurs, le LIA est présent dans le débat relatif à l'usage de la comparaison de voix dans le domaine criminalistique/judiciaire [1, 7]. Le LIA a coordonné le projet ANR FABIOLE et coordonne actuellement le projet ANR VoxCrim, deux projets dédiés à la mesure de la fiabilité en comparaison de voix.

Enfin, le LIA travaille sur la recommandation de voix par similarité pour la production de contenus dans le secteur de l'industrie créative en vue de faciliter le casting ou la génération de voix artificielles (projet ANR TheVoice).

Troubles de la parole et de la voix

Ils sont définis par les difficultés, voire l'incapacité, pour un locuteur de produire des sons articulés et modulés pour former des mots compréhensibles dans un acte de communication. Les maladies neurodégénératives (par ex. maladie de Parkinson, sclérose en plaque, accidents vasculaires cérébraux) touchant le système nerveux central (cerveau, tronc cérébral, cervelet et moelle épinière) et/ou périphérique (nerfs crâniens et spinaux) ainsi que les cancers des voies aéro-digestives supérieures, suivant la localisation de la tumeur, peuvent être la cause de troubles de la parole. En complément, la dysphonie est une altération de la voix qui touche de manière plus ou moins sévère l'un des trois paramètres acoustiques caractéristiques d'une voix, la hauteur, l'intensité ou le timbre, de manière isolée ou combinée. Si la parole n'est pas affectée en cas de dysphonie seule, l'acte de communication peut, pour sa part, être gravement perturbé.

Impliqué depuis 2004 dans des travaux de recherche pluridisciplinaires appliqués à la caractérisation et à l'évaluation des troubles de la parole et de la voix, les objectifs du LIA sont de répondre à une demande récurrente de la part des praticiens d'outils objectifs d'évaluation du niveau de sévérité des altérations de parole et/ou de la voix observées chez les patients et/ou du niveau d'intelligibilité. En effet, dans le cadre de la prise en charge thérapeutique ou du suivi longitudinal d'un patient, après traitement thérapeutique ou rééducation, le seul outil actuellement à disposition du praticien est l'évaluation perceptive (« à l'oreille »), dont le caractère subjectif et non reproductible est bien illustré dans la littérature. Dans les activités les plus récentes, nous citerons les travaux autour de (1) la détection automatique d'anomalies dans la parole dysarthrique pour la caractérisation et l'évaluation des troubles de parole [25, 26], (2) la modélisation par des i-vecteurs de productions de parole altérée pour une tâche de prédiction du niveau d'intelligibilité [23, 24]. Par ailleurs, nous travaillons de concert avec le Laboratoire Parole et Langage (LPL) sur un protocole original d'évaluation de l'intelligibilité en milieu clinique [17]. L'apport du LIA est ici de fournir des outils automatiques adaptés à ce protocole particulier [15]. Pour finir, le LIA est investi dans des travaux impliquant des approches de *deep learning* et d'interprétabilité dans le cadre du projet ANR/RUGBI portant sur la caractérisation de l'intelligibilité en partenariat avec les laboratoires IRIT, Lordat-Octogone et le CHU de Toulouse et le LPL.

Réseaux sociaux, réseaux complexes et TAL

Le Web, et plus spécifiquement les médias sociaux qu'il héberge, est devenu depuis sa création un espace d'échange mondialisé par lequel transitent et sont partagées d'énormes quantités de données multimédia (texte, audio et vidéo), qui de plus ne cessent de croître. Ce média est devenu un terrain de recherche privilégié pour de nombreuses études scientifiques exploitant ces données. Parmi les travaux entrepris au LIA, les échanges entre utilisateurs sous forme textuelle ont suscité l'intérêt dans plusieurs problématiques du TAL, comme la prédiction de buzz [32], la détection d'opinions [44, 11], ou encore l'analyse temporelle du contenu de messages courts [40]. Les différentes méthodes proposées dans ces travaux ont ainsi dû faire face à des problèmes nouveaux propres à ce mode de diffusion, en particulier par rapport aux textes écrits traités jusque-là en TAL. On peut notamment lister un vocabulaire souvent particulier et/ou non-standard, de nombreuses erreurs grammaticales ou orthographiques, et des contenus pouvant être très courts.

Ces difficultés rendent alors le traitement automatique du contenu textuel compliqué. Des travaux récents entrepris au LIA ont notamment montré,

dans le cadre de la détection d'abus dans des discussions en ligne, que l'analyse du contenu textuel n'était pas l'approche la plus efficace pour ce type de problème. En effet, nous avons montré dans [37] que la modélisation des échanges entre utilisateurs (*i.e.* qui parle à qui), sans tenir compte de leur contenu, pouvait permettre de réaliser de meilleures prédictions. Les caractéristiques issues de ces réseaux conversationnels apparaissent alors plus robustes que des caractéristiques textuelles, tout en ayant montré que ces deux sources d'information peuvent être complémentaires en termes d'information [9].

Le LIA a exploité le même type de réseau conversationnel dans un contexte très différent : celui de la génération de résumé automatique de séries TV [3]. Un type dynamique de réseau conversationnel a été proposé pour modéliser l'évolution des échanges entre les personnages de la série, et ainsi indirectement son intrigue. Ce réseau a ensuite pu être exploité pour déterminer les scènes-clés de la série, et ainsi construire un résumé extractif.

De nombreuses collaborations ont été entreprises sur l'exploitation de données textuelles ou multimédia issues des réseaux sociaux, que nous retrouvons dans des projets financés récents tels que les projets ANR RPM2 et GAFES. Plus récemment, certains travaux se concentrent sur une source différente, elle aussi en ligne : les données publiques ouvertes, telles que le Bulletin officiel des annonces des marchés publics, qui est actuellement exploité dans le cadre du projet ANR DéCoMaP. Certains travaux en cours visent à combiner différentes modalités d'information relatives au contenu (texte, multimédia) mais aussi à la structure (réseau social en ligne, interconnexion par hyperliens), au moyen de méthodes de plongements (embeddings) [10]. Celles-ci permettent d'élaborer automatiquement une représentation compacte multimodale pouvant ensuite être exploitée dans différentes tâches telles que la détection d'abus ou la recommandation.

Anonymisation de texte

Le LIA s'est engagé au mois de juin 2019 dans une collaboration avec une société aixoise et un cabinet d'avocats vauclusien sur le développement d'une application d'anonymisation automatique de décisions de justice.

Cette problématique va au-delà du cadre général de l'extraction d'entités nommées. En effet, il s'agit bien entendu de repérer dans un texte les informations permettant d'identifier des personnes : prénoms, noms, adresses mail, adresses postales, diverses immatriculations, *etc.* Cependant, si certaines de ces informations doivent impérativement être anonymisées (justiciables, témoins, professionnels de santé, *etc.*), d'autres informations doivent de préférence ou absolument être préservées (noms des auxiliaires de justice, noms et adresses des sociétés et institutions, numérotage des articles de lois, *etc.*) et l'étude du contexte dans lequel ces entités nommées apparaissent est évidemment déterminant.

L'absence de corpus francophone annoté syntaxiquement et sémantiquement dans le domaine juridique constitue une autre difficulté majeure pour le développement de l'application.

Enfin, si les fausses détections peuvent avoir un impact très négatif sur l'intelligibilité du texte anonymisé, une non-anonymisation erronée peut très concrètement mettre en danger des personnes.

Le défi est donc d'aboutir à une application qui vise un rappel le plus proche possible de la perfection et une précision suffisante pour permettre au lecteur de tirer profit du texte anonymisé dans son analyse de la jurisprudence. Nous utilisons actuellement une approche hybride qui utilise l'annoteur syntaxique Talismane [51], fondé sur les CRF (*conditional random fields*), des algorithmes classiques d'ingénierie documentaire et un système de règles.

Résumé automatique et génération de texte

Soit dans sa forme textuelle [48], soit multimédia [21, 45], le résumé automatique fait partie des activités de recherche au LIA. Il vise à créer une version condensée d'un document source ayant un genre reconnaissable et à donner à l'utilisateur une idée précise et concise de la source. Différents systèmes de résumé automatique du texte ont été développés [4, 50, 47].

Le résumé automatique multimédia et multilingue a fait l'objet du projet CHIST-ERA/ANR-AMIS [38, 39]. Ce projet a donné l'opportunité de traiter la langue arabe standard, l'anglais et le fran-

çais en développant des techniques de résumé automatique basées sur du texte et la parole [18]. Le résumé automatique de texte est très lié à la compression de phrases [30, 31] qui vise à produire une phrase de petite taille, à la fois grammaticalement correcte et informative. La compression multiphrase est une variation de la compression de phrases qui vise à combiner les informations d'un groupe de phrases similaires pour générer une nouvelle phrase, grammaticalement correcte, qui comprime les données les plus pertinentes de ce groupe. La détection de limite de phrase est une tâche intermédiaire entre la reconnaissance de la parole et le résumé automatique. Cette tâche est essentielle pour trouver les frontières de phrases entre deux mots qui ont été transcrits, et ainsi arriver à produire des résumés cohérents et informatifs. Nous traitons aussi cette tâche avec une approche multilingue [19, 20].

La génération automatique de phrases littéraires est un domaine peu exploré de la Créativité computationnelle. Au LIA nous avons développé des outils pour la génération de phrases littéraires suivant des traits psychologiques [35], des corpus spécialisés [33] et des outils d'évaluation de la qualité des émotions produites par ces phrases [34].

Finalement, l'évaluation de la qualité des résumés, sujet controversé dû à son subjectivité, a également été étudiée au LIA au moyen des méthodes qui n'ont pas besoin des références humaines mais qui sont basées sur les divergences et trivergences de probabilités [49, 6].

PartnersLIA et événements

Les activités du LIA sont largement soutenues par des contrats de collaboration avec des industriels locaux et nationaux, incluant start-up, PME et grands groupes industriels, avec une forte présence dans ce secteur de la thématique langage. Pour dynamiser et renforcer ces collaborations, le LIA a créé en 2017 le PartnersLIA (club des partenaires industriels du LIA) et organise des actions et journées thématiques (par ex. une journée sur le big data en 2018) auxquelles l'ensemble de ces partenaires sont conviés en tant qu'invités ou participants.

Le LIA a co-organisé le 28 novembre 2019 la première édition des journées « IA Région Sud », sous l'égide de l'Institut 3IA Côte d'Azur avec Aix-Marseille Université, Avignon Université et les universités de Sophia-Antipolis et de Toulon.

Références

[1] Jean-François Bonastre, Frédéric Bimbot, Louis Jean Boë, Joseph P. Campbell, Douglas A. Reynolds, and Ivan Magrin-Chagnolleau. Person authentication by voice : A need for caution. In *Proc. of Eurospeech*, Genova, 2003.

[2] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas WD Evans, Benoit GB Fauve, and John SD Mason. Alize/spkdet : a state-of-the-art open source software for speaker recognition. In *Odyssey*, page 20, 2008.

[3] X. Bost, S. Gueye, V. Labatut, M. Larson, G. Linarès, D. Malinas, and R. Roth. Remembering winter was coming : Character-oriented video summaries of tv series. *Multimedia Tools and Applications*, 78(24) :35373–35399, 2019.

[4] Florian Boudin and Juan Manuel Torres Moreno. Neo-cortex : A performant user-oriented multi-document summarization system. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 551–562. Springer, 2007.

[5] Pierre-Michel Bousquet and Mickaël Rouvier. On robustness of unsupervised domain adaptation for speaker recognition. In *Proc. of Interspeech'2019*, Austria, 2019.

[6] Luis Adrián Cabrera-Diego and Juan-Manuel Torres-Moreno. Summtriver : A new trivergent model to evaluate summaries automatically without human references. *Data Knowl. Eng.*, 113 :184–197, 2018.

[7] Joseph P. Campbell, Wade Shen, William M. Campbell, Reva Schwartz, Jean-François Bonastre, and Driss Matrouf. Forensic Speaker Recognition. *IEEE Signal Processing Magazine*, 26(2) :95–103, 2009.

[8] Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Ca-

melin, and Yannick Estève. Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *Interspeech*, 2019.

[9] N. Cécillon, V. Labatut, R. Dufour, and G. Linarès. Abusive language detection in online conversations by combining content-and graph-based features. *Frontiers in Big Data*, 2 :8, 2019.

[10] N. Cécillon, V. Labatut, R. Dufour, and G. Linarès. Graph embeddings for abusive language detection. *Springer Nature Computer Science*, 2 :37, 2021.

[11] Richard Dufour, Mickael Rouvier, Alexandre Delorme, and Damien Malinas. Lia@ clef 2018 : Mining events opinion argumentation from raw unlabeled twitter data using convolutional neural network. In *CLEF (Working Notes)*, 2018.

[12] Emmanuel Ferreira and Fabrice Lefèvre. Reinforcement-learning based dialogue system for human-robot interactions with socially-inspired rewards. *Computer Speech & Language, Special issue on Speech and Language for Interactive Robots*, 34(1) :256–274, 2015.

[13] Emmanuel Ferreira, Alexandre Reiffers Masson, Bassam Jabaian, and Fabrice Lefèvre. Adversarial bandit for online interactive active learning of zero-shot spoken language understanding. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, pages 6155–6159, Shanghai, China, mar 2016.

[14] Emmanuel Ferreira, Grégoire Milliez, Fabrice Lefèvre, and Rachid Alami. *Users' Belief Awareness in Reinforcement Learning-Based Situated Human-Robot Dialogue Management*, pages 73–86. Springer International Publishing, 2015.

[15] Corinne Fredouille, Alain Ghio, Imed Laaridh, Muriel Lalain, and Virginie Woisard. Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. In *Proceedings of Intl Congress of Phonetic Sciences (ICPhS'19)*, Melbourne, Australia, 2019.

[16] Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE, 2018.

[17] Alain Ghio, Muriel Lalain, Laurence Giusti, Gilles Pouchoulin, Danièle Robert, Marie Rebourg, Corinne Fredouille, Imed Laaridh, and Virginie Woisard. Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. In *Journée d'Etudes sur la Parole, JEP'18, Aix-en-Provence, France*, pages 285–293, 2018.

[18] Carlos-Emiliano González-Gallardo, Romain Deveaud, Eric Sanjuan, and Juan-Manuel Torres-Moreno. Audio Summarization with Audio Features and Probability Distribution Divergence. In *20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France, April 2019.

[19] Carlos-Emiliano González-Gallardo, Elvys Linhares Pontes, Fatiha Sadat, and Juan-Manuel Torres-Moreno. Automated sentence boundary detection in modern standard arabic transcripts using deep neural networks. *Procedia Computer Science*, 142 :339–346, 2018.

[20] Carlos-Emiliano González-Gallardo and Juan-Manuel Torres-Moreno. Sentence boundary detection for french with subword-level information vectors and convolutional neural networks. *arXiv preprint arXiv :1802.04559*, 2018.

[21] Michał Grega, Kamel Smaïli, Mikołaj Leszczuk, Carlos-Emiliano González-Gallardo, Juan-Manuel Torres-Moreno, Elvys Linhares Pontes, Dominique Fohr, Odile Mella, Mohamed Menacer, and Denis Jouvet. An integrated amis prototype for automated summarization and translation of newscasts and reports. In Kazimierz Choroś, Marek Kopel, Elżbieta Kukla, and Andrzej Siemiński, editors, *Multimedia and Network Information Systems*, pages 415–423, Cham, 2019. Springer International Publishing.

[22] Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. A methodology for turn-taking capabilities enhancement in Spoken Dialogue Systems using Reinforcement Learning. *Computer Speech & Language*, 47 :93–111, jan 2018.

[23] Imed Laaridh, Waad Ben Kheder, Corinne Fredouille, and Christine Meunier. Automatic prediction of speech evaluation metrics for dysarthric speech. In *Proceedings of Interspeech'17*, pages 1834–1838, 2017.

[24] Imed Laaridh, Corinne Fredouille, Alain Ghio, Muriel Lalain, and Virginie Woisard. Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers. In *Proceedings of Interspeech'17 18*, pages 2943–2947, Hyderabad, India, 2018.

[25] Imed Laaridh, Corinne Fredouille, and Christine Meunier. Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on accessible computing*, 6(3) :9 :1–9 :24, May 2015.

[26] Imed Laaridh, Christine Meunier, and Corinne Fredouille. Perceptual evaluation for automatic anomaly detection in disordered speech : Focus on ambiguous cases. *Speech Communication*, 105 :23–33, 2018.

[27] Itshak Lapidot and Jean-François Bonastre. Effects of waveform pmf on anti-spoofing detection. In *Proc. of Interspeech'2019*, Austria, 2019.

[28] Anthony Larcher, Jean-François Bonastre, Benoit GB Fauve, Kong-Aik Lee, Christophe Lévy, Haizhou Li, John SD Mason, and Jean-Yves Parfait. Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *Interspeech*, pages 2768–2772, 2013.

[29] Fabrice Lefèvre. En route to a better integration and evaluation of social capacities in vocal artificial agents. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents - ISIAA 2017*, pages 15–19, New York, USA, 2017. ACM Press.

[30] Elvys Linhares Pontes, Stéphane Huet, Thiago Gouveia da Silva, Andréa carneiro Linhares, and Juan-Manuel Torres-Moreno. Multi-sentence compression with word vertex-labeled graphs and integer linear programming. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 18–27, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics.

[31] Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, and Andréa Carneiro Linhares. Cross-language text summarization using sentence and multi-sentence compression. In Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems*, pages 467–479, Cham, 2018. Springer International Publishing.

[32] Mohamed Morchid, Georges Linares, and Richard Dufour. Characterizing and predicting bursty events : The buzz case study on twitter. In *LREC*, pages 2766–2771, 2014.

[33] Luis-Gil Moreno-Jiménez and Juan-Manuel Torres-Moreno. Lisss : A new corpus of literary spanish sentences for emotions detection. *Computación y Sistemas*, 24(3), 2020.

[34] Luis-Gil Moreno-Jiménez, Juan-Manuel Torres-Moreno, Hanifa Boucheneb, and Roseli S. Wedemann. FLE : A fuzzy logic algorithm for classification of emotions in literary corpora. In Ana L. N. Fred and Joaquim Filipe, editors, *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2020, Volume 1 : KDIR, Budapest, Hungary, November 2-4, 2020*, pages 202–209. SCITEPRESS, 2020.

[35] Luis-Gil Moreno-Jiménez, Juan-Manuel Torres-Moreno, and Roseli S. Wedemann. Literary natural language generation with psychological traits. In Elisabeth Métais, Farid Meziane, Helmut Horacek, and Philipp Cimiano, editors, *Natural Language Processing and Information Systems - 25th International Conference on Applications of Natural Language to Information Systems,*

NLDB 2020, Saarbrücken, Germany, June 24-26, 2020, Proceedings, volume 12089 of *Lecture Notes in Computer Science*, pages 193–204. Springer, 2020.

[36] Manh Ha Nguyen, Natalia Tomashenko, Marcely Zanon Boito, Antoine Caubrière, Fethi Bougares, Mickael Rouvier, Laurent Besacier, and Yannick Estève. On-trac consortium end-to-end speech translation systems for the iwslt 2019 shared task. In *16th International Workshop on Spoken Language Translation 2019 (IWSLT)*, 2019.

[37] E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems*, 6(1) :38–55, 2019.

[38] Elvys Linhares Pontes, Stéphane Huet, and Juan-Manuel Torres-Moreno. A multilingual study of compressive cross-language text summarization. In Ildar Batyrshin, María de Lourdes Martínez-Villaseñor, and Hiram Eredín Ponce Espinosa, editors, *Advances in Computational Intelligence*, pages 109–118, Cham, 2018. Springer International Publishing.

[39] Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, and Andréa Carneiro Linhares. Compressive approaches for cross-language multi-document summarization. *Data & Knowledge Engineering*, 2019.

[40] M. Quillot, C. Ollivier, R. Dufour, and V. Labatut. Exploring temporal analysis of tweet content from cultural events. In *International Conference on Statistical Language and Speech Processing*, pages 82–93, 2017.

[41] Matthieu Riou, Bassam Jabaian, Stéphane Huet, Thierry Chaminade, and Fabrice Lefèvre. Integration and evaluation of social competences such as humor in an artificial interactive agent. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents - ISIAA 2017*. ACM Press, 2017.

