AfIA

Association française
pour l'Intelligence Artificielle

Dossier N° 1

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

Collège TLH

Avril 2020 N° 1

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 dossier vise à recenser les équipes de recherche académiques et industrielles françaises menant des travaux à l'intersection du traitement automatique des langues, de la recherche d'information, de la communication parlée et de l'intelligence artificielle.

Les technologies du langage humain (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. Elles englobent traditionnellement le traitement automatique des langues (TAL), la communication parlée (CP) et leurs applications les plus emblématiques comme la recherche d'information (RI) et la traduction automatique.

Suite à un appel à participation communiqué sur les listes de diffusion françaises des domaines de recherche des TLH, nous avons reçu 20 contributions, dont 18 issues de laboratoires académiques, réparties sur 10 villes plus Paris et sa région (Figure 1).

Fig. 1 : Cartographie des TLH en France.

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, dans (4) un cadre éventuellement pluri ou transdisciplinaire, alors vous trouverez un interlocuteur dans ce dossier.

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 dossier que j'en ai pris à sa réalisation. Bonne lecture.

Gaël DIAS

AfIA
Association française
pour l'Intelligence Artificielle

BIBLIOME : Acquisition et Formalisation de Connaissances à partir de Textes

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

Claire NÉDELLEC
claire.nedellec@inra.fr

Robert BOSSY
robert.bossy@inra.fr

Louise DELÉGER
louise.deleger@inra.fr

Arnaud FERRÉ
arnaud.ferre@inra.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, microbiolologie, 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 [6] 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 CONTES [7] associe par apprentissage automatique 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-ST) 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'INRA 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) 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, 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).

[7] Arnaud Ferré, Pierre Zweigenbaum, and Claire Nédellec. Representation of complex terms in a vector space structured by an ontology for a normalization task. In *BioNLP 2017, Vancouver, Canada, August 4, 2017*, pages 99–106, 2017.

[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.

AfIA
Association française
pour l'Intelligence Artificielle

CARTEL : Corpus, Application, Ressources pour le Traitement et l'Étude du Langage

CLLE UMR 5263 / ERSS
CNRS et Université de Toulouse
https://w3.erss.univ-tlse2.fr

Ludovic TANGUY
ludovic.tanguy@univ-tlse2.fr

Membres Impliqués

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

Présentation générale

CLLE (Cognition, Langues, Langage, Ergonomie, UMR 5263) est un laboratoire pluridisciplinaire en sciences cognitives. Il est actuellement composé de deux équipes :

- l'ERSS (Équipe de Recherche en Syntaxe et Sémantique) qui travaille plus particulièrement des thématiques relevant de la linguistique (phonologie, morphologie, syntaxe, sémantique, discours, TAL, didactique des langues, psycholinguistique).
- le LTC (Laboratoire Travail et Cognition) qui couvre de nombreux champs de la psychologie (ergonomie cognitive, cognition sociale, développement du langage et de la communication, neurosciences).

L'axe Cartel de l'ERSS regroupe les membres du laboratoire dont les recherches se situent dans le domaine du traitement automatique des langues (TAL) et de la linguistique outillée. Les principaux objectifs de l'axe concernent la fertilisation mutuelle de la linguistique (modèles, approches sur corpus) et de l'ingénierie linguistique (méthodes et outils informatiques) autour de la manipulation, l'étude et l'exploitation de matériaux langagiers. Le recours à des traitements assistés ou automatisés permet aux membres de l'axe 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 Cartel 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'axe 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 Cartel 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 distri-

butionnelle (impact des corpus, évaluation qualitative, 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 de l'axe 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'axe 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 l'axe Cartel 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'axe 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

L'axe Cartel 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'axe 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'axe 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 l'axe Cartel 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 l'axe, 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

ERIC : Entrepôts, Représentation et Ingénierie des Connaissances

Laboratoire ERIC EA 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) 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 [5], 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 [11]. 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).

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 [6]. 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]. Des travaux en cours consistent à utiliser des mécanismes d'attention, mis en lumière par le succès de l'architecture du Transformer, dans ce formalisme [3].