[42] Matthieu Riou, Bassam Jabaian, Stéphane Huet, and Fabrice Lefèvre. Joint On-line Learning of a Zero-shot Spoken Semantic Parser and a Reinforcement Learning Dialogue Manager. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3072–3076. IEEE, 2019.

[43] Matthieu Riou, Bassam Jabaian, Stéphane Huet, and Fabrice Lefèvre. Reinforcement adaptation of an attention-based neural natural language generator for spoken dialogue systems. *Dialogue & Discourse*, 10 :1–19, 2019.

[44] Mickael Rouvier and Benoit Favre. Sensei-lif at semeval-2016 task 4 : Polarity embedding fusion for robust sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 202–208, 2016.

[45] Kamel Smaïli, Dominique Fohr, Carlos-Emiliano González-Gallardo, Michał Grega, Lucjan Janowski, Denis Jouvet, Arian Koźbiał, David Langlois, Mikołlaj Leszczuk, Odile Mella, et al. Summarizing videos into a target language : Methodology, architectures and evaluation. *Journal of Intelligent & Fuzzy Systems*, (Preprint) :1–12, 2019.

[46] Natalia Tomashenko, Antoine Caubrière, and Yannick Estève. Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech. In *Interspeech*, 2019.

[47] Juan-Manuel Torres-Moreno. Artex is another text summarizer. *arXiv preprint arXiv :1210.3312*, 2012.

[48] Juan-Manuel Torres-Moreno. *Automatic Text Summarization*, volume 1. John Wiley & Sons, 2014.

[49] Juan-Manuel Torres-Moreno, Horacio Saggion, Iria da Cunha, Eric SanJuan, and Patricia Velazquez-Morales. Summary Evaluation With and Without References. *Polibits : Research Journal on Computer Science and Computer Engineering with Applications*, 42 :13–19, 2010.

[50] Juan-Manuel Torres-Moreno, Patricia Velázquez-Morales, and Jean-Guy Meunier.

Cortex : un algorithme pour la condensation automatique de textes. *ARCo*, 2 :365, 2001.

[51] Assaf Urieli. *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, 2013. Thèse de doctorat dirigée par Tanguy, Ludovic Sciences du langage Toulouse 2 2013.

AfIA
Association française
pour l'Intelligence Artificielle

LIFAT : Laboratoire d'Informatique Fondamentale Appliquée de Tours

LIFAT EA 6300
Université de Tours
https://tln.lifat.univ-tours.fr/

Jean-Yves ANTOINE
jean-yves.antoine@univ-tours.fr

Nathalie FRIBURGER
nathalie.friburger@univ-tours.fr

Denis MAUREL
denis.maurel@univ-tours.fr

Associée
Agata SAVARY
agata.savary@universite-paris-saclay.fr

Introduction

Le traitement automatique des langues attire une attention de plus en plus marquée des scientifiques et de l'industrie depuis que les *big data* ont commencé à intégrer la fouille de texte orientée web dans leurs applications. Cette évolution nous permet ainsi de renforcer les collaborations internes avec les autres axes de recherche relevant de la fouille de données, de la recherche d'information et du web sémantique.

Une des évolutions marquantes de nos recherches a été celle du développement de plus en plus marqués de traitements utiles à l'extraction d'information et la recherche d'information dans les documents numériques dont la modalité peut aussi bien être l'écrit que l'oral (parole transcrite). Dans ce cadre, on peut dire que nos activités sont centrées autour des traitements et ressources mobilisées par le passage du MOT (langue), telle qu'il apparaît dans le document, au CONCEPT, c'est-à-dire à un niveau sémantique qui permet de faire le lien avec l'ingénierie des connaissances (web sémantique). Nous nous intéressons ainsi à différents niveaux de traitements sur lesquels notre équipe a atteint une visibilité réelle et qui constituent des barrières essentielles dans la mise en œuvre d'une recherche d'information précise et robuste.

Détection en documents de mots d'intérêt : reconnaissance des entités nommées, reconnaissance et analyse des entités polylexicales

Une première étape dans un processus de recherche d'information efficace est la détection intelligente des entités lexicales porteuses de sens, parmi lesquels les entités nommées sont un élément essentiel. La reconnaissance des entités nommées est une thématique fédératrice puisqu'elle a réuni autour du système à base de connaissance (cascades de transducteurs) CasEN la plupart des membres du groupe. Elle a également renforcé les collaborations internes avec les collègues travaillant sur la fouille de données dans le cadre d'un doctorat portant sur l'adaptation de techniques de recherche de motifs hiérarchiques de détection à cette problématique (système mXs). La pertinence de ces recherches a été démontrée dans le cadre de la campagne francophone d'évaluation ETAPE, où CasEN et mXs ont été bien classés. Ce domaine d'excellence de l'équipe sera renforcé au cours du prochain contrat dans le cadre de nos travaux sur l'identification et le parsage des entités polylexicales dans le cadre du projet ANR PARSEME_FR. Les mots d'intérêt pour la RI, parmi lesquels les entités nommées, sont très fréquemment des entités polylexicales qui posent des problèmes d'identification aux techniques de TAL. Nos recherches multilingues initiales sur l'analyse morphologique de telles unités (MultiFlex) ont été étendues à la question de leur analyse syntaxique et surtout à celle de leur

prise en compte précoce dans les analyseurs syntaxiques, dans le cadre d'un réseau européen COST (PARSEME) dont notre équipe a été un des pilotes. Cette activité de réseautage à forte visibilité se poursuit dans le cadre du projet PARSEME_FR.

Mise en relations des mots d'intérêt en documents : résolution de la coréférence, identification de relations temporelles et construction de prédicats

Une barrière scientifique importante pour la recherche d'information est de dépasser une simple approche par sacs de mots pour atteindre une réelle compréhension du document traité. La caractérisation des relations entre mots d'intérêts est une tâche essentielle de ce point de vue. Nous avons développé une réelle expertise dans le domaine de la résolution des coréférences (qui revient à regrouper tous les mots d'un ou plusieurs documents relevant de la même référence) et avons initié une recherche originale dans le domaine de l'analyse des relations temporelles présentes dans un document. Sur le domaine de la coréférence, le projet collaboratif ANCOR réalisé avec le LLL nous a permis d'atteindre une forte visibilité en permettant la réalisation du plus grand corpus mondial de parole spontané annoté en coréférence. Ces recherches se sont depuis poursuivies avec le laboratoire LATTICE (ENS Montrouge) sur la réalisation d'un des premiers systèmes francophones de résolution de la coréférence, entraîné sur le corpus ANCOR. Ce travail se poursuit désormais dans le cadre du projet ANR TALAD où la reprise par coréférence est utilisée pour appuyer l'étude des nominations en analyse du discours. TALAD nous conduit à nous focaliser sur la coréférence indirecte (reprise par une tête lexicale différente), question ignorée par la communauté car portant sur des phénomènes assez peu fréquents (un dixieme de toutes les reprises anaphoriques en moyenne). Il s'agit pourtant d'une problématique importante d'un point de vue applicatif et sans clairement la plus délicate dans le domaine.

Par ailleurs, nous avons initié avec les projets TEMPORAL puis ODIL des travaux sur la réalisation de ce qui devrait être à terme le plus grand corpus francophone annoté en relations temporelles.

Nous venons d'autre part de débuter le projet ANR Abliss en collaboration avec des biologistes afin de fouiller de grandes collections d'articles scientifiques du domaine de la biologie systémique et d'en extraire, sous forme de prédicats, les résultats des expériences décrites.

Pour conclure

Nous avons insisté sur la visibilité de nos travaux, qui est renforcée également par leur caractère résolument multilingue. Il s'agit ici d'une caractéristique forte de notre démarche, l'objectif n'étant pas simplement d'appliquer nos travaux sur différents langages, mais de les confronter pour atteindre un niveau de modélisation et de compréhension linguistique plus profond. C'est la raison pour laquelle nous sommes amenés à envisager des classes de langues variées, suivant les applications (français, anglais, polonais, serbe, arabe, allemand). Notre équipe a développé de ce point de vue une compétence multilingue rare qui nous permet de développer des modèles de traitement d'une grande généricité idiomatique.

Enfin, une dernière caractéristique des travaux menés au sein de notre groupe réside dans nos efforts constants de combiner développement de modules de traitement mais également de ressources linguistiques (corpus, lexiques) qui sont mises à la disposition de la communauté (licences LPGL ou Creative Commons). Depuis le dictionnaire de noms propres multilingue ProlexBase en passant par le corpus en coréférence ANCOR, notre expertise en matière de production est désormais largement reconnue. Il nous amène également à intervenir sur la question de la standardisation (LMF, ISO TimeML). Deux des membres de l'équipe sont ainsi experts du comité AFNOR X03A de normalisation des ressources linguistiques, relai du groupe TC37/SC4 de l'ISO.

En matière de traitements, notons enfin que nous avons développé des approches relevant aussi bien du paradigme centré connaissance que centré sur les données, et que nous comptons poursuivre dans cette voie. Dans le premier cas, les approches symboliques favorisent la comparaison multilingue citée ci-avant et s'appuient sur des partenariats régionaux structurants avec les collègues linguistes du LLL, mais aussi par les chercheurs en TAL du labo-

ratoire LIFO d'Orléans. Les travaux reposants sur des approches centrées données permettent de leur côté la mise en place de collaborations internes avec les collègues de l'équipe travaillant sur des questions de classification et plus généralement d'apprentissage automatique. Suivant les applications, un paradigme privilégié sera choisi, mais nous pouvons observer également que cette double compétence développée au sein de l'axe nous permet également d'envisager des solutions d'hybridations, ou de comparaison fructueuses.

Enfin, ces recherches sont et seront dirigées vers trois champs d'applications transverses qui orientent ces activités en matière d'expression des besoins. Outre l'extraction et la recherche d'informations déjà citées, il s'agit d'une part des humanités numériques (projets Renom et Biblimos par exemple), thématiques en émergence qui répond aux collaborations déjà actives en linguistique et en ingénierie des connaissances appliquées aux textes patrimoniaux, et d'autre part du domaine de l'aide au handicap sur lequel nous avons développé une expérience de plus de vingt années de recherche dans le domaine des systèmes de communication augmentée (système Sibylle). Cette problématique s'est accrue récemment d'une réflexion éthique sur l'impact des technologies numériques qui a donné lieu à la mise en place d'un Réseau Thématiques Régional (RTR Risque) auquel notre équipe et le LIFO INSA Bourges compte donner une dimension.

Références

Toutes nos publications sont sur la plateforme Hal. Nos projets et nos ressources sont décrits sur le web.

AfIA
Association française
pour l'Intelligence Artificielle

LINAGORA Labs

LINAGORA Labs
https://research.linagora.com

Julie HUNTER
jhunter@linagora.com

Jean-Pierre LORRÉ
jplorre@linagora.com

Ilyes REBAI
irebai@linagora.com

Kate THOMPSON
cthompson@linagora.com

Contexte de recherche

LINAGORA Labs est la branche recherche et développement de LINAGORA, une PME française spécialisée dans l'édition des logiciels libres et de plates-formes pour la collaboration. Nos solutions industrielles sont spécifiquement adaptées aux environnements professionnels, allant de notre plate-forme d'assistant LinTO, qui offre une gamme diversifiée de solutions pour développer des applications à commande vocale, à des plates-formes collaboratives qui proposent non seulement des services telles que courrier électronique et calendrier partagé, mais également d'autres fonctionnalités nécessaires à une collaboration fructueuse, telles que le chat et l'édition collaborative temps réel de documents.

La recherche au sein de LINAGORA Labs s'articule suivant deux axes principaux, d'une part l'amélioration de la reconnaissance et de l'interaction vocale et d'autre part le développement de modèles de compréhension des conversations qui peuvent être utilisés pour synthétiser les interactions orales ou écrites entre les individus ou pour extraire des informations pour des tâches telles que la recommandation en ligne. Dans tous les cas, nos recherches suivent la philosophie du LINAGORA, qui consiste à créer des produits open-source qui préservent la vie privée.

Reconnaissance automatique de la parole

Nos recherches en matière de reconnaissance vocale peuvent être divisées en deux grandes catégories : recherche de modèles dédiées aux interactions dites "de commande" et de modèles pour les tâches nécessitant l'appréhension d'un vocabulaire étendu (large vocabulaire).

Les modèles de commande font l'hypothèse d'un monde fermé et fonctionnent avec un vocabulaire limité. En effet leur objectif est de pouvoir reconnaître les intentions à partir d'un ensemble prédéfini de possibilités (par exemple, "Allumez les lumières" ou "Quel temps fait-il ?") et des entités optionnelles ("dans le salon", "à Toulouse", etc.). Ils doivent néanmoins relever certains défis, en particulier la prise en compte du bruit de fond potentiel dans l'environnement de l'utilisateur ainsi que la nécessité d'être adaptés à des contextes métiers spécifiques, car différentes entreprises peuvent nécessiter des intentions différentes et des dictionnaires différents pour les entités. Notre approche permet d'enrichir incrémentalement le modèle de langage en fonction des scénarios d'usages envisagés. Ceci présente l'avantage de générer des modèles de petite taille particulièrement performants. Dans un contexte respectueux de la vie privée des utilisateurs, pour lequel les données d'un utilisateur doivent rester confidentielles, la mise en oeuvre de modèles spécialisés peut être délicate. De ce fait, une autre ambition concerne également la possibilité de décoder de tels modèles dans un dispositif dédié à l'utilisateur afin que le signal vocal reste au plus près de son émetteur.

Nos modèles à large vocabulaire s'adressent aux mondes ouverts et sont disponibles soient en mode streaming (décodage au fils de l'eau), soit en mode hors ligne (à posteriori). Ils se concentrent sur la

tâche de transcription de conversations complètes, et en particulier des conversations multipartites et spontanées du type de celles que l'on rencontre lors des réunions d'affaires. Ces interactions posent un certain nombre d'obstacles aux systèmes de reconnaissance vocale de pointe, notamment des conditions d'enregistrement difficiles car les locuteurs peuvent être mal placés par rapport aux microphones, ils peuvent présenter un niveau élevé de disfluences, de chevauchement des paroles et d'autres types de distorsions de la conversation propres à la conversation spontanée. Tout cela s'avère difficile pour les modèles linguistiques formés à la parole grammaticalement correcte sans hésitations.

Le mode streaming nous permet de transcrire des réunions en temps réel pour faire le sous-titrage ou la traduction automatique lors d'une intervention. Il est également à la base des recommandations en temps réel. Le décodage offline produit un résultat plus précis parce qu'il s'appuie sur la connaissance de l'ensemble du message à transcrire. Il permet en particulier l'utilisation des i-vecteurs pour distinguer des traits phonétiques des locuteurs, pour faire des prédictions spécifiques et plus précises.

Notre démarche se base sur un modèle de reconnaissance de la parole hybride DNN-HMM combinant modèle acoustique (DNN, Deep Neural Network) et modèle de langage (HMM, Hidden Markov Model). L'architecture neuronale du modèle acoustique combine un réseau TDNN (Time Delay Neural Network) avec un mécanisme d'attention [10]. Cette approche permet d'obtenir des résultats intéressants notamment dans le cas ou des corpus de très grandes tailles ne sont pas disponibles ; nous l'implémentons grâce à la boite à outils Kaldi. Nous étudions en outre de nouvelles techniques issues des méthodologies telles que l'apprentissage multitâche et les modèles dis "end-to-end" [5].

Deux autres enjeux viennent compliquer nos travaux sur l'ASR. D'abord le besoin de travailler sur des langues européennes et en particulier le français. Or, la majorité des ressources disponibles sont en anglais. Pour cela, nous avons initié un projet ANR, SUMM-RE, pour créer un corpus de réunions en français. Un autre problème est le respect des données personnelles qui fait que le partage de données, qui aiderait à constituer des corpus d'une taille suffisante pour les modèles avancés de l'apprentissage automatique que l'on utilise pour la reconnaissance de la parole, est impossible. Pour cela, nous avons initié des travaux en collaboration avec l'association Le Voice Lab en vue d'explorer des approches telles que l'apprentissage fédéré.

Analyse du discours

La possibilité de transcrire de manière fiable une conversation orale ouvre la possibilité d'exploiter ces données pour des tâches qui exigent une compréhension linguistique plus avancée que la simple reconnaissance de commandes. Cela inclut les interactions naturelles en temps réel avec des agents artificiels ainsi que des tâches d'extraction d'informations a posteriori, comme la synthèse automatique et le compte rendu automatique de réunions.

De nombreuses approches de pointe en matière de résumé s'appuient fortement sur les mots utilisés tout au long d'un discours ou d'une conversation. Une grande partie de nos travaux sur le résumé se concentre quant à eux sur la construction de meilleures représentations de l'importance lexicale et de la similarité dans le discours afin d'identifier les énoncés les plus importants à résumer [11, 12].

Toutefois, pour rédiger des résumés de conversations détaillées ou d'interactions conversationnelles complexes, il ne suffit pas d'identifier les principaux sujets de discussion ou les mots clés qui pourraient aider à catégoriser un énoncé. Il faut aussi être capable de suivre le fil d'une conversation. Il faut pour cela reconnaître le rôle que chaque énoncé joue dans une interaction donnée : un énoncé sert-il à répondre à une question qui a été posée, à expliquer quelque chose qui a été dit, à corriger ou à contraster avec un argument qui a été avancé ? Un objectif central de notre équipe est de construire des modèles de dialogue qui nous permettent d'exploiter les relations structurelles et sémantiques qui existent entre les contenus des différents énoncés pour extraire automatiquement des informations plus détaillées des transcriptions de conversations.

L'identification de ces relations implique d'abord de segmenter un discours à peu près au niveau des propositions, c'est-à-dire des contenus pour des actes de dialogue individuels qui peuvent four-

nir des arguments aux relations discursives, comme une relation Question-Réponse, ou une Explication, un Contraste ou une Correction. Comme les algorithmes de pointe pour la segmentation du discours sont issus de texte ou de transcriptions de discours préparés [8], cela signifie qu'il faut trouver des moyens d'adapter les modèles actuels aux conversations spontanées et multipartites, qui contiennent des disfluences et des constructions syntaxiques ou lexicales particulières à un langage parlé moins formel.

Pour ce faire, nous travaillons, avec des partenaires de l'IRIT, dans le cadre d'un paradigme de supervision faible basé sur la programmation par les données (data programming) [9] dans lequel des annotateurs experts étudient un échantillon de données petit mais représentatif pour écrire des règles d'annotation qui peuvent être ensuite utilisées pour annoter automatiquement de grands ensembles de données. Cette approche nous permet également d'intégrer facilement des informations acoustiques qui peuvent être utiles pour indiquer les limites des segments ou des actes de dialogue. De tels indices acoustiques fournissent une aide précieuse lorsque l'expression est pleine de disfluences ou que les transcriptions sont imparfaites, ce qui implique que les indices basés sur le texte sont plus faibles.

Après la segmentation, la tâche suivante consiste à construire des structures de discours qui indiquent comment les contenus des différents actes de dialogue sont liés les uns aux autres [2, 7]. Pour cette tâche aussi, nous expérimentons la programmation par les données [3, 4] pour faire face au manque de données discursives annotées nécessaires aux algorithmes d'apprentissage automatique.

Un dernier aspect de notre travail sur la modélisation et la compréhension du dialogue parlé concerne la multimodalité des conversations en face à face, voire des vidéoconférences. Les gestes ou autres mouvements significatifs, ainsi que les objets et actions visibles dans le contexte, peuvent être sémantiquement pertinents. Comprendre comment le contexte non linguistique ajoute du contenu à une conversation et, inversement, comment le contenu d'une conversation peut nous guider vers des interprétations pertinentes de ce qui se passe dans la scène visuelle est crucial pour obtenir un modèle complet d'une conversation [1, 6]. La compréhension de ces interactions est donc nécessaire pour des interactions fructueuses entre les humains et les assistants ou les agents incorporés.

Références

[1] Nicholas Asher, Julie Hunter, and Kate Thompson. Comparing discourse structures between purely linguistic and situated messages in an annotated corpus. *Dialogue & Discourse*, 11(1) :89–121, 2020.

[2] Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

[3] Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. Data programming for learning discourse structure. In *Association for Computational Linguistics (ACL)*, 2019.

[4] Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. Weak supervision for learning discourse structure. In *Empirical Methods in Natural Language Processing*. 2019.

[5] Abdelwahab Heba, Thomas Pellegrini, Jean-Pierre Lorré, and Régine Andre-Obrecht. Char+CV-CTC : Combining Graphemes and Consonant/Vowel Units for CTC-Based ASR Using Multitask Learning. In *Interspeech*, pages 1611–1615, 2019.

[6] Julie Hunter, Nicholas Asher, and Alex Lascarides. A formal semantics for situated conversation. *Semantics and Pragmatics*, 11, 2018.

[7] William Mann and Sandra Thompson. Rhetorical structure theory : A framework for the analysis of texts. *International Pragmatics Association Papers in Pragmatics*, 1 :79–105, 1987.

[8] Philippe Muller, Chloé Braud, and Mathieu Morey. ToNy : Contextual embeddings for accurate multilingual discourse segmentation for full documents. In *Discourse Relation Parsing and Treebanking*, pages 115–124. 2019.

[9] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming : Creating large training sets,

quickly. *Advances in neural information processing systems*, 29 :3567–3575, 2016.

[10] Ilyes Rebai, Kate Thompson, Sami Benhamiche, Zied Sellami, Damien Laine, and Jean-Pierre Lorré. LinTO platform : A smart open voice assistant for business environments. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 89–95, Marseille, France, May 2020. European Language Resources Association.

[11] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine J.P. Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *The 56th Annual Meeting of the Association for Computational Linguistics*. 2018.

[12] Guokan Shang, Antoine Jean-Pierre Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. Energy-based self-attentive learning of abstractive communities for spoken language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020.

AfIA
Association française
pour l'Intelligence Artificielle

LISN : Laboratoire Interdisciplinaire des Sciences du Numérique

LISN UMR 9015
CNRS, Université Paris-Saclay
https://www.lisn.upsaclay.fr

https://www.limsi.fr/fr/recherche/iles

https://www.limsi.fr/fr/recherche/tlp

Gilles ADDA
Responsable du département STL
adda@lisn.upsaclay.fr

Annelies BRAFFORT
Responsable de l'équipe ILES
annelies.braffort@lisn.upsaclay.fr

Jean-Luc GAUVAIN
Responsable de l'équipe TLP
gauvain@lisn.upsaclay.fr

Membres au 1/1/2021

Permanents (ILES)

- Annelies BRAFFORT (CNRS)
- Michael FILHOL (CNRS)
- Sahar GHANNAY (Université Paris Saclay)
- Cyril GROUIN (CNRS)
- Thierry HAMON (Université Paris-Nord)
- Gabriel ILLOUZ (Université Paris Saclay)
- Thomas LAVERGNE (Université Paris Saclay)
- Anne-Laure LIGOZAT (ENSIIE)
- Aurélie NÉVÉOL (CNRS)
- Patrick PAROUBEK (CNRS)
- Sophie ROSSET (CNRS)
- Anne VILNAT (Université Paris Saclay)
- Pierre ZWEIGENBAUM (CNRS)

Permanents (TLP)

- Gilles ADDA (CNRS)
- Philippe BOULA DE MAREÜIL (CNRS)
- Caio CORRO (Université Paris-Saclay)
- Marc EVRARD (Université Paris-Saclay)
- Laurence DEVILLERS (Univ. Paris-Sorbonne)
- Kim GERDES (Université Paris-Saclay)
- Jean-Luc GAUVAIN (CNRS)
- Camille GUINAUDEAU (Université Paris-Saclay)
- Lori LAMEL (CNRS)
- Jean-Sylvain LIÉNARD (CNRS)
- Joseph-Jean MARIANI (CNRS)
- Hélène MAYNARD (Université Paris-Saclay)
- Albert RILLIARD (CNRS)
- Ioana VASILESCU (CNRS)
- François YVON (CNRS)

Doctorant·e·s, Université Paris-Saclay, *ED STIC* : ILES : Nesrine BANNOUR, Valentin BELISSEN, Alexandra BENAMAR, Félix BIGAND, Hugo BOULANGER, Hannah BULL, Oralie CATTAN, Hussein CHAABAN, Juan Manuel CORIA, Hicham EL BOUKKOURI, Léo GALMANT, Marion KACZMAREK, Corentin MASSON, Tsanta RANDRIATSITOHAINA, Lisa RAITHEL, Léon-Paul SCHAUB, Mathilde VERON, TLP : Hugues Ali MEHENNI, Marc BENZAHRA, Aman Zaid BERHE, François BUET, Théo DESCHAMPS-BERGER, Caroline ÉTIENNE, Aina GARI SOLER, Natalia KALASHNIKOVA, Margot LACOUR, Paul LERNER, Saulo MENDES SANTOS, Anh Khoa NGO HO, Alban PETIT, Minh Quang PHAM, José Carlos ROSALES, Rémi URO, Jitao XU, Yajing FENG.

Contractuel·le·s et post-doctorant·e·s : Claire DANET, Meghan DOWLING, Benjamin ELIE, Lucie GIANOLA, Mathilde HUTIN, Lucas ONDEL, Sadaf Abdul RAUF, Yaru WU.

Introduction

Le département Sciences et Technologies de la Langue (STL) regroupe les activités du LISN (ex-LIMSI) en *traitement de la langue* au sein des équipes ILES et TLP. Le spectre des recherches conduit dans ces deux équipes est très large et s'étend à plusieurs modalités de la langue, *orale, écrite ou signée*. Certaines des recherches autour de la visualisation sont faites en collaboration avec les équipes du département Interaction du laboratoire et donnent lieu à des co-encadrements de thèse.

L'équipe ILES (Information, Langue Écrite et Signée) se centre sur l'étude de la langue écrite ou signée, aussi bien dans ses fonctions de communica-

tion ou de support d'information, motivant l'étude de méthodes pour extraire et rechercher des informations précises dans des documents variés, articles de presse, publications scientifiques ou dossiers médicaux, ou pour dialoguer avec un locuteur humain.

La communication parlée constitue le noyau des recherches développées dans l'équipe TLP (Traitement du Langage Parlé), avec des activités qui se déploient depuis le traitement du signal jusqu'à la structure narrative, en passant par toutes les étapes d'analyse et d'enrichissement automatique de l'entrée vocale : idenfication de la langue et des locuteurs, transcription de la parole, reconnaissance des émotions, traduction de la parole.

Thèmes de recherche

Reconnaissance de la parole (TLP)

La reconnaissance vocale consiste à convertir la forme d'onde de la parole, un signal acoustique, en une séquence de mots. Aujourd'hui, les approches les plus performantes sont fondées sur une modélisation statistique du signal vocal. Nos recherches portent sur les principaux problèmes de la reconnaissance de la parole : modélisation du langage, représentation lexicale, modélisation acoustique-phonétique et décodage. La réalisation de chaque mot dépend fortement du locuteur, du contexte social et de l'environnement acoustique (cf. thème « Perception et traitement automatique de la variation dans la parole »). Les systèmes automatiques de conversion parole-texte doivent être capables de gérer de tels effets contextuels variant dans le temps et d'évoluer pour gérer les changements de style et de sujet, en adaptant leur vocabulaire. La recherche sur la reconnaissance de la parole est menée dans un contexte multilingue, en étudiant et en développant des modèles pour une multitude de langues et de variantes dialectales. En collaboration avec le thème « Perception et traitement automatique de la variation dans la parole », des études linguistiques sur corpus sont réalisées pour quantifier et découvrir les tendances linguistiques, et les erreurs du système sont étudiées pour identifier les faiblesses technologiques potentielles et sont comparées aux performances humaines de référence.

Modélisation et traitement automatique des langues des signes (ILES)

Les langues des signes (LS) sont des langues naturelles pratiquées au sein des communautés de sourds et la Langue des Signes Française (LSF) est celle utilisée en France. Ce sont des langues visuo-gestuelles : une personne s'exprime en LS en utilisant de nombreuses composantes corporelles (les mains et les bras, mais aussi les expressions du visage, le regard, le buste, *etc.*) et son interlocuteur perçoit le message par le canal visuel. Le système linguistique des LS exploite ces canaux spécifiques : de nombreuses informations sont exprimées simultanément et s'organisent dans l'espace, et l'iconicité joue un rôle central. Les LS sont encore peu décrites, peu dotées et ne disposent pas d'outillage dédié. Les recherches sur ces langues sont récentes en linguistique et en sont encore aux balbutiements en traitement automatique. La communauté scientifique étudiant les LS est réduite.

Nous avons réalisé des travaux concernant l'étude et la modélisation de la LSF, en nous plaçant dans une approche résolument pluridisciplinaire impliquant la linguistique, les sciences du mouvement, la psychologie et l'informatique. Un objectif est d'appuyer les modèles ou descriptions formelles que nous élaborons sur les résultats d'études basées sur des analyses statistiques solides. Nos travaux couvrent les axes de recherche suivants : (1) l'étude et la modélisation de la LSF, en linguistique, en sciences du mouvement (en collaboration avec le CIAMS de l'Université Paris Saclay) et en perception visuelle (en collaboration avec l'équipe "Cognition, Perceptions et Usages" du LISN) ; (2) l'élaboration de ressources linguistiques et d'outils permettant de manipuler ces ressources (*i.e.* aide à l'annotation par traitement d'images) ; et (3) les principaux thèmes de recherche en traitement automatique de la LSF : la reconnaissance (en collaboration avec l'équipe "Architectures et Modèles pour l'Interaction"), la génération et la traduction.

Perception et traitement automatique de la variation dans la parole (TLP)

Les activités autour de ce thème ont comme objectif de circonscrire et de modéliser la variation pré-

sente dans la parole, qu'il s'agisse de variation diatopique, diastratique, diaphasique ou diachronique. La méthode adoptée comprend une analyse statistique de grands corpus oraux (utilisant notamment des systèmes de reconnaissance de la parole comme outils d'exploration linguistique) et l'exploitation de la composante perceptive, via des comparaisons humain/machine dans différentes configurations expérimentales. Ces dernières années, nous avons concentré nos efforts autour de deux axes. D'une part, nous avons abordé la variation orale dans des grands corpus multilingues, dans différentes langues et notamment dans les langues romanes (travaux de Ioana VASILESCU). D'autre part, nous avons poursuivi des activités de documentation des accents et langues régionales via l'acquisition de données permettant de cartographier la variation diatopique (en particulier en français). Le fruit de cette seconde activité prend de plus en plus la forme d'atlas dialectologiques des accents et langues régionales de France.

Caractérisation du locuteur dans un contexte multimédia (TLP)

Les activités de ce thème se sont développées principalement selon trois grands axes. Elles concernent premièrement des travaux sur la segmentation et le regroupement en locuteurs dans les documents audio. En particulier, il s'agit de repenser les approches classiquement utilisées pour le traitement des journaux radio- ou télé-diffusés, qui atteignent leurs limites quand elles sont appliquées à d'autres types de contenus (films, séries TV, enregistrements de réunions). Deuxièmement, une composante *multimédia* a émergé avec la tâche « Multimodal Person Discovery in Broadcast TV » que nous avons organisée lors des campagnes d'évaluation MediaEval 2015 et 2016 en lien avec le projet CHIST-ERA/CAMOMILE (2012–2016). Enfin, une nouvelle activité portant sur la structuration sémantique de contenus audio-visuels (films, séries TV) a vu le jour, où la composante « traitement automatique de la langue » prend une place importante. Un axe transverse portant sur la question de l'évaluation des technologies multimédia rapproche ces trois grands axes thématiques.

Dimensions affectives et sociales des interactions parlées avec des (ro)bots et enjeux éthiques (TLP)

Les activités autour de ce thème se concentrent sur trois axes : le premier axe porte sur la robustesse de la détection des émotions à partir d'indices paralinguistiques et linguistiques et l'utilisation de ces systèmes dans les interactions avec des agents conversationnels et des robots sociaux.

Le deuxième axe porte sur l'interaction affective avec des machines et le nudge en utilisant des théories en linguistique sur l'interaction, en sociologie sur les rites sociaux, en psychologie cognitive sur les modèles d'évaluation et la théorie des états mentaux.

Le troisième axe porte sur le besoin de réflexions éthiques autour de la modélisation affective et le pouvoir de manipulation par les machines vocales (chatbot, robots sociaux, objets vocaux connectés) dans la société. Les sujets de recherche principaux sont la perception et l'interprétation des signaux émotionnels et sociaux en contexte dans l'interaction orale avec des chatbots ou des robots sociaux.

La chaire hors 3IA HUMAAINE : HUman-MAchine Affective Interaction Ethics portée par Laurence DEVILLERS s'interesse sur le nudge numérique des agents conversationnels affectifs. Ce sujet est également expliqué dans le rapport demandé par le premier ministre sur les chatbots et l'éthique (mission CNPEN).

Multilinguisme et paraphrase (ILES)

L'un des problèmes auxquels s'attaque le traitement automatique des langues est l'existence d'énoncés distincts dont le sens est proche voire équivalent : synonyme d'un terme, paraphrase, version simplifiée ou traduction d'une phrase, phrase qui en implique une autre, *etc.* Ces questions sont au cœur de la sémantique. Le présent thème s'attaque aux problématiques qui en dérivent : l'identification de la relation qui existe entre deux tels énoncés, ou inversement la production d'un énoncé cible étant donné un énoncé source et une relation (par ex. traduction, simplification). Cette dernière problématique s'étend au cas du transfert de systèmes de TAL mis au point pour une variété de

langue à une autre variété de langue, par exemple leur portage à une autre langue.

Les travaux menés dans ce thème s'articulent ainsi autour des trois problématiques suivantes : (1) similarité sémantique et implication textuelle ; (2) production de paraphrases, notamment de simplifications ; (3) traduction et alignement ; et (4) adaptation de systèmes à une autre langue. Ce thème interagit de façon transverse avec chacun des trois autres thèmes du groupe ILES, ainsi qu'avec l'activité de traduction du groupe TLP.

Traduction, apprentissage automatique (TLP)

L'activité de ce thème recouvre un large spectre de thématiques relatives à l'apprentissage automatique en traitement automatique des langues, avec une focalisation particulière sur les tâches d'apprentissage structuré, et comme terrain d'application principal la traduction automatique (TA) de texte et de parole. L'amélioration des systèmes et des modèles de traduction automatique est restée au cœur de nos préoccupations et nous avons continué de contribuer activement au développement d'architectures computationnelles pour la TA (au sens large), en explorant deux directions : d'une part, l'étude de systèmes plus interactifs et plus réactifs ; d'autre part en poursuivant nos travaux sur les architectures neuronales pour la TA, qui ont, au cours de la période, radicalement transformé l'état de l'art en TA statistique et éliminé du paysage les méthodes antérieures (TA à base de segments). Ces études s'étendent également aux problèmes d'alignement, avec des applications à l'apprentissage cross-lingue par transfert ou encore à la documentation semi-automatique de données collectées par des linguistes de terrain, ainsi qu'à la traduction monolingue (correction, transfert de style, simplification). Les travaux en apprentissage automatique se développent dans deux directions principales. Elles s'intéressent d'une part à l'études de méthodes génériques pour aborder des tâches complètement ou partiellement supervisées d'apprentissage structuré « de bas niveau » (normalisation, étiquetage en parties du discours ou en *chunks*, segmentation en unités sous-lexicales (morphologiques ou non), *parsing*), avec comme ambition de développer des techniques, par exemple en matière d'adaptation au domaine ou d'intégration de ressources lexicales, qui pourront ensuite être transférées à des tâches de TA. Plus récemment, nous avons également étudié des tâches plus complexes comme l'analyse sémantique profonde.

Elles explorent d'autre part les questions relatives au traitement du multilinguisme, que ce soit du point de vue de l'apprentissage de représentations lexicales, de la représentation des corpus multilingues non-annotés et annotés (par exemple dans une perspective de comparaison cross-lingues). Ces activités impliquent des collaborations resserrées avec le thème « Multilinguisme et paraphase » (groupe ILES), ainsi qu'avec le thème « Reconnaissance de la parole » pour ce qui concerne la traduction de parole et le thème « Caractérisation du locuteur dans un contexte multimédia » qui aborde des problèmes formellement proches.

Extraction et reconnaissance d'informations précises, dialogue (ILES)

Devant la production massive de documents sous forme numérique, sur l'Internet, dans des entreprises ou hôpitaux, il est essentiel de disposer d'outils d'analyse automatique afin de pouvoir extraire, représenter ou accéder aux informations qu'ils contiennent. Autrement dit, comment transformer une information exprimée en langage naturel, donc sous forme non structurée, en une connaissance structurée, manipulable par une machine, et par quel modèle d'analyse ?

Les analyses que nous proposons visent à (1) produire des représentations sémantiques d'énoncés et de documents, (2) extraire et stocker des informations dans une base de connaissances, (3) restituer une information à un utilisateur en fonction d'un besoin qu'il exprime, par l'analyse d'un texte ou l'interrogation de bases de connaissances (les données liées du web sémantique) ou (4) gérer un dialogue en langue naturelle avec un locuteur. Elles contribuent aux tâches de détection d'entités et de relations, de catégorisation de textes, de liaison référentielle et peuplement de bases de connaissances, et de recherche de réponses précises à des questions en langage naturel. Nos travaux portent une diversité de textes incluant une langue éditée (par exemple, des articles scien-

tifiques) ou une langue non standard (par exemple, dans les productions écrites issues des réseaux sociaux).

Par ailleurs, les « chatterbots » connaissent un fort développement ces derniers temps. Nos recherches visent à élaborer des systèmes de dialogue autorisant une interaction naturelle avec un utilisateur, en le laissant libre d'utiliser sa langue, et qui soient capable de retrouver l'information cherchée quelle qu'en soit la représentation. Il s'agit alors de modéliser le processus d'interaction lui-même afin de développer un dialogue naturel.

Ressources langagières, corpus et représentations (ILES, TLP)

L'évaluation comparative est un élément moteur du traitement de la parole depuis plus de 30 ans. Les corpus sont au cœur de ces deux grands paradigmes. Alors que dans le passé, l'utilisation des grands corpus s'est limitée à quelques domaines et langues, la dernière décennie a connu une vraie expansion vers le multilinguisme et la multimodalité. Le développement de corpus et l'organisation de campagnes d'évaluations (DEFT, CLEF eHealth 2015-2018, Covid-19 @ MLIA, etc.) sont cruciaux pour la communauté linguistique et posent à leur tour des problèmes scientifiques qui doivent être résolus, tels que les corpus à collecter et comment ils devraient être annotés, ainsi que des questions scientifiques sur la façon de récompenser leurs promoteurs et la façon d'assurer l'éthique dans le processus de collecte. Ce thème traite de l'aspect théorique et des problèmes pratiques concernant la collecte, l'annotation et la diffusion de grands corpus multilingues.

Une activité importante de ce thème est la proposition de représentations des énoncés langagiers écrits ou signés et la production de corpus les instanciant. Définir la représentation requise par un traitement automatique du langage (par exemple la reconnaissance d'entités nommées, la fouille d'opinion ou la génération de texte) est une étape fondamentale dans l'étude de la tâche et de ses fondements linguistiques. La création de corpus annotés avec les représentations associées aux traitements fournit le matériau indispensable au développement et à l'évaluation de systèmes d'analyse, de transformation ou de production du langage.

Références

[1] Bissan Audeh, Cyril Grouin, Pierre Zweigenbaum, Cédric Bousquet, Marie-Christine Jaulent, Mehdi Benkhebil, and Agnès Lillo-Le Louët. French levothyrox® crisis : retrospective analysis of social media. In *International Society of Pharmacovigilance*, 2019.

[2] Valentin Belissen, Annelies Braffort, and Michèle Gouiffès. Experimenting the automatic recognition of non-conventionalized units in sign language. *Algorithms*, 13(12) :310, 2020.

[3] Aman Berthe, Camille Guinaudeau, and Claude Barras. Détection de scènes remarquables dans un contexte de séries TV. In *Conférence en Recherche d'Information et Applications*, 2021.

[4] Félix Bigand, Elise Prigent, and Annelies Braffort. Retrieving Human Traits from Gesture in Sign Language : The Example of Gestural Identity. In *International Symposium on Movement and Computing*, Tempe, United States, 2019.

[5] Georgeta Bordea, Tsanta Randriatsitohaina, Fleur Mougin, Natalia Grabar, and Thierry Hamon. Query selection methods for automated corpora construction with a use case in food-drug interactions. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 115–124, Florence, Italy, August 2019. Association for Computational Linguistics.