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 [4].

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 [8]. Nous avons également intégré le contexte dans les cubes de textes afin d'obtenir des analyses OLAP plus pertinentes [7]. 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) [9]. 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 [10].

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 permettant à des utilisateurs non-experts d'effectuer des analyses OLAP, ainsi que des regroupements de documents similaires [9] 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 *The World Wide Web Conference*, pages 2587–2593. ACM, 2019.

[3] R. Brochier, A. Guille, and J. Velcin. Link prediction with mutual attention for text-attributed networks. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 283–284. ACM, 2019.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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

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

- Jean-Michel DAUBE (PRAG)
- Kata GABOR (MCF)
- Marie-Anne MOREAUX (MCF)
- Damien NOUVEL (MCF)
- Frédérique SEGOND (PAST)
- François STUCK (IGR)
- Mathieu VALETTE (PR)

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, mais aussi de la linguistique générale (lexicologie textuelle, sémantique textuelle, morphologie lexicale).

L'équipe est structurée selon les axes :

- *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 participer au développement d'outils de fouille de textes, d'analyse et d'interprétation de textes assistées. Les applications visées sont celles de la recherche d'information, la classification de documents et la fouille de textes.
- *Acquisition des connaissances*. Élaboration et 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, d'ontologies, *etc.*).
- *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 de corpus et de TAL.
- *Corpus et multilinguisme*. Les thèmes abordés sont les enjeux théoriques et pratiques des corpus multilingues (parallèle et comparable), la problématique du multilinguisme dans le traitement automatique du document numérique et la prise en compte technique des spécificités associées (écritures, encodages).

Projets

Sémantique textuelle et analyse du discours

- *TALAD (2018-2022)*. Adaptation des techniques issues du TAL pour apporter à l'analyse du discours des jeux de descripteurs plus complexes, en particulier pour l'étude des nominations par utilisation des entités nommées et des chaînes de coréférences.

- *ANR Contint ACCORDYS (2012-2016)*. Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas de DYSmorphologie foetale (INSERM, LIMSI, INALCO, Hôpital Trousseau, ANTIDOT).

Multilinguisme et langues peu dotées

- *INaLCO MANTAL (2014-2017)*. Analyse morpho-syntaxique du bambara à partir d'un corpus partiellement désambiguisé et de techniques d'apprentissage automatique.
- *MultiTAL (2015-2016)*. Plateforme de documentation et d'expertise des outils et ressources pour le traitement automatique des langues orientales et des langues peu dotées.
- *SPC Blanc APRECIADO (2013-2016)*. Analyse et spatialisation des Perceptions et Représentations sociales des Changements environnementaux en Afrique De l'Ouest sahélo-soudanienne (Paris Nord, Paris Diderot, INALCO) (français, anglais, peul, wolof, djerma).
- *GAELL (2014-2015)*. Réalisation et mise en ligne d'un générateur automatique d'exercices d'estonien, issus du CoPEF, un Corpus Parallèle Franco-Estonien d'environ 65 millions de mots (AideMoi, dispositif d'aide à la lecure en L2).

Acquisition de connaissances

- *Labex EFL Axe 5 : Analyse sémantique computationnelle (2011-2024)*. Dans le champ de la linguistique computationnelle, l'axe 5 du Labex met l'accent sur l'analyse sémantique et son application dans divers outils d'accès au contenu, dont l'extraction d'informations et de connaissances. Les membres de l'équipe participent aux recherches en « Extraction de relations sémantiques dans des corpus de spécialité » en tant que coresponsable de l'opération. L'extraction de relations sémantiques est un composant clé dans l'identification des connaissances de domaine et leur structuration dans des bases de connaissances. Dans le cadre du même projet, des membres d'ERTIM participent aussi à l'opération « Étude de la variation et du changement lexical ».
- *CNES - TALREX (2018-2020)*. Exploitation de rapports techniques liés aux lancement de fusées afin de les numériser puis de mettre en place des outils de recherche d'information et de détection de signaux faibles.