[6] Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. A French clinical corpus with comprehensive semantic annotations : development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation*, 2017.

[7] Leonardo Campillos-Llanos, Catherine Thomas, Eric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. Designing a virtual patient dialogue system based on terminology-rich resources : challenges and evaluation. *Natural Language Engineering*, pages 1–38, 2019.

[8] Caio Corro. Span-based discontinuous constituency parsing : a family of exact chart-based algorithms with time complexities from O(n^6) down to O(n^3). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2753–2764, Online, November 2020. Association for Computational Linguistics.

[9] Caio Corro. Sur l'impact des contraintes structurelles pour l'analyse en dépendances profondes fondée sur les graphes. In *6e conférence conjointe JEP-TALN-RECITAL. Volume 2 : Traitement Automatique des Langues Naturelles*, pages 197–204. ATALA ; AFCP, 2020.

[10] Philippe Boula de Mareüil, Frédéric Vernier, Gilles Adda, Albert Rilliard, and Jacques Vernaudon. A speaking atlas of indigenous languages of france and its overseas. In *Language Technologies for All (LT4All) Enabling Language Diversity & Multilingualism Worldwide*, volume 1, pages 155–159. European Language Resources Association (ELRA), 2019.

[11] Laurence Devillers. *Human–Robot Interactions and Affective Computing : The Ethical Implications*, pages 205–211. Springer International Publishing, Cham, 2021.

[12] Laurence Devillers, Françoise Fogelman-Soulié, and Ricardo Baeza-Yates. *AI & Human Values*, pages 76–89. Springer International Publishing, Cham, 2021.

[13] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. CharacterBERT : Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proc. 28th International Conference on Computational Linguistics*, pages 6903–6915, December 2020.

[14] Michael Filhol. A human–editable sign language representation inspired by spontaneous productions... and a writing system ? *Sign Language Studies*, 21(1) :98–136, 2020.

[15] Léo Galmant, Hervé Bredin, Camille Guinaudeau, and Anne-Laure Ligozat. "Hé Manu, tu descends ?" : identification nommée du locuteur dans les dialogues. In *Conférence en Recherche d'Information et Applications*, Lyon, France, 2019.

[16] Kim Gerdes, Sylvain Kahane, and Xinying Chen. Typometrics : From implicational to quantitative universals in word order typology. *Glossa : a journal of general linguistics*, 6(1), 2021.

[17] Sahar Ghannay, Antoine Neuraz, and Sophie Rosset. What is best for spoken language understanding : small but task-dependant embeddings or huge but out-of-domain embeddings ? In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8114–8118. IEEE, 2020.

[18] Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Maynard, Markus Mueller, Annie Rialland, Sebastian Stueker, François Yvon, and Marcely Zanon-Boito. A very low resource language speech corpus for computational language documentation experiments. In *Proc. 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018.

[19] Julia Ive, Aurélien Max, and François Yvon. Reassessing the proper place of man and machine in translation : a pre-translation scenario. *Machine Translation*, 32(4) :31p, 2018.

[20] Elena Knyazeva, Guillaume Wisniewski, and François Yvon. Les méthodes " apprendre à chercher " en traitement automatique des langues : un état de l'art. *Traitement Automatique des Langues (TAL)*, 59(1) :39–63, 2018.

[21] Jean-Sylvain Liénard. Quantifying vocal effort from the shape of the one-third octave long-term-average spectrum of speech. *The Journal of the Acoustical Society of America*, 146(4) :EL369–EL375, 2019.

[22] Rasa Lileikyté, Lori Lamel, Jean-Luc Gauvain, and Arseniy Gorin. Conversational telephone speech recognition for Lithuanian. *Computer Speech and Language*, 49 :71–82, 2018.

[23] Joseph J Mariani, Gil Francopoulo, and Patrick Paroubek. Reuse and Plagiarism in Speech and

Natural Language Processing. *International Journal on Digital Libraries*, 18 :1–14, 2017.

[24] Fabio Martínez, Antoine Manzanera, Michèle Gouiffès, and Annelies Braffort. A Gaussian mixture representation of gesture kinematics for on-line Sign Language video annotation. In *International Symposium on Visual Computing ISVC'15*, Las Vegas, United States, 2015.

[25] Hugues Ali Mehenni, Sofiya Kobylyanskaya, Ioana Vasilescu, and Laurence Devillers. Nudges with conversational agents and social robots : A first experiment with children at a primary school. In *Conversational Dialogue Systems for the Next Decade*, pages 257–270. Springer, 2021.

[26] Luma Miranda, Marc Swerts, João Moraes, and Albert Rilliard. The role of the auditory and visual modalities in the perceptual identification of Brazilian Portuguese statements and echo questions. *Language and speech*, page 0023830919898886, 2020.

[27] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. Clinical natural language processing in languages other than English : opportunities and challenges. *Journal of biomedical semantics*, 9(1) :1–13, 2018.

[28] Christopher R Norman, Mariska MG Leeflang, Raphaël Porcher, and Aurélie Névéol. Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *Systematic reviews*, 8(1) :1–18, 2019.

[29] Minh Quang Pham, Josep-Maria Crego, and François Yvon. Revisiting Multi-Domain Machine Translation. *Transactions of the Association for Computational Linguistics*, 9 :17–35, 2021.

[30] Albert Rilliard, Christophe d'Alessandro, and Marc Evrard. Paradigmatic variation of vowels in expressive speech : Acoustic description and dimensional analysis. *The Journal of the Acoustical Society of America*, 143(1) :109–122, 2018.

[31] Ioana Vasilescu, Ioana Chitoran, Bianca Dimulescu-Vieru, Martine Adda-Decker, Lori Lamel, Oana Niculescu, and P Langlais. Studying variation in Romanian : deletion of the definite article -l in continuous speech. *Linguistic Vanguard*, 5(1) :17p, 2018.

[32] Ioana Vasilescu and Lori Lamel. Synchronic variation and sound change in romance languages : a corpus-based study of lenition phenomena in romanian and spanish. *Journal of the Acoustical Society of America*, 130(6) :3980–3991.

[33] Mathilde Veron, Sahar Ghannay, Anne-Laure Ligozat, and Sophie Rosset. Lifelong learning and task-oriented dialogue system : what does it mean ? In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction : 10th International Workshop on Spoken Dialogue Systems*, pages 347–356. Springer Singapore, 2021.

[34] Yuming Zhai, Gabriel Illouz, and Anne Vilnat. Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. In *Proc. 28th International Conference on Computational Linguistics*, pages 5944–5956, 2020.

AfIA
Association française
pour l'Intelligence Artificielle

TETIS/MISCA : Modélisation Information Spatiale, Extraction de Connaissances et Analyse

TETIS / MISCA
AgroParisTech, Cirad, Cnrs, Inrae
https://umr-tetis.fr

Mathieu ROCHE
mathieu.roche@cirad.fr

Pascal DEGENNE
pascal.degenne@cirad.fr

Membres de l'équipe liés aux problématiques TLH (Technologies du Langage Humain) :

- Rémy Decoupes (INRAE)
- Hugo Deleglise (doctorant)
- Roberto Interdonato (CIRAD)
- Rodrique Kafando (doctorant)
- Urcel Kalenga (doctorant)
- Martin Lentschat (doctorant)
- Mathieu Roche (CIRAD)
- Lucile Sautot (AGROPARISTECH)
- Camille Schaeffer (doctorante)
- Mehtab Alam Syed (doctorant)
- Maguelonne Teisseire (INRAE)

Introduction

Un des objectif de l'équipe MISCA est de développer des méthodes de gestion de l'information permettant de répondre aux grands enjeux sociétaux liés à l'environnement et à l'agriculture, qu'il s'agisse de stocker, de gérer, de partager ou d'analyser de gros volumes de données. Les données et informations, décrites par des caractéristiques spatiales, temporelles et/ou thématiques, sont de surcroît hétérogènes ouvrant de nouvelles problématiques de recherche. Dans ce contexte, des contributions méthodologiques sont proposées et mises en place pour consolider la chaîne de l'information et les processus d'extraction de connaissances, en particulier :

- L'intégration de données (numériques, textuelles, géographiques, issues de la télédétection, etc.) hétérogènes, structurées ou non sous forme de connaissances.
- L'association et le traitement de données et connaissances couvrant plusieurs échelles spatiales et temporelles.
- L'étude de méthodologies hybrides combinant des approches orientées données avec des approches orientées processus.
- La mise en œuvre d'outils et de méthodes pour faciliter la représentation des connaissances et de leur sémantique sous la forme de modèles, permettant des traitements automatiques ou des simulations.

Ces champs de recherche sont mobilisés dans des projets d'application terrain sur une grande diversité de domaines thématiques :

- Appui au diagnostic et à la gestion des territoires,
- Épidémiologie spatiale en santé humaine, animale et végétale,
- Biodiversité, services éco-systémiques,
- Sécurité alimentaire et alerte précoce,
- Impact du changement climatique sur les cultures,
- Hydrologie et gestion de l'eau.

Enfin, les questions scientifiques autour des TLH sont abordées à travers différents projets d'envergure nationale et internationale :

- Projets régionaux (Région Occitanie et MUSE) : *SONGES (Science des dONnées hétéroGènES), MeDO (Mégadonnées, Données liées et fouille de données pour les réseaux d'assainissement), TEXT4LOD (n-ARy relaTions EXTraction for Linked Open Data).*
- Projets nationaux (ANR) : *HERELLES (Hétérogénéité des données - Hétérogénéité des méthodes), BEYOND (Building epidemiological surveillance and prophylaxis with observations both near and distant).*
- Projets européens (H2020) : *MOOD (MOnitoring Outbreak events for Disease surveillance in a data science context), LEAP4FNSSA (Long-*

term Europe-Africa Research and Innovation Partnership for Food and Nutrition Security and Sustainable Agriculture).

Activités de MISCA et Technologies du Langage Humain

La multitude et la variété des données textuelles ainsi que l'émergence de nouvelles formes d'écriture rendent difficile l'extraction automatique d'information à partir de données textuelles souvent hétérogènes et/ou de domaines spécialisés. Afin de relever ces défis, l'équipe MISCA propose des approches originales de fouille de textes permettant l'identification automatique des informations spatio-temporelles et thématiques et leur mise en relation à partir de corpus mis à disposition auprès de la communauté scientifique sur des infrastructures de mutualisation et de partage de données numériques (Dataverse [5], Human-Num [6], Ortolang [7]).

Extraction d'entités spatiales et thématiques

Une partie des travaux de l'équipe MISCA consiste à proposer de nouvelles méthodes d'identification des entités spatiales (absolues et relatives) à partir de corpus peu standardisés [14] dans les domaines de l'agronomie [5], de l'hydrologie [3] et de l'épidémiologie [1]. Ces informations spatiales peuvent être désambiguisées par des méthodes d'apprentissage supervisé et d'apprentissage actif [4]. Les travaux propres à l'identification d'informations thématiques reposent sur l'extraction de la terminologie (logiciel BioTex) via la proposition de nouvelles fonctions de rang [10], le labelling [13] et l'induction d'informations sémantiques [11]. Certaines méthodes mises en œuvre reposent sur la définition de nouveaux descripteurs linguistiques adossés à des méthodes d'apprentissage supervisé et non supervisées [10, 8] et de nouvelles approches de gestion et de stockage [9].

Mise en relation des entités

La mise en relation des différentes entités extraites est alors proposée. Dans ce cadre, des structures appelées STR (Spatial Textual Representation) permettent de représenter, de manière automatique, la configuration spatiale d'un document par des graphes dont les nœuds sont les entités spatiales désambiguisées et les arcs les différentes relations spatiales (adjacence, inclusion, etc.) [6]. Les relations entre entités sont aussi modélisées et extraites sous forme de relations n-aires en combinant des méthodes de fouille de données (extraction de motifs et règles séquentiels) et d'analyse syntaxique [2]. Enfin, l'ensemble des entités (spatio-temporelles et thématiques) mises en relation constituent des évènements pour des applications en veille épidémiologique réalisées dans un cadre pluridisciplinaire [1, 12]. Ainsi, des logiciels de veille en épidémiologie animale (PADI-Web, EpidNews, EpidVis) accompagnés de nouvelles visualisations ont été produits [1, 12, 7].

Références

[1] Elena Arsevska, Sarah Valentin, Julien Rabatel, Jocelyn de Goer de Hervé, Sylvain Falala, Renaud Lancelot, and Mathieu Roche. Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. *PloS One*, 13(8), 2018.

[2] Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie, and Mathieu Roche. Xart : Discovery of correlated arguments of n-ary relations in text. *Expert Systems with Applications*, 73 :115 – 124, 2017.

[3] Nanée Chahinian, Thierry Bonnabaud La Bruyère, Francesca Frontini, Carole Delenne, Marin Julien, Rachel Panckhurst, Mathieu Roche, Lucile Sautot, Laurent Deruelle, and Maguelonne Teissere. Weir-p : An information extraction pipeline for the wastewater domain. In *Proc. International Conference on Research Challenges in Information Science (RCIS), LNBIP, Springer*, 2021.

[4] Amal Chihaoui, Asma Bouhafs Hafsia, Mathieu Roche, and Maguelonne Teisseire. Désa-

5. https ://dataverse.cirad.fr/dataverse/tetis
6. http ://88milsms.huma-num.fr
7. https ://repository.ortolang.fr/api/content/comere/v3.2/cmr-88milsms.html

mbiguïsation des entités spatiales par apprentissage actif. *Rev. Int. Géomatique*, 28(2) :163–190, 2018.

[5] Brett Drury and Mathieu Roche. A survey of the applications of text mining for agriculture. *Computers and Electronics in Agriculture*, 163 :104864, 2019.

[6] Jacques Fize, Mathieu Roche, and Maguelonne Teisseire. Could spatial features help the matching of textual data ? *Intell. Data Anal.*, 24(5) :1043–1064, 2020.

[7] Rohan Goel, Sarah Valentin, Alexis Delaforge, Samiha Fadloun, Arnaud Sallaberry, Mathieu Roche, and Pascal Poncelet. Epidnews : Extracting, exploring and annotating news for monitoring animal diseases. *Journal of Computer Languages*, 56 :100936, 2020.

[8] Roberto Interdonato, Jean-Loup Guillaume, and Antoine Doucet. A lightweight and multilingual framework for crisis information extraction from twitter data. *Soc. Netw. Anal. Min.*, 9(1) :65 :1–65 :20, 2019.

[9] Rodrique Kafando, Rémy Decoupes, Lucile Sautot, and Maguelonne Teisseire. Spatial data lake for smart cities : From design to implementation. *AGILE : GIScience Series*, 1 :8, 2020.

[10] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical term extraction : overview and a new methodology. *Information Retrieval Journal*, 19(1) :59–99, 2016.

[11] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. A novel framework for biomedical entity sense induction. *Journal of Biomedical Informatics*, 84 :31 – 41, 2018.

[12] Sarah Valentin, Elena Arsevska, Sylvain Falala, Jocelyn de Goër, Renaud Lancelot, Alizé Mercier, Julien Rabatel, and Mathieu Roche. Padiweb : A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169 :105163, 2020.

[13] Julien Velcin, Antoine Gourru, Erwan Giry-Fouquet, Christophe Gravier, Mathieu Roche, and Pascal Poncelet. Readitopics : Make your topic models readable via labeling and browsing. In *Proc. of IJCAI*, pages 5874–5876, 2018.

[14] Sarah Zenasni, Eric Kergosien, Mathieu Roche, and Maguelonne Teisseire. Spatial information extraction from short messages. *Expert Systems with Applications*, 95 :351 – 367, 2018.

AfIA
Association française
pour l'Intelligence Artificielle

LIP6/MLIA : Machine Learning for Information Access

LIP6 UMR 7606 / MLIA
CNRS et Sorbonne Université
https://mlia.lip6.fr

Patrick GALLINARI
patrick.gallinari@lip6.fr

Vincent GUIGUE
vincent.guigue@lip6.fr

Sylvain LAMPRIER
sylvain.lamprier@lip6.fr

Benjamin PIWOWARSKI
benjamin.piwowarski@lip6.fr

Laure SOULIER
laure.soulier@lip6.fr

Introduction

L'équipe MLIA du LIP6 est spécialisée dans l'apprentissage statistique (*machine Learning*), dont l'apprentissage profond (*deep learning*) avec un accent particulier sur les aspects algorithmiques et les applications impliquant l'analyse sémantique de données. Ces dernières années, la plus grande partie de notre recherche en recherche d'information et traitement automatique du langage se focalisent sur l'utilisation d'architecture neuronales.

Aperçu de l'état actuel de notre domaine de recherche

En *recherche d'information* (RI), nous nous intéressons aux approches neuronales pour l'ordonnancement de documents, et plus particulièrement les approches évitant l'étape de pré-ordonnancement [9, 8]. Un autre intérêt est également centré sur le futur des moteurs de recherche, en étudiant comment un utilisateur peut dialoguer avec avec système de RI afin de trouver des informations pertinentes pour des besoins d'information complexes (ANR COST et ANR JCJC SESAMS). Ces derniers se caractérisent par le fait que leurs durées s'étendent à plusieurs sessions de recherche, qu'ils incluent l'apprentissage de l'utilisateur au fur et au mesure du déroulement de la recherche, et qu'ils nécessitent la lecture de plusieurs documents de natures différentes. Notre approche se focalise d'une part sur la modélisation fine d'utilisateurs, impliquant la modification de systèmes de RI pour capturer un contexte la recherche en fonction des actions de l'utilisateur [12]. D'autre part, nous nous intéressons à proposer des systèmes proactifs basés sur l'apprentissage par renforcement permettant ainsi de guider l'utilisateur dans sa démarche de recherche [1, 7].

Un autre axe repose sur l'augmentation sémantique des modèles de recherche d'information en combinant la sémantique distributionnelle issue des modèles d'apprentissage de représentation et la sémantique relationnelle recensée dans les bases de connaissances [13, 23].

En *extraction d'information*, nous avons travaillé sur l'apport des modèles de langue pour la détection d'entités [21] et l'extraction de relation de bout en bout [22]. Nous nous intéressons aussi à l'apprentissage non (ou faiblement) supervisé, en exploitant une généralisation des hypothèses du type « si une paire d'entité apparaît dans deux phrases, alors ces deux phrases expriment la même relation » [20], qui permettent d'apprendre de manière non supervisée à détecter des relations avec des modèles très expressifs comme les réseaux de neurones – contrairement aux approches génératives utilisées jusque-là.

En *génération du langage*, nous nous intéressons à deux thèmes : comment transformer une donnée structurée (par ex. un tableau) en texte, et comment générer un résumé abstractif d'un document. Dans les deux cas, nous utilisons des techniques basées sur les réseaux de neurones, et développons des modèles permettant de guider l'appren-

tissage (en résumé automatique [19, 18, 17]), capables de comprendre la hiérarchie des informations structurées pour d'identifier les éléments saillants à retranscrire (en *data-to-text*) [15] ou encore d'effectuer un contrôle sur la génération [16, 14].

La *propagation d'information* sur les réseaux sociaux est au coeur de nombreuses recherches en apprentissage statistique.

Nous nous focalisons sur l'apprentissage de représentations continues, basées sur des méthodes neuronales, pour la modélisation des dynamiques de transmission de contenu en jeu dans ces réseaux. Depuis [3] qui a posé les bases de l'utilisation de ce genre de techniques pour la prédiction de diffusion dans les réseaux, différents modèles ont été développés au sein de l'équipe, notamment [4] ou [11] avec prises en compte de dépendances temporelles plus complexes. Et puisque la propagation sur les réseaux ne concerne pas uniquement des évènements binaires de transmission d'items, nos recherches se sont plus récemment portées sur la diffusion de modèles de langue dans les communautés d'auteurs [5].

La *recommandation*, au sens large, concerne tous les systèmes de personnalisation des interfaces. Il s'agit de comprendre les différentes facettes d'un item ou du profil d'un individu pour évaluer une affinité. Cette tâche est donc intrinsèquement liée à l'apprentissage de représentation, l'enjeu étant de modéliser à la fois le contenu, les interactions et le contexte des acteurs pour faire des propositions pertinentes. La modélisation du contexte temporel offre par exemple de nouvelles opportunités en matière de recommandation [10]. L'étape suivante consiste à expliquer les suggestions pour dépasser le paradigme de la boite noire. Dans cette optique, il est possible d'analyser les données textuelles associées aux items et aux personnes pour générer du texte explicatif associé aux suggestions [6].