AfIA
Association française
pour l'Intelligence Artificielle

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

Laurent BESACIER
Responsable d'équipe
laurent.besacier@univ-grenoble-alpes.fr

Membres permanents de l'équipe

- Véronique AUBERGÉ (CR)
- Valérie BELLYNCK (MCF)
- Laurent BESACIER (PR)
- 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 (MCF)
- Fabien RINGEVAL (MCF)
- Solange ROSSATO (MCF)
- Didier SCHWAB (MCF)
- Gilles SÉRASSET (MCF)
- Michel VACHER (IR)

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 [1]. 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.

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

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

fond pour l'informatique affective. On peut notamment citer l'utilisation de GANs (*Generative Adversarial 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 actuellement 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 diction-

naires 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 Lexi-*

cal 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

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

- Houssam AKHMOUCH (doctorant)
- Céline ALEC (MCF)
- Judith Jeyafreeda ANDREW (doctorante)
- Nihal Yagmur AYDIN (IGR)
- Anaëlle BALEDENT (doctorante)
- Pierre BEUST (MCF HDR)
- Gaël DIAS (PR)
- Stéphane FERRARI (MCF HDR)
- Emmanuel GIGUET (CR HDR)
- Govind (post-doctorant)
- Amit KUMAR (doctorant)
- Yann MATHET (MCF HDR)
- Fabrice MAUREL (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) [4] et sur l'identification de relations lexico-sémantiques entre unités de sens (par apprentissage profond [21] et par approche statistique [7]) 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 [3].

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 [17], à partir de

laquelle de nombreuses applications ont pu émerger [18, 19].

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) [2].

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 [5] 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 [6] 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 [16].

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 [14]. 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 [26].

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 [12]. Ainsi, les informations sur les entités peuvent ensuite être utilisées pour de nombreuses applications, comme par exemple la datation automatique de photographies [25], la représentation des contenus du web par *empreinte sémantique* [10] ou l'étude de la viralité de l'information [9].

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) [20]. Dans ce cadre, les chercheurs du GREYC ont développé plusieurs systèmes qui permettent (1) de prédire l'évolution des taxonomies [23], (2) d'aligner automatiquement des bases de connaissances structurées [24], ou hétérogènes dans différentes langues [22], (3) de prédire la diffusion d'un évènement dans des communautés parlant une langue étrangère [11], et (4) d'analyser des documents web en fonction des entités nommées qu'ils contiennent [8].

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 [32, 31] 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 [32]).

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 [13].

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 [15] ont été développés. En particulier, un dispositif haptique permet d'appréhender la structure d'un document à partir du toucher sur une tablette tactile [30] (voir figure ci-dessous). Parallèlement, une transposition orale de la structure visuelle des contenus textuels est possible grâce à une architecture TTS concurrente [1].

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 [28]. Ces modèles ont ensuite été améliorés par l'apport de différentes stratégies d'apprentissage multitâches, notamment par le couplage classification/régression [29] et le couplage régression/classification de la dépression/régression des émotions [27].

Références

[1] 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.

[2] 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.

[3] Cleuziou G. and Dias G. Learning pretopological spaces for lexical taxonomy acquisition. In *Proceedings of the European Confe-*

rence on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2015), 2015.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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*, pages 482–486, 2018.

[10] 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.

[11] 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*, pages 111–120, 2017.

[12] 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*, pages 782–792, 2011.

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

[14] 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. [CORE=A+].

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

[24] 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*, pages 93–94, 2015.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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*, 2019.

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

[32] 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 : 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

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 19 millions d'heures de contenus audiovisuels, il assure le dépôt légal de la radio, de la télévision et du web média, 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 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 usagers (internes et externes) et des clients 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 clients (journalistes, producteurs) et les usagers (chercheurs, grand public). Les thématiques de recherche couvrent plusieurs champs disciplinaires : 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.e.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 et l'analyse des locuteurs sont indispensables pour mener à bien les travaux de recherche sur l'analyse et la compréhension des médias. Ces thématiques se trouvent au cœur de plusieurs des projets de recherche de l'Institut.