L'*ancrage visuel* pour le TAL s'intéresse à l'utilisation de média non textuels (par ex. des images, des vidéos) pour *ancrer* les systèmes de TAL dans le concret : en effet, beaucoup d'informations sont bien plus exprimées dans des médias non textuels (par ex. la position relative de deux objets) que dans un texte. Exploiter cette information permet de construire des représentations de mots ou de phrases qui capturent cette réalité (par ex. en représentant de manière similaire des phrases qui représentent une même scène visuelle). Plus en détail, nos travaux ont porté sur l'utilisation du contexte visuel d'un objet [25], la définition d'un espace sémantique spécifique [2], et l'étude des informations visuelles (fréquence *a priori*, co-occurrence, et apparence visuelle) contenues dans les représentations purement textuelles de mots [24].

Références

[1] Wafa Aissa, Laure Soulier, and Ludovic Denoyer. A reinforcement learning-driven translation model for search-oriented conversational systems. In *SCAI@EMNLP*, pages 33–39, 2018.

[2] Patrick Bordes, Éloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Incorporating Visual Semantics into Sentence Representations within a Grounded Space. In *EMNLP*.

[3] Simon Bourigault, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. Learning social network embeddings for predicting information diffusion. In *WSDM*, pages 393–402, 2014.

[4] Simon Bourigault, Sylvain Lamprier, and Patrick Gallinari. Representation learning for information diffusion through social networks : an embedded cascade model. In *WSDM*, pages 573–582, 2016.

[5] Edouard Delasalles, Sylvain Lamprier, and Ludovic Denoyer. Learning dynamic author representations with temporal language models. In *ICDM'19*, 2019.

[6] Charles-Emmanuel Dias, Vincent Guigue, and Patrick Gallinari. Personalized attention for textual profiling and recommendation. In *EARS@SIGIR*, 2019.

[7] Pierre Erbacher and Laure Soulier. État de l'art des approches de modélisation et de simulation utilisateur pour la recherche d'information conversationnelle. In Antoine Doucet and Adrian-Gabriel Chifu, editors, *CORIA 2021*. ARIA, 2021.

[8] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE : sparse lexical

and expansion model for first stage ranking. In *SIGIR*, pages 2288–2292. ACM, 2021.

[9] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. A white box analysis of colbert. In *ECIR*, pages 257–263. Springer, 2021.

[10] Elie Guàrdia-Sebaoun, Vincent Guigue, and Patrick Gallinari. Latent trajectory modeling : A light and efficient way to introduce time in recommender systems. In *RecSys*, pages 281–284, 2015.

[11] Sylvain Lamprier. A recurrent neural cascade-based model for continuous-time diffusion. In *ICML*, pages 3632–3641, 2019.

[12] Agnès Mustar, Sylvain Lamprier, and Benjamin Piwowarski. Using BERT and BART for query suggestion. In Iván Cantador, Max Chevalier, Massimo Melucci, and Josiane Mothe, editors, *CIRCLE*, volume 2621 of *CEUR Workshop Proceedings*, 2020.

[13] Gia-Hung Nguyen, Laure Soulier, Lynda Tamine, and Nathalie Bricon-Souf. DSRIM : A deep neural information retrieval model enhanced by a knowledge resource driven representation of documents. In *ICTIR*, pages 19–26, 2017.

[14] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. Controlling hallucinations at word level in data-to-text generation. *CoRR*, abs/2102.02810, 2021.

[15] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. A hierarchical model for data-to-text generation. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *ECIR*, volume 12035 of *Lecture Notes in Computer Science*, pages 65–80. Springer, 2020.

[16] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. Parenting via model-agnostic reinforcement learning to correct pathological behaviors in data-to-text generation. In Brian Davis, Yvette Graham, John D. Kelleher, and Yaji Sripada, editors, *INLG*, pages 120–130. Association for Computational Linguistics, 2020.

[17] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Coldgans : Taming language gans with cautious sampling strategies. In *NeurIPS*, 2020.

[18] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Discriminative adversarial search for abstractive summarization. In *ICML*, pages 8555–8564. PMLR, 2020.

[19] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Answers Unite ! Unsupervised Metrics for Reinforced Summarization Models. In *EMNLP*, 2019.

[20] Etienne Simon, Vincent Guigue, and Benjamin Piwowarski. Unsupervised Information Extraction : Regularizing Discriminative Approaches with Relation Distribution Losses. In *ACL*, 2019.

[21] Bruno Taillé, Vincent Guigue, and Patrick Gallinari. Contextualized embeddings in named-entity recognition : An empirical study on generalization. In *European Conference on Information Retrieval*, pages 383–391. Springer, 2020.

[22] Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. Let's stop incorrect comparisons in end-to-end relation extraction ! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, 2020.

[23] Lynda Tamine, Laure Soulier, Gia-Hung Nguyen, and Nathalie Souf. Offline versus online representation learning of documents using external knowledge. *ACM Trans. Inf. Syst.*, 37(4), 2019.

[24] Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Context-Aware Zero-Shot Learning for Object Recognition. In *ICML*, 2019.

[25] Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Learning Multi-Modal Word Representation Grounded in Visual Context. In *AAAI*, 2018.

LORIA/MULTISPEECH : *Speech Modeling for Facilitating Oral-Based Communication*

LORIA UMR 7503/MULTISPEECH
CNRS, INRIA et Université de Lorraine
https://team.inria.fr/multispeech/

Denis JOUVET
denis.jouvet@inria.fr

Emmanuel VINCENT
emmanuel.vincent@inria.fr

Membres

- Denis JOUVET (DR Inria)
- Yves LAPRIE (DR CNRS)
- Emmanuel VINCENT (DR Inria)
- Anne BONNEAU (CR CNRS)
- Antoine DELEFORGE (CR Inria)
- Dominique FOHR (CR CNRS)
- Mostafa SADEGHI (ISFP Inria)
- Vincent COLOTTE (MCF)
- Irina ILLINA (MCF)
- Slim OUNI (MCF)
- Agnès PIQUARD (MCF)
- Romain SERIZEL (MCF)
- Md SAHIDULLAH (chercheur)
- Théo BIASUTTO-LERVAT (post-doctorant)
- Félix GONTHIER (post-doctorant)
- Imran SHEIKH (post-doctorant)
- Tulika BOSE (doctorante)
- Pierre CHAMPION (doctorant)
- Stéphane DILUNGANA (doctorant)
- Sandipana DOWERAH (doctorante)
- Ashwin Geet D'SA (doctorant)
- Adrien DUFRAUX (doctorant)
- Raphaël DUROSELLE (doctorant)
- Nicolas FURNON (doctorant)
- Seyed Ahmad HOSSEINI (doctorant)
- Ajinkya KULKARNI (doctorant)
- Michel OLVERA ZAMBRANO (doctorant)
- Manuel PARIENTE (doctorant)
- Vinicius Souza RIBEIRO (doctorant)
- Shakeel Ahmad SHEIKH (doctorant)
- Prerak SRIVASTAVA (doctorant)
- Nicolas ZAMPIERI (doctorant)

Mots clés

- Parole et audio
- Apprentissage automatique
- Modélisation statistique
- Réseaux de neurones et apprentissage profond
- Rehaussement de la parole
- Reconnaissance de la parole
- Synthèse de la parole
- Synthèse articulatoire
- Traitement du signal
- Traitement de la langue
- Multimodalité (acoustique et visuelle)
- Perception
- Privacité
- Construction de corpus spécifiques (parole, multimodalité, IRM, *etc.*)

Introduction

Les thématiques développées concernent la modélisation de la parole pour faciliter la communication orale ; avec une attention particulière pour les aspects multisources, multilingues et multimodaux (d'où le nom *Multispeech*) :

- *Multisources*, car le signal de parole capté par un microphone est fréquemment bruité ou inclut des superpositions de voix. Dans ce contexte une partie des travaux porte sur la séparation de sources, en particulier pour une prise de son avec plusieurs microphones. Ces travaux contribuent au rehaussement de la parole et à la reconnaissance de parole robuste.
- *Multilingues*, car la parole non-native est influencée par la langue maternelle, ce qui complique notablement son traitement. L'un des domaines applicatifs concernés est l'aide à l'apprentissage de langues étrangères.
- *Multimodaux*, avec la prise en compte des modalités visuelles et acoustiques de la communication vocale pour la synthèse audiovisuelle expressive.

Quelques domaines applicatifs concernés sont les suivants.

- L'*interaction multimodale* avec la synthèse expressive et audiovisuelle, pour améliorer, grâce à l'apport de la composante visuelle, la communication avec des personnes malentendantes et pour aider à l'apprentissage de langues.
- L'*annotation et le traitement de documents audio* avec par exemple la transcription enrichie de documents audio, l'alignement texte-parole (segmentation en mots et/ou en phonèmes) entre autres pour des études linguistiques, et le traitement de documents multimédia.
- Le *monitoring* et la *communication assistée* permettant d'apporter une aide dans des situations de handicap ou pour améliorer l'autonomie. Un exemple concerne la commande vocale mains-libres et le monitoring d'événements sonores dans le cadre de la maison intelligente.
- L'*apprentissage de langues assisté par ordinateur* dont l'objectif est de fournir des retours vers l'apprenant sur la qualité de ses prononciations pour l'articulation des sons comme pour la prosodie. Cela repose sur une analyse des prononciations de l'apprenant. La qualité des diagnostics est conditionnée par la fiabilité de la segmentation phonétique et des paramètres prosodiques calculés.

Thématique générale de l'équipe

Le programme de recherche est structuré selon trois axes.

Le premier axe traite de défis fondamentaux liés à l'apprentissage profond, et vise à aller *au-delà de l'apprentissage supervisé en boîte noire*.

Un premier point concerne l'*intégration de connaissances du domaine*. Les bons résultats empiriques de l'apprentissage profond cachent plusieurs limitations : fonctionnement en boîte noire, gros besoins en données, spécificité à une tâche. Nous explorons des méthodes hybrides combinant l'apprentissage profond d'une part et la modélisation statistique ou le raisonnement symbolique d'autre part, afin de réduire les besoins en données et d'accroître l'interprétabilité. Nous travaillons également sur des modèles génératifs réutilisables pour diverses tâches.

Le deuxième aspect est relatif à l'*apprentissage faiblement supervisé*, c'est-à-dire à partir de données étiquetées de façon incomplète ou potentiellement erronée, et à l'*apprentissage par transfert*, dont le potentiel reste actuellement peu exploré par rapport à l'apprentissage supervisé ou non supervisé.

Le dernier aspect concerne la *préservation de la confidentialité*. Le traitement de la parole dans le cloud soulève des problèmes de confidentialité. Notre objectif est d'anonymiser les données afin d'assurer la confidentialité tout en permettant l'apprentissage de modèles acoustiques [7] et de modèles de langage. Nous explorons également des méthodes d'apprentissage semi-décentralisées et la personnalisation de ces modèles.

Le deuxième axe est relatif à la *production et à la perception de la parole*, et il exploite la dimension physique de celle-ci. La parole résulte du mouvement des articulateurs – mâchoire, lèvres, langue, *etc.* – et se traduit aussi par des déformations visibles sur le visage. De plus la parole ne se limite pas uniquement à une suite de mots : la prosodie joue un rôle important pour structurer l'énoncé vocal et véhiculer l'expressivité (emphase, émotion, *etc.*).

Dans ce cadre, la *modélisation articulatoire* précise les liens entre le signal de parole et la position et le mouvement des articulateurs. L'acquisition et l'analyse des données IRM (Imagerie par Résonance Magnétique), tant statiques que dynamiques, permettent d'améliorer la synthèse articulatoire, c'est-à-dire la production de parole à partir de la connaissance du conduit vocal [2]. Les travaux s'étendent à la modélisation des coarticulations pour une animation précise du visage et des articulateurs.

La *synthèse audiovisuelle expressive* concerne la production d'une synthèse de parole bi-modale (composantes audio et visuelle), avec la prise en compte de l'expressivité sur les deux composantes [1]. Les développements considèrent l'animation de la partie inférieure du visage relative à la parole et de la partie supérieure relative à l'expression faciale, et se poursuivront vers une tête parlante multilingue.

La *catégorisation des sons et de la prosodie* porte sur l'étude des contrastes au niveau prosodique et phonétique, et les relations avec la production et la perception de la parole, tant pour la parole native, y compris dans des situations de han-

dicap [6], que pour la parole non-native en utilisant un corpus bilingue de parole non-native [8].

Le troisième axe est dédié à la *parole dans son environnement* et concerne l'analyse de signaux audio et la reconnaissance vocale.

La *caractérisation de l'environnement acoustique* concerne la localisation de sources sonores [5] y compris en présence d'échos, l'estimation des propriétés acoustiques de salles, et la détection d'événements sonores [9]. Au-delà de la communication parlée, cela a de nombreuses applications comme le monitoring sonore, l'audition robotique, l'acoustique du bâtiment ou la réalité augmentée.

Le *rehaussement de la parole* est particulièrement étudié dans un contexte multicanal [4] et les travaux en cours portent sur le traitement de plusieurs distorsions (écho, réverbération, bruit, parole superposée), et l'utilisation de réseaux de microphones distribués. Les travaux sur la modélisation acoustique robuste aux distorsions tant pour la reconnaissance de la parole que pour la reconnaissance du locuteur ou de la langue, reposent sur la recherche de représentations invariantes, sur l'adaptation de domaine, et sur l'extension de notre approche de propagation de l'incertitude statistique [3] à des modèles plus avancés.

Les aspects *linguistiques et sémantiques* sont également considérés, avec l'utilisation de plongements sémantiques pour, d'une part, rendre la reconnaissance de la parole encore plus robuste, et d'autre part, détecter et classifier les discours haineux dans les médias sociaux (haineux, agressif, insultant, ironique, *etc.*).

Projets marquants

Pour terminer, nous citons ici quelques projets collaboratifs en cours, ou récemment terminés, en lien avec les thèmes décrits ci-dessus.

AI4EU – A European AI On Demand Platform and Ecosystem (H2020 ICT, 2019-2021).

AMIS – Access Multilingual Information opinionS (CHIST-ERA, 2015-2018).

ARTSPEECH – Phonetic articulatory synthesis (ANR, 2015-2019).

BENEPHIDIRE – Le Bégaiement : la Neurologie, la Phonétique, l'Informatique pour son Diagnostic et sa Rééducation (ANR, 2019-2022).

COMPRISE – Cost-effective, Multilingual, Privacy-driven voice-enabled Services (H2020 ICT, 2018-2021).

CONTNOMINA – Exploitation of context for proper names recognition in diachronic audio documents (ANR Blanc SIMI 2, 2013-2016).

COREXp – Acquisition, Processing and Analysis of a Corpus for the Synthesis of Expressive Audiovisual Speech (Région Lorraine, 2014-2016).

CPS4EU – Cyber Physical Systems for Europe (PSPC + ECSEL, 2019-2022).

DEEP-PRIVACY – Apprentissage distribué, personnalisé, préservant la privacité pour le traitement de la parole (ANR, 2019-2022).

DiSCogs – Antennes acoustiques hétérogènes et non contraintes pour la communication parlée (ANR jeunes chercheurs, 2018-2022).

DYCI2 – Creative Dynamics of Improvised Interaction (ANR, 2015-2018).

HAIKUS – Artificial Intelligence applied to augmented acoustic Scenes (ANR, 2019-2023).

HARPOCRATES – Open data, tools and challenges for speaker anonymization (ANR Flash Open Science, 2019-2021).

IFCASL – Individualized Feedback for Computer-Assisted Spoken Language Learning (Programme franco-allemand en SHS, ANR+DFG, 2013-2016).

KAMoulox – Kernel additive modelling for the unmixing of large audio archives (ANR Jeunes Chercheurs, 2015-2019).

LCHN – Langues, connaissances et humanités numériques (CPER, Contrat Plan Etat-Région, 2015-2020).

LEAUDS – Apprentissage statistique pour la compréhension de scènes audio (ANR, 2019-2022).

METAL – Modèles et Traces au service de l'Apprentissage des Langues (e-FRAN, Programme Investissement d'Avenir 2, 2016-2020).

M-PHASIS – Migration et discours haineux dans les médias sociaux Une perspective cross-culturelle (ANR+DFG, 2019-2022).

ORFEO – Tools and ressources for written and spoken French (ANR Corpus, 2013-2016).

ORTOLANG – Open Resources and TOols for LANGuage (EQUIPEX, ANR investissements d'avenir, 2012-2016).

RAPSODIE – Automatic speech recognition for hard of hearing and handicapped people (FUI + FEDER, 2012-2016).

ROBOVOX – Identification vocale robuste pour les robots de sécurité mobiles (ANR, 2019-2023).

VOCADOM – Commande vocale robuste adaptée à la personne et au contexte pour l'autonomie à domicile (ANR, 2017-2020).

VOICEHOME – Robust voice control system for smart home and multimedia applications (FUI, 2015-2017).

Références

[1] Sara Dahmani, Vincent Colotte, Valérian Girard, and Slim Ouni. Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis. In *INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, September 2019.

[2] Benjamin Elie and Yves Laprie. Extension of the single-matrix formulation of the vocal tract : consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink. *Speech Communication*, 82 :85–96, September 2016.

[3] Karan Nathwani, Emmanuel Vincent, and Irina Illina. DNN Uncertainty Propagation using GMM-Derived Uncertainty Features for Noise Robust ASR. *IEEE Signal Processing Letters*, January 2018.

[4] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10) :1652–1664, June 2016.

[5] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin. CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 13(1) :22 – 33, February 2019.

[6] Agnès Piquard-Kipffer and Tamara Léonova. Scolarité et handicap : parcours de 170 jeunes dysphasiques ou dyslexiques- dysorthographiques âgés de 6 à 20 ans. *ANAE - Approche Neuropsychologique des Apprentissages Chez L'enfant*, October 2017.

[7] Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Privacy-Preserving Adversarial Representation Learning in ASR : Reality or Illusion ? In *INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, September 2019.

[8] Jürgen Trouvain, Anne Bonneau, Vincent Colotte, Camille Fauth, Dominique Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, Odile Mella, Bernd Möbius, and Frank Zimmerer. The IFCASL Corpus of French and German Non-native and Native Read Speech. In *LREC'2016, 10th edition of the Language Resources and Evaluation Conference*, Proceedings LREC'2016, Portorož, Slovenia, May 2016.

[9] Nicolas Turpault, Romain Serizel, and Emmanuel Vincent. Semi-supervised triplet loss based learning of ambient audio embeddings. In *ICASSP*, Brighton, France, 2019.

AfIA
Association française
pour l'Intelligence Artificielle

LIPN/RCLN : Représentation des Connaissances et Langage Naturel

LIPN/RCLN
Université Sorbonne Paris Nord, CNRS UMR 7030
https://lipn.univ-paris13.fr/accueil/equipe/rcln/

Thierry CHARNOIS
charnois@lipn.univ-paris13.fr

Nathalie PERNELLE
pernelle@lipn.univ-paris13.fr

Membres impliqués

- Laurent Audibert (MCF)
- Davide Buscaldi (MCF)
- Thierry Charnois (PR)
- Emmanuel Cartier (MCF)
- Jorge Garcia Flores (IR)
- Aude Grezka (IR)
- Joseph Leroux (MCF)
- Francois Levy (PR émérite)
- Adeline Nazarenko (PR)
- Nathalie Pernelle (PR)
- Antoine Rozenkhop (MCF)
- Sylvie Salotti (MCF)
- Nadi Tomeh (MCF)
- Haifa Zargayouna (MCF)
- Manel Zarrouk (MCF)

Introduction

Les activités de recherche menées au Laboratoire d'Informatique de Paris-Nord (LIPN) s'articulent autour d'axes forts s'appuyant sur les compétences de ses membres, notamment en combinatoire, en optimisation combinatoire, en logique et vérification, en langage naturel, en apprentissage. Thème de recherche historique du LIPN, l'Intelligence Artificielle mobilise près de 40% des 150 membres du laboratoire, ce qui fait du LIPN un des principaux groupes de recherche en IA au niveau national.

L'équipe RCLN (Représentation des Connaissances et Langage Naturel) réunit des compétences en traitement automatique des langues, ingénierie des connaissances textuelles, web sémantique, linguistique de corpus, fouille de données et apprentissage automatique à partir de données textuelles. Cette combinaison de compétences lui confère un positionnement original et lui permet de mener des travaux novateurs sur la combinaison de méthodes d'analyse linguistique, de fouille de données, d'apprentissage automatique et de gestion de connaissances hétérogènes pour l'analyse, l'exploitation et l'exploration des corpus à l'échelle du web des données. Ces recherches ont abouti à des résultats publiés dans les conférences TAL mais aussi dans les domaines du web sémantique et du data mining. L'équipe est fortement impliquée dans l'animation et les travaux du LabEx EFL *Empirical Foundations of Linguistics* en particulier dans l'axe Analyse sémantique computationnelle. L'équipe est structurée autour de trois axes de recherche fortement liés :

- L'axe analyse syntaxique et sémantique qui tire parti des méthodes d'apprentissage profond et d'optimisation combinatoire pour proposer des approches d'analyse en syntaxe profonde ou des approches combinant analyse syntaxique et sémantique
- L'axe annotation sémantique et exploration textuelle qui s'intéresse au développement de nouvelles méthodes d'annotation et montre l'intérêt de ces annotations pour l'exploration sémantique,
- L'axe acquisition de connaissances qui développe des méthodes d'acquisition à partir de sources textuelles hétérogènes, mais aussi à partir de graphes de connaissances.

Les axes ci-dessus sont complémentaires, l'analyse syntaxico-sémantique de corpus sert de base à l'annotation sémantique et à l'acquisition de connaissances, et réciproquement l'analyse tire parti des connaissances et annotations. L'équipe travaille à l'intégration de ces différents axes de recherche dans le cadre du projet transverse de fouille de la littérature scientifique.

Analyse syntaxique et sémantique

En dépit des progrès accomplis, l'analyse des textes soulève toujours de nouveaux défis, du fait du volume et de la diversité des textes à analyser,

mais aussi du niveau d'analyse attendu : au-delà de l'analyse de surface, le projet de l'équipe RCLN est d'effectuer de l'analyse en syntaxe profonde et de combiner analyses syntaxique et sémantique. L'objectif est d'améliorer l'analyse en combinant différents systèmes d'analyse, soit sur un même niveau, soit de différents niveaux (i.e. morphologique, lexical, syntaxique) [5, 11]. La combinaison des niveaux d'analyse est modélisée comme un problème d'optimisation combinatoire avec une formulation en terme de théorie des graphes et pour lequel sont développés des algorithmes efficaces (e.g. relaxation lagrangienne [2], programmation dynamique) : analyse en constituants discontinus via le problème de l'arborescence couvrante généralisée, algorithme efficace et exact pour l'analyse en dépendances incorporant directement les scores affectés aux arcs dans les systèmes en transitions [11], systèmes joints pour l'analyse 'easy-first' considérant plusieurs schémas d'annotations et plusieurs analyses en parallèle. Enfin, un travail en cours qui s'intéresse à la question de l'interprétabilité des résultats issus de modèles neuronaux pour l'analyse syntaxique montre qu'il est possible d'apprendre des représentations - en partie - interprétables par l'utilisation d'auto-encodeurs variationnels et l'étude du comportement des variables latentes.

Annotation sémantique et exploration textuelle

L'annotation sémantique consiste à enrichir les documents par des métadonnées qui relient le texte à des connaissances externes. L'équipe RCLN développe de nouveaux modèles et de nouvelles méthodes d'annotation automatique, et montre leur intérêt pour l'exploration textuelle sur différents types de corpus. L'équipe s'intéresse aux modèles d'annotation dédiés a l'annotation des règles issues des textes juridiques et réglementaires [8]. Un langage contrôlé abstrait a été proposé comme langage d'annotation et nous avons montré comment les outils de TAL peuvent aider à traduire les règles écrites en langage naturel dans ce langage. D'autres travaux portent sur l'annotation de textes issus des réseaux sociaux. L'objectif est de repérer les messages malveillants ou ironiques [9] ou encore leur géolocalisation dans le cas de situation d'urgence [10]. Il s'agit de problèmes particulièrement difficiles du fait du peu de données disponibles pour un apprentissage supervisé. Les derniers résultats portent sur le problème de la géolocalisation pour lequel (i) une architecture neuronale originale par transfer learning en cascades a été conçue pour la géolocalisation du pays puis de la zone plus précise, (ii) une approche sémantique fondée sur des connaissances sur les POI a également été proposée pour des zones géographiques à plus petite échelle [10].

Un autre volet des travaux de l'équipe porte sur l'exploration textuelle pour extraire des connaissances. Le projet RENFO s'intéresse ainsi à la recherche d'experts et à l'extraction de leur parcours à partir de requêtes sur le web. Afin de minimiser l'espace de recherche, une approche à base d'apprentissage profond par renforcement a été développée. Un autre travail s'intéresse à la recherche d'experts en concevant une approche fondée sur l'annotation des textes scientifiques et leur représentation en graphes pour appliquer des techniques d'abstraction de graphes avec contraintes topologiques et découvrir des experts et leur expertise associée. Enfin, d'autres résultats portent sur le résumé automatique abstractif : (i) une approche originale basée sur les réseaux de neurones de type Transformers a été proposée pour traiter des textes longs tout en obtenant des performances à l'état de l'art ; (ii) la méthode d'évaluation automatique SERA, adaptée aux résumés abstractifs mais conçue pour le domaine biomédical, a été étendue au domaine général.

Acquisition de connaissances

Cet axe porte sur l'analyse et la fouille de textes pour acquérir des connaissances structurées et permettre le développement d'applications TAL prenant en compte la dimension sémantique des textes. Les travaux menés reposent sur une combinaison originale de techniques d'acquisition de connaissances, d'analyses linguistiques et de technologies du web sémantique. L'équipe s'intéresse à la détection et la caractérisation de nouveaux usages en contexte ainsi qu'à la caractérisation textuelle appliquée à la stylistique. Elle travaille ainsi sur la détection et la modélisation des néologies de formes

ou des néologies sémantiques. Les approches développées se basent sur une première annotation manuelle constituant un premier jeu de référence pour un apprentissage (faiblement) supervisé, ou combinent modèles neuronaux et fouille de motifs. Ces travaux ont conduit à la réalisation de l'outil Neoveille de détection et de suivi des néologismes sur le web [1], et au dictionnaire morphologique évolutif et collaboratif Morfetik. L'équipe a par ailleurs proposé différentes approches fondées sur la découverte de motifs pour la caractérisation des styles d'auteurs ou de genres littéraires [6]. Les travaux de l'équipe portent également sur l'extraction de relations sémantiques entre concepts et propose des méthodes non supervisées combinant des techniques de clustering, de sémantique distributionnelle et de fouille de motifs [4]. Ces approches qui extraient des relations lexicales et contextuelles ont permis de produire des corpus annotés en sources ouvertes qui nous ont conduit à créer le premier challenge international portant sur l'évaluation de méthodes d'extraction de relations, dans le cadre de la campagne SemEval 2018 [3].

Enfin, l'équipe s'intéresse également à la découverte de connaissances à partir de graphes de données volumineux. Un grand nombre de données accessibles sur le web sont décrites sous forme de graphes qui contiennent des milliards de faits mais ceux-ci sont incomplets et contiennent souvent des données erronées. Différents algorithmes de découverte efficace de règles expressives à partir de graphes de données volumineux ont été définis et l'un d'eux a permis d'obtenir la première place au track SpimBench-Liage de données lors de la compétition internationale OAEI 2020 [7].

Fouille de la littérature scientifique

L'exploitation de la littérature scientifique est un enjeu majeur aujourd'hui pour la recherche elle-même mais aussi pour tout le champ de l'innovation ouverte. Au-delà de l'accès aux sources bibliographiques, il s'agit de concevoir des outils d'exploration sémantique interactifs qui permettent de retrouver des documents, de faire émerger des notions ou relations latentes, de cartographier un domaine, d'analyser des tendances, ou encore de rechercher des experts sur un domaine particulier. L'équipe RCLN travaille sur différents projets de fouille de la littérature scientifique et éprouve les méthodes développées sur différents types de corpus et domaines (e.g. philosophie, informatique).

Références

[1] Emmanuel Cartier. Neoveille, a web platform for neologism tracking. In *EACL, Software Demonstrations*, pages 95–98. ACL, 2017.

[2] Caio Corro, Joseph Le Roux, Mathieu Lacroix, Antoine Rozenknop, and Roberto Wolfler Calvo. Dependency parsing with bounded block degree and well-nestedness via lagrangian relaxation and branch-and-bound. In *ACL (1)*, 2016.

[3] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. Semeval-2018 task 7 : Semantic relation extraction and classification in scientific papers. In *SemEval@NAACL-HLT*, pages 679–688. ACL, 2018.

[4] Kata Gábor, Haïfa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois. Exploring vector spaces for semantic relations. In *EMNLP*, pages 1814–1823. ACL, 2017.

[5] Yash Kankanampati, Joseph Le Roux, Nadi Tomeh, Dima Taji, and Nizar Habash. Multi-task easy-first dependency parsing : Exploiting complementarities of different dependency representations. In *COLING*, pages 2497–2508, 2020.

[6] Dominique Legallois, Thierry Charnois, and Meri Larjavaara. Trends in Linguistics. Studies and Monographs, 248 pages. 2018.

[7] Armita Khajeh Nassiri, Nathalie Pernelle, Fatiha Saïs, and Gianluca Quercini. Generating referring expressions from RDF knowledge graphs for data linking. In *ISWC*, pages 311–329, 2020.

[8] Adeline Nazarenko, François Lévy, and Adam Z. Wyner. An annotation language for semantic search of legal sources. In *LREC, Japan*. (ELRA), 2018.

[9] Diego Reforgiato Recupero, Mehwish Alam, Davide Buscaldi, Aude Grezka, and Farideh

Tavazoee. Frame-based detection of figurative language in tweets [application notes]. *IEEE Comput. Intell. Mag.*, 14(4) :77–88, 2019.

[10] Laura Di Rocco, Federico Dassereto, Michela Bertolotto, Davide Buscaldi, Barbara Catania, and Giovanna Guerrini. Sherloc : a knowledge-driven algorithm for geolocating microblog messages at sub-city level. *Int. J. Geogr. Inf. Sci.*, 35(1) :84–115, 2021.

[11] Joseph Le Roux, Antoine Rozenknop, and Mathieu Lacroix. Representation learning and dynamic programming for arc-hybrid parsing. In *CoNLL*, pages 238–248. ACL, 2019.

AfIA
Association française
pour l'Intelligence Artificielle

IRIT/SAMoVA : Structuration, Analyse et Modélisation de documents Vidéo et Audio

IRIT / SAMoVA
Université Toulouse III - Paul Sabatier
www.irit.fr/departement/signaux-images/samova/

Régine ANDRÉ-OBRECHT
regine.andre-obrecht@irit.fr

Julien PINQUIER
julien.pinquier@irit.fr

Membres permanents impliqués

- Régine ANDRE-OBRECHT (PR émérite)
- Hervé BREDIN (CR CNRS)
- Jérôme FARINAS (MCF)
- Isabelle FERRANE (MCF)
- Philippe JOLY (PR)
- Julie MAUCLAIR (MCF)
- Thomas PELLEGRINI (MCF)
- Julien PINQUIER (MCF, Responsable)
- Christine SENAC (MCF)

Historique de l'équipe

L'équipe ART.ps "Analyse, Reconnaissance et Traitement de la parole et des sons se constitue en 1999 autour d'Isabelle Ferrané, Christine Sénac et Régine André-Obrecht. Le thème de recherche privilégié était centré sur le traitement du signal de parole à des fins de reconnaissance avec comme objectif privilégié, l'identification automatique des langues et du locuteur. Une des retombées applicatives émergentes était l'indexation de documents sonores. L'équipe devient l'équipe SAMoVA en 2002, avec l'arrivée de Philippe Joly ; le cœur des recherches reste l'analyse et la modélisation, dans un but d'identification et de structuration, mais il s'y ajoute une dimension supplémentaire : l'information n'est plus issue d'un seul média, mais dérive de l'analyse conjointe des deux média que sont l'audio et la vidéo.

Thématique de l'équipe

Les travaux de l'équipe SAMoVA se placent dans le contexte de l'indexation et de la recherche de contenus des documents audio et vidéo (AV). Plus précisément, la recherche s'appuie sur le traitement du signal, la modélisation ainsi que sur la structuration du contenu audiovisuel.

- En **Analyse du signal**, l'équipe possède une grande expertise et savoir faire en segmentation du signal audio et vidéo. Nombre d'algorithmes originaux permettent d'accéder à une segmentation multi-échelle (segmentation phonétique ou prosodique, localisation et regroupement de locuteurs...). Une fois extraites, ces informations sont en général utilisées comme entrées dans les tâches de reconnaissance, de structuration de haut niveau.
- Les études en **modélisation** sont développés à des fins de reconnaissance (parole, musique, locuteur, langue, bruits saillants). Initialement très centrées sur les Modèles de Markov Cachés, les recherches actuelles s'inscrivent dans la mouvance des réseaux bayésiens, des réseaux de neurones (Apprentissage Profond) et celle des approches de bout en bout ("End-to-End"). Les problèmes d'apprentissage en fonction du volume des données d'apprentissage (augmentation de données) sont explorés. Les études d'analyse et de modélisation se rejoignent pour définir des mesures objectives d'intelligibilité voire de compréhension de la parole, en y intégrant la dimension perceptive et productive, et ce en situation dégradée (environnement sonore, handicap).
- Afin de caractériser l'organisation d'un document AV et sa **structuration**, sont automatiquement localisés des segments relevant éventuellement de différentes échelles temporelles, fournissant un accès direct au contenu et à une information pertinente [6]. Ces travaux sont également menés afin de comparer et d'évaluer la similarité en termes de contenus avec comme but, celui d'organiser des bases de données hétérogènes en collections par exemple.
  Une des approches récentes consiste à propo-

ser une structuration automatique de contenus audiovisuels en unités centrées sur les situations d'interaction conversationnelle personnes/personnes (dialogue, conversations) ou personnes/système (robot ou assistant vocal) et à caractériser ces unités conversationnelles ; des mesures de leur qualité en terme d'intelligibilité, de fluidité, de compréhensibilité, et de complexité des échanges sont élaborées. Cette approche est basée sur la multimodalité et notamment sur la fusion de caractéristiques issues des contenus audio, vidéo et texte.

Activités contractuelles

La majorité des recherches est menée dans un cadre contractuel (projets ANR, Région, entreprises) avec l'aide incontournable de nombreux doctorants (allocations ministérielles, ANR, CIFRE, Région). Une forte coopération scientifique s'est installée avec l'Université Toulouse Jean Jaurès (UT2J, équipe Octogone-Lordat) et avec le CHU de Toulouse (Hopital Larrey, IUCT). Les principales relations industrielles s'articulent autour des sociétés Linagora et Archean Technologies. Quelques exemples des projets en cours (2020) sont donnés ci-après.

- Projet Evolex : né en 2013 d'une collaboration entre SAMoVA et ToNIC, il permet d'envisager un traitement informatisé des réponses enregistrées sur une tâche de fluence lexicale, puis sur deux tâches linguistiques : la dénomination d'images et la génération sémantique verbale. La quantité de données récoltées et la complexité des processus lexicaux et sémantiques impliqués a nécessité une ouverture vers l'équipe Octogone-Lordat [5]. Une plateforme disposant de fonctionnalités de reconnaissance automatique de la parole avec des techniques à l'état de l'art est actuellement en cours de déploiement, afin de pouvoir produire de nouvelles normes dans ces tâches et permettre de meilleures analyses des fonctions cognitives.
- La contribution au projet LinTo porte sur la structuration automatique de réunions en séquences, sur la base de l'étude des interactions entre participants, de la fusion multimodale (audio, vidéo) des données perçues, et de la caractérisation des séquences en type d'interaction : question-réponse, présentation, tour de table... type de "parole" des participants... L'objectif de l'extraction de ces indicateurs est de rendre possible une analyse plus fine des conversations pour aider à la constitution automatique de comptes-rendus. Ce travail est réalisé en collaboration avec le LAAS-CNRS [7] et l'équipe MELODI de l'IRIT.
- Dans le cadre du LabCom ALAIA (ANR) en partenariat avec Archean, est définie une mesure de la qualité de la parole de locuteurs non-natifs apprenants d'une langue cible ; le focus porte sur la prononciation avec la détection et la localisation d'erreurs de prononciation (répétition de mots et de phrases) et sur la compréhensibilité d'énoncés produits plus librement ou en fonction d'un contexte (parole guidée ou semi-spontanée).
- La recommandation musicale est le contexte dans lequel s'inscrit le projet Région ECREME (en partenariat avec l'UT2J). Il s'agit de définir automatiquement un modèle spécifique à chaque utilisateur, sur la base de critères cognitifs et contextuels. Pour ce faire, des techniques d'apprentissage automatique sont utilisées et le challenge est d'atteindre une information cognitive de haut niveau malgré une faible quantité de données provenant de l'analyse cognitive [3].
- L'analyse de la voix pathologique est un des sujets prioritaires de l'équipe impliquée dans deux projets : production de corpora et analyses de la voix dans le cadre de maladies de type cancer (projet ANR RUGBI) ou de type Parkinson ou Atrophie multisystématisée (projet Voice4PD-MSA) [9, 1].
- L'intelligibilité de la parole est au cœur des préoccupations de plusieurs chercheurs de l'équipe. Celle-ci est étudiée aussi bien en production de la parole (réseau européen H2020 TAPAS et projet ANR RUGBI) [2] qu'en perception de la parole (projet région Occitanie AUDIOCAP) [4].
- Le projet ANR Jeunes Chercheurs LUDAU a pour cible l'étude des algorithmes d'apprentissage profond pour la découverte d'événements sonores dans un mode peu supervisé [8].
- Avec le projet Région INGPRO, est étudié l'impact des gestes correctifs (réalisés par un avatar) dans l'apprentissage de la prononciation de

langue étrangère.

Pour accompagner ces recherches et faciliter leur intégration dans un contexte d'indexation multimédia, un important travail d'ingénierie a conduit à la réalisation d'outils informatiques mis à disposition : https://www.irit.fr/SAMOVA/site/ressources/

Références

[1] Mathieu Balaguer, Jérôme Farinas, Pascale Fichaux-Bourin, Michèle Puech, Julien Pinquier, and Virginie Woisard. Validation of the french versions of the speech handicap index and the phonation handicap index in patients treated for cancer of the oral cavity or oropharynx. *Folia Phoniatrica et Logopaedica*, novembre 2019.

[2] Mathieu Balaguer, Timothy Pommée, Jérôme Farinas, Julien Pinquier, Virginie Woisard, and Renée Speyer. Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis : Systematic review. *Head & Neck*, 42(1) :111–130, janvier 2020.

[3] Nicolas Dauban, Christine Senac, Julien Pinquier, Pascal Gaillard, Ludovic Florin, and Paul Albenge. Automatic Analysis and Musicological Interpretation of Human Free Sorting of Musical Excerpts (regular paper). In *International Conference on Advances in Multimedia (MMEDIA 2019), Valencia, Espagne, 24-28/04/2019*, pages 42–48. IARIA, avril 2019.

[4] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, Julien Tardieu, Cynthia Magnen, Pascal Gaillard, Xavier Aumont, and Christian Füllgrabe. Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language and Hearing Research*, 60 :2394–2405, septembre 2017.

[5] Bruno Gaume, Ludovic Tanguy, Cécile Fabre, Lydia-Mai Ho-Dac, Bénédicte Pierrejean, Nabil Hathout, Jérôme Farinas, Julien Pinquier, Lola Danet, Patrice Peran, Xavier De Boissezon, and Mélanie Jucla. Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures (regular paper). In *Natural Language Processing and Cognitive Science, Krakow, Poland, 11–12/09/18*, pages 19–26. Jagiellonian Library, 2018.

[6] Zein Al Abidin Ibrahim, Isabelle Ferrané, and Philippe Joly. Temporal relation algebra for audiovisual content analysis. *Multimedia Tools and Applications*, 78(11) :15275–15316, 2019.

[7] Francisco Madrigal, Frederic Lerasle, Lionel Pibre, and Isabelle Ferrané. Audio-Video detection of the active speaker in meetings. In *25th International Conference on Pattern Recognition (ICPR2021)*, January 2021.