Ainsi, développé dans le cadre du projet H2020 MeMAD « Methods for Managing Audiovisual Data », le logiciel open source InaSpeechSegmenter permet la détection de la parole, de la musique et du genre des locuteurs. Le logiciel a été appliqué à un corpus d'un million d'heures de télévision et de radio pour mesurer l'évolution du taux d'apparition des femmes dans les médias. Dans la continuité de cette étude, le projet ANR *Gender Equality Monitor*, mené en partenariat avec le LIUM, le LIMSI, le CARISM, le LERASS, le Centre Max Weber et la société Deezer, a pour objectif d'accomplir la plus vaste étude sur la place des hommes et des femmes dans les médias jamais réalisée, fondée sur l'analyse de plusieurs millions de documents échantillonnés sur une période de plus de 80 ans.

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 utilisateurs ainsi que des portails web entièrement paramétrables.

Références

[1] E. 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 Numerique*, 20, 2018.

[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, 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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

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

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

- Leonor BECERRA-BONACHE (MCF)
- Marc BERNARD (MCF)
- Mathias GERY (MCF)
- Christophe GRAVIER (MCF HDR)
- Amaury HABRARD (PR)
- François JACQUENET (PR)
- Charlotte LACLAU (MCF)
- Christine LARGERON (PR)
- Pierre MARET (PR)
- Fabrice MUHLENBACH (MCF)
- Julien SUBERCAZE (MCF)

Contexte

Le laboratoire Hubert Curien est un laboratoire pluridisciplinaire qui développe une activité de recherche en informatique qui sera principalement regroupée au sein de l'équipe Data Intelligence à l'horizon 2021. 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 du laboratoire, 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 [19]. 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 [21] ainsi qu'à leur représentation par vecteurs de bits [22]. Nous avons également développé des techniques à base d'apprentissage profond pour la production automatisée de résumés de textes [23, 15, 13] ou la génération automatique de questions [11].

Nous intégrons des techniques du web sémantique dans des travaux en *question answering* [8], 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 [9]. 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 [18], modèles de langages personnalisés [2], combinaison de données de type texte, de variables descriptives et d'informations issues de traitement du signal [17]) à développer des systèmes plus efficaces [20, 4], permettant de découvrir des relations avec un rappel élevé [10] 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 in-

attendue dans des corpus de textes [14], 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 [16] ou d'identifier automatiquement des auteurs d'articles [12].

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, 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.

[9] 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.

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

[11] 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.

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

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

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

[15] Lucie-Aimée Kaffee, Hady ElSahar, Pavlos Vougiouklis, Christophe Gravier, et al. Learning to generate wikipedia summaries for un-

derserved languages from wikidata. In *Proc. of NAACL-HLT*, pages 640–645, 2018.

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

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

[18] 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.

[19] 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.

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

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

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

[23] 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

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@ens.fr

Membres

- Pascal AMSILI (PU P3)
- Frédéric LANDRAGIN (DR CNRS)
- Frédérique MÉLANIE-BECQUET (IE CNRS)
- Clément PLANCQ (IE CNRS)
- Thierry POIBEAU (DR CNRS)

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) le développement de techniques fondamentales pour le TAL, notamment dans le domaine de l'apprentissage artificiel, ii) la mise en application de ces techniques, pour les humanités numériques, et iii) une réflexion historique et épistémologique sur le TAL.

Aperçu des recherches

Ces dernières années ont été marquées par l'émergence extrêmement rapide de nouvelles méthodes d'apprentissage (notamment les approches neuronales, dites aussi d'*apprentissage profond*). L'application de ces méthodes à certaines tâches de TAL ont permis d'obtenir des progrès très conséquents. Il est donc important de les comprendre, de les tester et de les faire évoluer quand nécessaire.