[8] Thomas Pellegrini and Léo Cances. Cosine-similarity penalty to discriminate sound classes in weakly-supervised sound event detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[9] Virginie Woisard, Corine Astésano, Mathieu Balaguer, Jérôme Farinas, Corinne Fredouille, Pascal Gaillard, Alain Ghio, Laurence Giusti, Imed Laaridh, Muriel Lalain, Benoît Lepage, Julie Mauclair, Olivier Nocaudie, Julien Pinquier, Gilles Pouchoulin, Michèle Puech, Danièle Robert, and Vincent Roger. C2SI corpus : a database of speech disorder productions toassess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, June 2020.

AfIA
Association française
pour l'Intelligence Artificielle

LORIA/SEMAGRAMME : Analyse sémantique des langues naturelles

Inria Nancy Grand Est, Univ. de Lorraine,
LORIA/ Sémagramme
Inria
semagramme.loria.fr

Philippe DE GROOTE
Philippe.deGroote@inria.fr

Membres permanents de l'équipe

- Maxime Amblard (MCF)
- Karën Fort (MCF)
- Philippe de Groote (DR Inria)
- Bruno Guillaume (CR Inria)
- Michel Musiol (PU, associé)
- Guy Perrier (PU émérite)
- Sylvain Pogodalla (CR Inria)

Grammaires catégorielles abstraites

L'équipe Sémagramme développe depuis plusieurs années un modèle linguistique basé sur la théorie des types : les Grammaires Catégorielles Abstraites [3]. Ces grammaires, qui peuvent être vues comme un modèle logique de l'interface syntaxe-sémantique, permettent d'encoder de nombreux formalismes grammaticaux existants [4, 8]. Elles bénéficient du développement d'un outil logiciel [7] et ont donné lieu à un transfert industriel [9].

Sémantique graphique

L'un des objectifs de l'équipe est de proposer une représentation de la sémantique de la langue naturelle qui soit utile par la qualité de sa représentation tout en restant accessible pour les non spécialistes. Il s'agit d'intégrer dans la formalisation proposée des éléments repris de formalismes installés (AMR, DMRS). Les fondements se trouvent dans la représentation logique.

Réécriture de graphes

Un des modèles de calcul proposé par Sémagramme est la réécriture de graphes. Une implémentation de ce modèle (Grew, grew.fr) est disponible Plusieurs applications à des transformations linguistiques ont été développées : conversion d'annotations entre différents formats syntaxiques et production d'analyses syntaxiques ou sémantiques. Ces applications montrent que le formalisme générique de la réécriture de graphes peut être pertinent dans le contexte des application en TAL (gestion des structures de traits, de l'interface avec un lexique et des ambiguïtés possibles dans certaines transformations). Le livre [1] décrit le modèle de calcul, les applications linguistiques et les aspects mathématiques de la réécriture de graphes. À noter également qu'un outil d'exploration de corpus utilisant la notion de matching de graphes au cœur de Grew est disponible en ligne (match.grew.fr) sur un grand nombre de corpus. En 2020, plus de 100 000 requêtes ont été faites sur cet outil.

Production de ressources langagières

Sémagramme est impliquée dans la production de ressources syntaxiques en français dans les projets Sequoia (deep-sequoia.inria.fr) et Universal Dependencies (universaldependencies.org).

L'équipe a également créé des jeux ayant un but pour la création de ressources langagières liées à la syntaxe, en particulier ZombiLingo [6] et ZombiLUDik, pour l'annotation en syntaxe de dépendances, et RigorMortis [5] pour l'annotation d'unités polylexicales (MWE).

Analyse du discours pathologique

ODiM et MePheSTO sont des projets interdisciplinaires, à l'interface de : la psychiatrie-psychopathologie, la linguistique, la sémantique formelle et les sciences du numérique.

ODiM tend à remplacer le paradigme des Troubles du Langage et de la Pensée (TLP) tel qu'on l'utilise dans le secteur de la Santé mentale par un modèle sémantico-formel des Troubles du Discours (TDD). Il s'agit de traduire ces troubles en signes diagnostiques ainsi que de dépistage des

personnes vulnérables dites « à risques ». Le projet se décline selon les trois axes suivants : 'recueil de données', 'développement du modèle théorique', 'outillage informatique du modèle'.

Le projet MePheSTO est un projet DFKI-Inria. Il a été conçu pour faire progresser de manière significative les connaissances et les bases méthodologiques du développement technologique dans le domaine des diagnostics assistés par ordinateur et du traitement des maladies pour les troubles psychiatriques tels que les troubles mentaux, affectifs et de l'humeur. En s'appuyant sur les méthodes modernes d'intelligence artificielle, ce projet combine des techniques issues de domaines tels que, mais sans s'y limiter, la vision par ordinateur, la linguistique informatique, l'ingénierie de la connaissance, ainsi que les sciences du comportement, et les applique dans plusieurs cas d'utilisation clinique. L'objectif global du projet est de développer un cadre/une base méthodologique pour une validation scientifiquement solide des phénotypes numériques pour les troubles psychiatriques, basée sur une entrée multimodale, y compris la parole, la vidéo et les bio-signaux provenant des interactions sociales cliniques.

Éthique et IA

Certains membres de l'équipe sont à l'origine du groupe "éthique et TAL", et du blog du même nom (http://www.ethique-et-tal.org/), ainsi que d'un certain nombre d'actions sur le sujet : organisation de l'atelier ETeRNAL (éthique et TAL) à TALN 2015 et 2020, numéro spécial de la revue TAL (TAL et éthique, 57 (2), 2016), comité d'éthique de EMNLP 2020 et NAACL 2021, track *ethics and NLP* à ACL 2021. Ces travaux d'animation de la communauté ont également donné lieu à des publications et à un projet ANR, *artificial text COrpus DEsIgNed Ethically : automatic synthesis of clinical document* (CODEINE), qui démarre en 2021. Par ailleurs, une membre de l'équipe participe en tant que *Project Ethics Officer* au projet européen AI-Proficient et est co-créatrice du tutoriel sur la relecture d'articles en TAL, présenté [2] à ACL 2020 et à EACL 2021.

Références

[1] Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. *Application de la réécriture de graphes au traitement automatique des langues*, volume 1 of *Série Logique, linguistique et informatique*. ISTE editions, September 2018.

[2] Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes, and Aurélie Névéol. Reviewing Natural Language Processing Research. Annual Meeting of the Association for Computational Linguistics (ACL), July 2020.

[3] Ph. de Groote. Towards Abstract Categorial Grammars. In *Association for Computational Linguistics, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference*, pages 148–155, 2001.

[4] Ph. de Groote and S. Pogodalla. On the expressive power of abstract categorial grammars : Representing context-free formalisms. *Journal of Logic, Language and Information*, 13(4) :421–438, 2004.

[5] Karën Fort, Bruno Guillaume, Yann-Alan Pilatte, Mathieu Constant, and Nicolas Lefèbvre. Rigor Mortis : Annotating MWEs with a Gamified Platform. In *LREC 2020 - Language Resources and Evaluation Conference*, Marseille, France, May 2020.

[6] Bruno Guillaume, Karën Fort, and Nicolas Lefebvre. Crowdsourcing complex language resources : Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2016.

[7] Sylvain Pogodalla. ACGTK : un outil de développement et de test pour les grammaires catégorielles abstraites. In Laurence Danlos and Thierry Hamon, editors, *Actes de la 23ème Conférence sur le Traitement Automatique des Langues Naturelles, 31ème Journées d'Études sur la Parole, 18ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (JEP-TALN-RECITAL 2016)*, pages 1–2, Paris, France, July 2016. Association pour le Traitement Automatique des Langues. Démonstration.

[8] Sylvain Pogodalla. A syntax-semantics interface for Tree-Adjoining Grammars through Abstract Categorial Grammars. *Journal of Language Modelling*, 5(3) :527–605, 2017.

[9] R. Salmon. *Natural language generation using Abstract Categorial Grammars*. Thèse de doctorat, Université Paris Diderot, 2017.

LORIA/SMART : *Speech Modelisation and Text*

LORIA UMR 7503 / SMaRT
CNRS, INRIA et Université de Lorraine
smart.loria.fr

Kamel SMAILI
kamel.smaili@loria.fr

Membres impliqués

- Kamel SMAÏLI (PR)
- David LANGLOIS (MCF)
- Joseph DI MARTINO (MCF)
- Salima HARRAT (chercheuse associée)
- Karima MEFTOUH (chercheuse associée)
- Youness MOUKAFIH (doctorant)
- Fadi EL GHAWNMEH
- Karima ABIDI (post-doc)

Mot-clés

- traduction automatique
- reconnaissance automatique de la parole
- corpus parallèles
- corpus comparables
- traitement automatique de la langue arabe
- dialectes arabes
- code-switching
- analyse de sentiments
- fouille d'opinions
- réseaux sociaux
- correction de voix
- Réponse musicale automatique par des techniques de TAL

Thèses issues de l'équipe à partir de 2014

- 2011-2015 : Motaz Khaled SAAD, Fouille de documents et d'opinions multilingues.
- 2013-2018 : Salima HARRAT, Traduction automatique fondée sur des méthodes statistiques : application aux langues peu dotées en ressources.
- 2015-2019 : Ameur DOUIB, Algorithmes bio-inspirés pour la traduction automatique statistique.
- 2014-2019 : Imen BEN OTHMANE, Conversion de la voix : approches et applications.
- 2016-2019 : Karima ABIDI, La construction automatique de ressources multilingues à partir des réseaux sociaux : application aux données dialectales du Maghreb (soutenance prévue en décembre).
- 2016-2020 : Mohamed Amine MENACER, Reconnaissances et traduction automatique de la parole de vidéos arabes et dialectales.
- 2016-2020 : Nouha OTHMAN, Learning to Retrieve Relevant Passages and Questions in Open Domain and Community Question Answering
- 2019- : Younness MOUKAFIH, Identification of comparable segments using multi-task neural training for informal and poorly endowed languages.
- 2019- : Fadi AL-GHAWANEMEH, NLP based automatic composition with sentimental control.

Thématiques de l'équipe

L'objectif principal de l'équipe SMaRT est de développer des modèles de représentation linguistique pour les systèmes de traduction automatique et de reconnaissance vocale. Cette modélisation implique l'utilisation de méthodes mathématiques pour identifier, extraire et proposer des associations entre deux ou plusieurs langues pour la traduction et la reconnaissance vocale. Les langues sont étudiées à travers des corpus monolingues, parallèles ou comparables pour les langues à faibles ressources (dialectes arabes), l'arabe, le français et l'anglais.

SMaRT contribue à la recherche dans le domaine du multilingue selon plusieurs axes : construction de corpus, traduction automatique, mesures de comparabilité entre documents, analyse de sentiments et fouille d'opinion. L'idée qui sous-tend nos travaux est le fait que les opinions sur des sujets sensibles peuvent varier fortement d'une culture à l'autre, et donc d'une langue à l'autre. Nous cherchons donc à proposer des travaux et projets aidant l'être humain à rechercher des informations sur des sujets

proches en plusieurs langues, et à comparer ces informations sur le plan des opinions.

SMaRT traite aussi la question du code-switching dans les langues en développant des méthodes de traduction et de reconnaissance vocale pour les langues concernées [3, 10].

Nous travaillons dans SMaRT sur un sujet original et qui concerne le problème de la réponse musicale automatique appliquée à la musique arabe orientale en utilisant des techniques de traitement automatique de la langue. En fait, dans le contexte de la musique maqamique arabe, le mawwāl est une improvisation vocale non métrique et est souvent appliquée à la poésie narrative. À la fin de chaque phrase vocale du mawwāl, l'instrumentiste effectue une récapitulation ou une traduction de la phrase vocale. Notre objectif est de proposer une réponse musicale automatique à l'image de ce que pourrait faire un instrumentiste expérimenté [4].

Un autre axe de recherche de SMaRT concerne la reconnaissance automatique de voix pathologiques. Il s'agit de transcrire des voix très déformées par la maladie, ou encore d'améliorer le signal pour renforcer l'intelligibilité de telles voix. Pour ce faire, l'équipe SMaRT développe des modèles pour améliorer l'intelligibilité de voix œsophagiennes via des techniques de conversion de voix. Ces techniques peuvent être considérées comme une sorte de traduction du signal en un autre.

Principaux travaux de l'équipe

Corpus. Notre approche étant celle de l'apprentissage automatique, les corpus sont une ressource vitale. Ces corpus peuvent être parallèles (pour chaque phrase d'une langue donnée, le corpus propose sa traduction dans une ou plusieurs langues) ou comparables (les corpus sont alignés au niveau du document, et deux documents comparables abordent le même sujet, sans nécessairement être traduction l'un de l'autre). Or, notre équipe s'intéresse fortement à la traduction des dialectes arabes. Les dialectes arabes, contrairement à l'arabe standard, ou à l'arabe classique, sont des langues orales, donc sans corpus écrits. Pourtant, de nos jours, sur les réseaux sociaux, les usagers arabophones utilisent leur dialecte sous une forme écrite. Il faut donc créer des corpus pour appliquer les techniques d'apprentissage automatique. Dans ce cadre, l'équipe a proposé deux corpus : PADIC (**P**arallel **A**rabic **DI**alect **C**orpus) [8] est un corpus parallèle de 6400 phrases en arabe standard, algérien, tunisien, marocain, syrien et palestinien ; et CALYOU (**C**omparable spoken **AL**gerian corpus extracted from **YOU**tube) [1], qui est un corpus comparable de commentaires algériens (caractères arabes et latins) et français issus de YouTube.

Un deuxième moyen de contribuer à la construction de corpus est de proposer des méthodes permettant de mesurer la comparabilité entre documents multilingues. Ces mesures peuvent alors être utilisées pour retrouver sur Internet des documents comparables. Notre équipe a proposé et comparé des méthodes en ce sens, fondées sur des dictionnaires bilingues ou sur l'approche *Latent Semantic Indexing*, et appliquées à des corpus issus de Wikipédia, Euronews et Al Jazeera [7].

Moteurs et modèles de traduction. La traduction automatique implique de proposer des « moteurs » permettant de traduire une phrase d'une langue en une autre. Un tel algorithme utilise des modèles de traduction.

En ce qui concerne le moteur de traduction, nous avons proposé une approche fondée sur les algorithmes génétiques : au lieu de construire incrémentalement des hypothèses de traduction, l'algorithme manipule à tout moment un ensemble (population) d'hypothèses (modélisées sous forme de chromosomes). La qualité de traduction de la population des hypothèses s'améliore au fur et à mesure via des opérations de mutation et de croisement, que nous avons créées en suivant l'approche de l'algorithmique génétique [6].

Mesures de confiance et estimation de qualité. Un des problèmes de la traduction automatique est qu'on ne sait par ce qu'est une bonne traduction. Pour une phrase donnée, il existe plusieurs traductions, plus ou moins correctes, ou tout aussi correctes que les autres. Comment mesurer la qualité d'une traduction ? L'estimation de qualité cherche à répondre à cette question. Depuis 2012, cette problématique est montée en puissance via l'organisation d'une campagne d'évaluation liée à la confé-

rence Machine Translation. Nous avons participé trois fois à cette campagne, en proposant des caractéristiques de la phrase traduite fondées sur les mesures de confiance issues de l'équipe (utilisées initialement pour guider l'algorithme de traduction [13]), en enrichissant le corpus d'apprentissage, ou encore en comparant la phrase traduite à évaluer aux traductions de systèmes état de l'art.

Analyse de sentiments et fouilles d'opinions. Les corpus comparables multilingues peuvent être analysés sur le plan des sentiments et des opinions exprimées. Il s'agit surtout de détecter les différences d'opinions. Dans ce cadre, au niveau textuel, nous avons comparé et amélioré plusieurs méthodes de classification [7], et nous avons proposé une approche utilisant la théorie de l'*appraisal* [2] permettant de donner une analyse plus fine des sentiments présents dans un document que ce qui est fait habituellement.

Amélioration de la voix. Depuis 2014 nous travaillons dans l'équipe SMaRT sur l'amélioration de la voix pathologique. La correction vocale est le terme que nous utilisons dans le cadre du rehaussement de la voix pathologique. Nous avons pu obtenir une amélioration sensible de la voix œsophagienne grâce à des techniques de conversion vocale [12]. Ces travaux se poursuivent dans le cadre de trois autres thèses.

Projets

Notre équipe a été à l'origine depuis 2014 de deux projets :

- AMIS [14] (**A**ccess to **M**ultilingual **I**nformation and Opinion**S**) est un projet ChistEra financé par l'Union Européenne (de décembre 2015 à novembre 2019). Il s'agit de proposer des méthodes pour obtenir un résumé audio et vidéo de vidéos issues de YouTube ainsi qu'une analyse comparative des sentiments et opinions des vidéos multilingues parlant du même sujet [2].
- TRAM (**TR**anslation of **A**rabic **M**usic) est un projet financé par l'AUF de 2016 à 2018. Ce projet a pour objectif de fournir automatiquement un accompagnement musical à une ligne mélodique proposée par un chanteur arabe. Pour ce faire, nous utilisons des méthodes issues de la traduction automatique.

Conclusion et perspectives

Cet article ne donne qu'un bref aperçu des activités de l'équipe SMaRT. En effet, notre investissement depuis plusieurs années sur les dialectes arabes permet d'aborder plusieurs problèmes spécifiques dont la recherche de données sur les réseaux sociaux, la prise en compte du *code-switching*, l'identification du dialecte [9], etc. Sur ces sujets, nous reportons le lecteur à la page de publications de l'équipe. La section Références ci-dessous liste un échantillon représentatif de nos publications. Nous avons entamé des travaux sur la détection de *fake news* [5] et sur le *multitask learning* pour apprendre à traduire et à développer des systèmes de reconnaissance automatique de la parole pour les dialectes [11].

Références

[1] K. Abidi, M. A. Menacer, and K. Smaili. CALYOU : A Comparable Spoken Algerian Corpus Harvested from YouTube. In *18th Annual Conference of the International Communication Association (Interspeech)*, 2017.

[2] Karima Abidi, Dominique Fohr, Denis Jouvet, David Langlois, Odile Mella, and Kamel Smaïli. A Fine-grained Multilingual Analysis Based on the Appraisal Theory : Application to Arabic and English Videos. In *ICALP : International Conference on Arabic Language Processing*, volume Communications in Computer and Information Science book series (CCIS, volume 1108), pages 49–61, Nancy, France, October 2019. Springer.

[3] Karima Abidi and Kamel Smaïli. CESAR : A new metric to measure the level of code-switching in corpora -Application to Maghrebian dialects. In *IntelliSys2021*, Springer series "Advances in Intelligent Systems and Computing", Amsterdam, Netherlands, September 2021.

[4] Fadi M Al-Ghawanmeh, Melissa J Scott, Mohamed-Amine Menacer, and Kamel Smaïli. Predicting and Critiquing Machine Virtuosity :

Mawwal Accompaniment as Case Study. In *International Computer Music Conference*, Santiago, Chile, July 2021.

[5] Maysoon Alkhair, Karima Meftouh, Nouha Othman, and Kamel Smaïli. An Arabic Corpus of Fake News : Collection, Analysis and Classification. In *Arabic Language Processing : From Theory to Practice7th International Conference, ICALP 2019, Nancy, France, October 16–17, 2019, Proceedings*, volume Communications in Computer and Information Science book series (CCIS, volume 1108), pages 292–302. October 2019.

[6] A. Douib, D. Langlois, and K. Smaïili. Genetic-based Decoder for Statistical Machine Translation. In *Springer LNCS series, Lecture Notes in Computer Science*. 2016.

[7] D. Langlois, M. Saad, and K Smaïli. Alignment of comparable documents : comparison of similarity measures on French-English-Arabic data. *Natural Language Engineering*, 2018.

[8] K. Meftouh, S. Harrat, and K. Smaïli. PADIC : extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey, 2018.

[9] Karima Meftouh, Karima Abidi, Salima Harrat, and Kamel Smaïli. The SMarT Classifier for Arabic Fine-Grained Dialect Identification. In *The Fourth Arabic Natural Language Processing Workshop co-located with ACL*, Florence, Italy, August 2019.

[10] Mohamed Menacer, David Langlois, Denis Jouvet, Dominique Fohr, Odile Mella, and Kamel Smaïli. Machine Translation on a parallel Code-Switched Corpus. In *Canadian AI 2019 - 32nd Conference on Canadian Artificial Intelligence*, Lecture Notes in Artificial Intelligence, Ontario, Canada, May 2019.

[11] Mohamed Amine Menacer and Kamel Smaïli. Investigating data sharing in speech recognition for an underresourced language : the case of algerian dialect. In *7th International Conference on Natural Language Processing - NATP 2021*, Vienna, Austria, March 2021.

[12] I. Othmane Ben, J. Di Martino, and K. Ouni. Enhancement of esophageal speech obtained by a voice conversion technique using time dilated fourier cepstra. *International Journal of Speech Technology*, 22(1) :99–110, 2019.

[13] S. Raybaud, D. Langlois, and K. Smaïli. "This sentence is wrong." Detecting errors in machine-translated sentences. *Machine Translation*, 25(1) :p. 1–34, 2011.

[14] K. Smaïli, D. Fohr, C.-E. González-Gallardo, M. L. Grega, L. Janowski, D. Jouvet, A. Koźbial, D. Langlois, M. Leszczuk, O. Mella, M. A. Menacer, A. Mendez, E. Pontes Linhares, E. Sanjuan, J.-M. Torres-Moreno, and B. Garcia-Zapirain. Summarizing videos into a target language : Methodology, architectures and evaluation. *Journal of Intelligent and Fuzzy Systems*, 1 :1–12, 2019.

AfIA
Association française
pour l'Intelligence Artificielle

LORIA/SyNaLP : *Symbolic and Statistical Natural Language Processing*

LORIA UMR 7503/SYNALP
CNRS, INRIA et Université de Lorraine
http://synalp.loria.fr

Christophe CERISARA
christophe.cerisara@loria.fr

Claire GARDENT
claire.gardent@loria.fr

Membres :

- Nadia BELLALEM (MCF)
- Lotfi BELLALEM (PRAG)
- Ilias BENJELLOUN (doctorant)
- Paul CAILLON (doctorant)
- Christophe CERISARA (CR CNRS)
- Emilie COLIN (doctorante)
- Samuel CRUZ-LARA (MCF)
- Angela FAN (doctorante)
- Christine FAY-VARNIER (MCF)
- Claire GARDENT (DRCE CNRS)
- Timothy GARWOOD (doctorant)
- Jean-Charles LAMIREL (MCF)
- Bart LAMIROY (MCF)
- Guillaume LE BERRE (doctorant)
- Anna LEDNIKOVA (doctorante)
- Joël LEGRAND (MCF)
- Thien Hoa LE (doctorant)
- Yannick PARMENTIER (MCF)
- Anastasia SHIMORINA (doctorante)

Introduction

Synalp (Symbolic and Statistical Natural Language Processing) est une équipe de recherche en traitement automatique des langues (TAL). Nous nous intéressons en particulier à la génération automatique et à la simplification de texte, à la modélisation syntaxique et sémantique, aux systèmes de question-réponse, à la classification de textes en thèmes et sentiments, et au dialogue. Pour aborder ces défis, nous nous appuyons sur les outils théoriques que sont :

- les grammaires formelles,
- l'apprentissage profond,
- l'apprentissage faiblement supervisé.

Thématiques de recherche

Les travaux de recherche actuels de l'équipe se concentrent sur les questions liées à la génération de texte, à l'analyse sémantique et au dialogue. Nous appliquons ces recherches à une variété de types de données, qui vont de la transcription de la parole spontanée aux bases de publications, en passant par les microblogs et les documents écrits. Lorsque cela est pertinent, nous considérons et modélisons l'information contextuelle et paralinguistique. Par exemple, l'émotion, l'opinion et les actes de dialogue sont des domaines d'application de nos recherches.

Plus généralement, les informations linguistiques font presque toujours partie d'un ensemble plus large d'informations connexes : hyperliens et références à des bases de connaissances sur les pages web, structures de documents et métadonnées dans les rapports scientifiques, géolocalisation, horodatages, graphes des réponses, des renvois et des utilisateurs sur les réseaux sociaux (par ex. Twitter), *etc.* Ces multiples dimensions de l'information disponible doivent être intégrées à l'analyse sémantique pour mieux interpréter le langage naturel.

De plus, la forme de l'entrée linguistique elle-même évolue, et nous devons rendre nos modèles plus robustes à de telles évolutions. Par exemple, de nouveaux mots et expressions apparaissent constamment sur le web, des phrases non-grammaticales sont rencontrées fréquemment dans des dialogues oraux, des abréviations et des émoticônes apparaissent et acquièrent de nouvelles fonctions sémantiques et pragmatiques sur Twitter. De tels phénomènes sont mieux modélisés au niveau du caractère qu'au niveau lexical.

Dans le passé l'équipe a utilisé des outils formels de description du langage naturel, en particulier les grammaires d'arbres adjoints lexicalisées, et des approches hybrides combinant modèles formels

et méthodes statistiques d'apprentissage automatique. Depuis plusieurs années, nous travaillons essentiellement avec des réseaux neuronaux profonds que nous construisons à partir de patrons classiques – convolutions, séquences, attention, transformers, *etc.* – et adaptons à nos objectifs de recherche.

Nous proposons également des algorithmes d'apprentissage faiblement supervisés et non-supervisés [11], pour entraîner ces modèles.

Projets marquants

Génération de textes. Dans le domaine de la génération de textes, l'équipe Synalp a travaillé sur différents aspects de la micro-planification, une tâche qui consiste à convertir des données en texte. L'approche privilégiée est de combiner des modèles linguistiques (grammaires et lexiques) avec des modèles statistiques. Afin de faciliter le développement manuel de grammaires pour la langue naturelle, nous avons ainsi développé un langage de spécification et un compilateur ainsi que des méthodes de fouilles d'erreur qui permettent la détection semi-automatique des erreurs et omissions introduites lors de la spécification manuelle. Afin de gérer l'explosion combinatoire des choix possibles, nous avons proposé des algorithmes combinant des méthodes d'apprentissage automatique (champs aléatoires conditionnels) avec des algorithmes d'analyse syntaxique inversée [3].

Plus récemment, nous avons exploré la micro-planification pour différents types d'entrées : arbres de dépendances, requêtes OWL sur des bases de connaissances et ensembles de triplets RDF. Nous avons notamment mis au point une technique permettant de produire des corpus d'apprentissage pour la génération à partir de données RDF et organisé une campagne d'évaluation internationale sur ce sujet [4]. Plusieurs projets collaboratifs sont actuellement en cours (Chaire IA xNLG, Projet ANR Quantum, Projet H2020 NL4XAI), qui portent sur les approches neuronales de type encodeur-décodeur pour la génération de texte à partir de représentations sémantiques abstraites [12], d'arbres de dépendances [13], de textes [2] ainsi que pour les systèmes de question-réponse [2].

Syntaxe, sémantique et résumé. Nous avons abordé l'analyse syntaxique et sémantique du langage naturel via des modèles supervisés, en particulier les modèles à noyaux d'arbres et les réseaux neuronaux profonds, tels que les auto-encodeurs récursifs ou les modèles *seq2seq* pour la tâche de question-réponse [14, 2]. Dans cette dernière approche, nous proposons d'utiliser une grammaire hors contexte et des automates à états finis pour guide un modèle neuronal et ainsi mieux prendre en compte les contraintes syntaxiques et lexicales qui résultent de la phrase d'entrée.

Nous exploitons également l'analyse des structures syntaxiques ainsi que les approches d'apprentissage par renforcement appliquées à des modèles de type encodeur-décodeur avec attention et copy-pointer afin d'améliorer la qualité des résumés automatiques [9].

Dialogues et discours. Dans le domaine de l'analyse des dialogues humain-humain sur les réseaux sociaux, nous avons notamment travaillé sur l'analyse conjointe des sentiments et la détection des émotions. Ces tâches soulèvent des difficultés concernant par exemple la couverture des modèles linguistiques sous-jacents à tous les niveaux (lexical, syntaxique, sémantique ou pragmatique). Dans ce cadre, nous avons proposé un modèle neuronal d'apprentissage multitâches qui présente de très bonnes performances en transfert d'information entre ces deux tâches [1] et permet de limiter les coûts d'annotation. L'étude des interactions entre sentiments et dialogues a également permis d'analyser l'évolution des interactions dans les dialogues sur les réseaux sociaux de manière plus qualitative. Nous poursuivons également des travaux sur l'influence des représentations lexicales multilingues sur la classification des actes de dialogue en partenariat avec l'université tchèque de Plzen [10]. Des travaux sur l'analyse du discours sont également menés en partenariat avec l'équipe Orpailleur du LORIA et le laboratoire IRIT [5].

Classification de textes. Un premier domaine d'application concerne l'analyse des sentiments dans les textes, domaine dans lequel nous avons notamment comparé l'importance relative des re-

présentations en mots et en caractères, et montré le faible impact obtenu en adaptant des réseaux très profonds inspirés des modèles état de l'art en image [8].

Dans le domaine de la classification non supervisée de textes, depuis plusieurs années, nous focalisons nos recherches sur le développement et l'exploitation de nouvelles métriques adaptées au traitement de grands corpus de données et applicables à des contextes variés. Nous avons proposé ainsi la métrique de maximisation des caractéristiques qui peut être substituée aux distances classiques de clustering tout en facilitant l'analyse qualitative du processus de clustering, contrairement aux méthodes à base de noyaux. Notre métrique, qui a l'avantage d'être non-paramétrique, a été appliquée notamment à l'analyse diachronique de grands corpus de publications scientifiques. Nous avons également développé un nouvel algorithme de clustering neuronal dérivé de l'algorithme IGNG mais intégrant cette nouvelle métrique au sein d'un algorithme d'apprentissage de type Estimation-Maximisation.

Nous avons montré que la méthode résultante, dénommée IGNGF, permettait d'obtenir des performances supérieures à celles de l'état de l'art pour le clustering incrémental de données hétérogènes. IGNGF a également obtenu de meilleurs résultats que d'autres méthodes de TAL dont l'analyse en concepts formels (AFC) et l'analyse spectrale pour la classification automatique de verbes en français appliquée à l'analyse de rôles sémantiques [7].

Nous avons également adapté la métrique de maximisation des caractéristiques aux modèles supervisés afin de proposer de nouvelles solutions au problème des classes déséquilibrées dans les grands corpus. Nous avons exploité cette métrique afin d'inférer automatiquement le modèle optimal de clustering parmi un ensemble de modèles possibles [6].

Références

[1] Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T Le. Multi-task dialog act and sentiment recognition on Mastodon. In *COLING*, Santa Fe, United States, August 2018.

[2] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4184–4194, Hong Kong, China, November 2019. Association for Computational Linguistics.

[3] Claire Gardent and Laura Perez-Beltrachini. A statistical, grammar-based approach to microplanning. *Computational Linguistics*, 43(1) :1–30, April 2017.

[4] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 179–188, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[5] Laurine Huber, Yannick Toussaint, Charlotte Roze, Mathilde Dargnat, and Chlo'e Braud. Aligning Discourse and Argumentation Structures using Subtrees and Redescription Mining. In *6th International Workshop on Argument Mining*, Florence, Italy, August 2019.

[6] Jean-Charles Lamirel, Nicolas Dugué, and Pascal Cuxac. New efficient clustering quality indexes. In *International Joint Conference on Neural Networks (IJCNN 2016)*, Vancouver, Canada, July 2016.

[7] Jean-Charles Lamirel, Ingrid Falk, and Claire Gardent. Federating clustering and cluster labelling capabilities with a single approach based on feature maximization : French verb classes identification with igngf neural clustering. *Neurocomputing*, 147 :136 – 146, 2015.

[8] Hoa T. Le, Christophe Cerisara, and Alexandre Denis. Do Convolutional Networks need to be Deep for Text Classification ? In *AAAI Workshop on Affective Content Analysis*, New Orleans, United States, February 2018.

[9] Hoa T Le, Christophe Cerisara, and Claire Gardent. How much can Syntax help Sentence Compression ? In *ICANN 2019*, Proceedings of ICANN 2019, Munich, Germany, September 2019.

[10] Jiří Martínek, Pavel Král, Ladislav Lenc, and Christophe Cerisara. Multi-Lingual Dialogue Act Recognition with Deep Learning Methods. In *Proc. Interspeech 2019*, pages 1463–1467, 2019.

[11] Hubert Nourtel, Christophe Cerisara, and Samuel Cruz-Lara. Deep unsupervised system log monitoring. In *PROFES 2019 - 20th International Conference on Product-Focused Software Process Improvement*, Barcelona, Spain, November 2019.

[12] Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. Enhancing AMR-to-text generation with dual graph representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3181–3192, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Anastasia Shimorina and Claire Gardent. LORIA / lorraine university at multilingual surface realisation 2019. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 88–93, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Chunyang Xiao, Marc Dymetman, and Claire Gardent. Symbolic priors for rnn-based semantic parsing. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4186–4192, 2017.

AfIA
Association française
pour l'Intelligence Artificielle

LS2N/TALN : Traitement Automatique du Langage Naturel

LS2N UMR 6004 / TALN
CNRS et Université de Nantes
https://www.ls2n.fr/equipe/taln/

Emmanuel MORIN
emmanuel.morin@ls2n.fr

Membres impliqués

- Adrien BAZOGE(doctorant)
- Denis BÉCHET (MCF)
- Florian BOUDIN (MCF)
- Mérième BOUHANDI (doctorante)
- Béatrice DAILLE (PR)
- Victor CONNES (doctorant)
- Colin DE LA HIGUERA (PR)
- Richard DUFOUR (PR)
- Chantal ENGUEHARD (MCF)
- Kévin ESPASA (doctorant)
- Esther FÉLIX (ingénieur)
- Corentin FOLLENFAN (ingénieur)
- Ygor GALLINA (doctorant)
- Nicolas HERNANDEZ (MCF)
- Christine JACQUIN (MCF)
- Martin LAVILLE (doctorant)
- Bastien MASSE (ingénieur)
- Laura MONCEAUX-CACHARD (MCF)
- Emmanuel MORIN (PR)
- Timothée POULAIN (doctorant)
- Solen QUINIOU (MCF)
- Andréane ROQUES (ingénieur)

Présentation

La masse de données langagières qui est maintenant disponible permet de mettre en œuvre des techniques robustes, indépendantes des langues. Néanmoins, quantité de données ne signifie pas toujours qualité, robustesse et finesse d'analyse. L'équipe TALN tente de concilier ces aspects antagonistes en proposant des méthodes d'analyses de textes robustes adaptables à la diversité des données langagières écrites s'exprimant sur des nouveaux supports communicationnels (e.g. blogs, réseaux sociaux, forums) ou encore dans des langues différentes. Nos travaux sont par nature multidisciplinaires, au cœur des données, en interaction avec les sciences humaines et sociales (e.g. linguistique, terminologie, traduction), les sciences de l'éducation et avec d'autres thématiques de l'informatique comme l'apprentissage, la fouille de données ou la recherche d'information.

Thématiques

Les travaux de l'équipe s'organisent autour de trois thématiques :

Analyse sémantique et discursive. Nous nous intéressons aux problématiques de modélisation et d'analyse d'information discursive sur plusieurs plans d'organisation : thématique, intentionnel, attentionnel, subjectif (e.g. opinion, sentiment, émotion). Il s'agit, par exemple, du suivi automatique de la dynamique d'une conversation en termes d'actes du dialogue. Du point de vue du traitement linguistique, cette tâche n'est pas simple car, d'une part, le marquage linguistique peut prendre plusieurs formes (e.g. morphologique, lexicale) et, d'autre part, par essence, ces informations sont composites et distribuées dans le texte. Du point de vue numérique, la tâche est également complexe car il s'agit de concevoir des représentations qui permettent à la fois de capturer et d'intégrer différentes dimensions du sens (e.g. thématique, intentionnelle) ainsi que soutenir des opérations de comparaison entre ces représentations (e.g. similarité, complémentarité) tout en restant interprétables par l'homme.

Apprentissage et fouille de textes. Nous utilisons l'apprentissage automatique dans de nombreuses applications, comme la recherche d'interactions entre organismes biologiques, à partir d'articles scientifiques, la détection d'éléments implicites dans les textes littéraires, ou la construction de modèles permettant l'alignement de lexiques bilingues. Nous travaillons également à l'élaboration de nouveaux algorithmes

et à l'amélioration de méthodes existantes. Par exemple, nous cherchons à prédire les dépendances itérées dans les grammaires de dépendances, et nous développons des modèles de graphes pour tirer parti de ces connaissances structurelles des textes.

Alignement multilingue et multimodal. Nous nous intéressons aux méthodes de rapprochement de diverses sources de données pour pouvoir bénéficier d'informations complémentaires : les alignements. Nous travaillons sur les alignements de corpus comparables, c'est-à-dire des textes dans deux langues sans rapport de traduction, ainsi que sur les alignements de corpus multimodaux, c'est-à-dire des textes provenant de différentes modalités, telles que l'oral, l'écriture manuscrite et différents types de textes écrits (e.g. textes bien formés, forums, chats, tweets).

Domaines applicatifs

Les travaux de l'équipe visent principalement quatre domaines applicatifs.

Multilinguisme. Nos travaux dans le champ du multilinguisme concernent deux axes complémentaires. Le premier est celui des méthodes d'extraction de lexiques bilingues, à partir de corpus comparables (BLI - Bilingual Lexicon Extraction), en proposant des approches d'alignement permettant d'obtenir des résultats en domaines de spécialité comparables à ceux obtenus en langue générale. Le second est celui de l'exploitation des résultats d'alignement en TAO (Traduction Assistée par Ordinateur), en nous intéressant notamment à la personnalisation d'un système de traduction neuronale.

Indexation documentaire. Nos travaux portent sur l'indexation sémantique, et en premier lieu sur les modèles d'indexation documentaire dans les bibliothèques numériques d'articles scientifiques. Nous concentrons nos efforts sur le développement de modèles permettant d'enrichir l'indexation (c'est-à-dire par l'ajout de mots-clés n'apparaissant pas dans le document source) et sur la mise au point d'approches non supervisées, en nous appuyant sur la profusion d'articles scientifiques disponibles pour construire des modèles performants.

Ingénierie des ressources éducatives. Nos travaux dans le domaine de l'enrichissement des ressources éducatives concernent des cas d'usage ciblés correspondant, par exemple, au profilage de groupes, de la mesure de l'engagement cognitif et émotionnel d'un apprenant, du soutien à la conception de nouveaux outils de navigation dans les ressources éducatives, voire de l'assistance à la conception de cours en ligne. Nous nous intéressons également à l'ordonnancement automatique des ressources éducatives, à leur recommandation et la construction de playlists.

Médical et santé. Le TAL appliqué au domaine du médical est un champ recherche à part entière. Par comparaison à d'autres domaines de spécialité, le TAL médical bénéficie de nombreuses ressources terminologiques et d'une abondance de données textuelles : articles scientifiques, dossiers électroniques patients, réseaux sociaux alimentés par les patients, etc. Le TAL médical propose de nombreuses applications qui ont un impact direct sur la pratique clinique, et améliorent la dissémination des connaissances auprès des professionnels de la santé mais aussi du grand public. Parmi ces applications, nos travaux portent sur la fouille de textes des compte-rendus hospitaliers pour recueillir et croiser à grande échelle des données cliniques en collaboration avec la clinique des données du CHU de Nantes.

AfIA
Association française
pour l'Intelligence Artificielle

Comité de pilotage du collège TLH

Florian BOUDIN (L2SN - Université de Nantes)
Davide BUSCALDI (LIPN - Université Paris XIII)
Gaël DIAS (GREYC - Université de Caen Normandie)
Corinne FREDOUILLE (LIA - Université d'Avignon)
José MORENO (IRIT - Université Paul Sabatier)
Aurélie NEVEOL (LIMSI - Université Paris Sud)
Yannick PARMENTIER (LORIA - Université de Lorraine)
François PORTET (LIG - Institut Polytechnique de Grenoble)
Mathieu ROCHE (TETIS - CIRAD)
Serena VILLATA (I3S - Université Côte d'Azur)

Pour contacter le collège TLH

Responsable
Mathieu ROCHE
UMR TETIS
AgroParisTech, Cirad, Cnrs, Inrae
500, rue J.F. Breton
34093 Montpellier Cedex 5, France
mathieu.roche@cirad.fr

Site WEB
https://afia-tlh.loria.fr/