Analyse syntaxique à large couverture. Nous nous sommes particulièrement intéressés à l'analyse syntaxique ces dernières années. Le développement conjoint des méthodes d'apprentissage artificiel déjà évoquées et surtout la mise à disposition de corpus annotés pour plus de 60 langues dans un même format (*universal dependencies*) ont permis la mise au point d'analyseurs performants pour de nombreuses langues, en un temps extrêmement réduit. Le laboratoire a ainsi pu participer aux campagnes d'évaluation CoNLL 2017 et 2018, avec des résultats très encourageants, au niveau des meilleures équipes internationales [3]. Les enjeux consistent aujourd'hui à traiter des langues complexes (notamment des langues à morphologie riche) pour lesquelles on dispose de peu de données. Des résultats intéressants ont été obtenus pour des langues finno-ougriennes comme le komi, pour lesquelles on ne disposait pas d'outils jusqu'ici [2]. Des techniques d'analyse multilingues ont dû être mises au point, afin de contourner la question de la rareté des données face à des techniques nécessitant d'ordinaire des corpus annotés de très grande taille pour fournir des résultats fiables.

TAL et humanités numériques. L'application des techniques de TAL au domaine des Humanités numériques est un domaine d'investigation particulièrement prometteur. Il oblige à tester les techniques de TAL sur des corpus réels, souvent difficiles, car touchant à des états de langue anciens et/ou à des questions de recherche complexes. Jusqu'ici le laboratoire a surtout travaillé sur des cas liés aux sciences sociales (négociations climatiques [5], corpus PoliInformatics) et au domaine philosophie (projet Mapping Bentham). Les recherches plus récentes portent sur l'exploration de techniques pour l'analyse (semi-)automatique du discours littéraire. On s'intéresse par exemple à la détection automatique 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 [1]. D'autres

éléments du domaine littéraire seront étudiés dans les années à venir, comme la notion de suspense ou de romans à succès (peut-on détecter le suspense ? Peut-on prédire le succès d'un roman ?).

Épistémologie du TAL. Les outils de TAL se répandent de plus en plus dans le monde courant, aussi bien dans les milieux professionnels qu'auprès du grand public. C'est particulièrement le cas de l'application emblématique du TAL à savoir la traduction automatique [4]. L'évolution des usages de cette technologie est particulièrement intéressante à observer et pose de multiples questions, d'acceptabilité, d'utilisabilité et, tout simplement, de mise en concurrence avec l'humain, en l'occurrence avec les traducteurs professionnels. Il s'agit d'un cas emblématique pour l'étude des relations entre technique et société.

Références

[1] 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.

[2] 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.

[3] KyungTae Lim and Thierry Poibeau. A system for multilingual dependency parsing based on bidirectional LSTM feature representations. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 63–70, 2017.

[4] Thierry Poibeau. *Babel 2.0. Où va la traduction automatique ?* Odile Jacob, 2019.

[5] Pablo Ruiz, Clément Plancq, and Thierry Poibeau. More than word cooccurrence : Exploring support and opposition in international climate negotiations with semantic parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2016.

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
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 [31] 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 [14] ou d'extraction de concepts sémantiques [41]. 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 [7].

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 [27]. 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 [11, 37].
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 [10]. La méthode étendue permet la prise en compte d'informations situées, obtenues par d'autres modalités [12], ou une gestion plus naturelle des tours de parole [20].
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 [38]. 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 [36].

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, 26]. 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 [25], 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, 6]. 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 [23, 24], (2) la modélisation par des i-vecteurs de productions de parole altérée pour une tâche de prédiction du niveau d'intelligibilité [21, 22]. 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 [15]. L'apport du LIA est ici de fournir des outils automatiques adaptés à ce protocole particulier [13]. Pour finir, le LIA a initié récemment des travaux impliquant des approches de *deep learning* dans le cadre du projet ANR/RUGBI portant sur la caractérisation de l'intelligibilité en partenariat avec les laboratoires IRIT et Lordat-Octogone, le CHU de Toulouse et le LPL.

Réseaux sociaux, réseaux complexes et TAL

Le web, et plus spécifiquement les média 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 [30], la détection d'opinions [39, 9], ou encore l'analyse temporelle du contenu de messages courts [35]. 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, des nombreuses erreurs grammaticales et orthographiques, et des contenus pouvant être très courts.

Ces difficultés rendent alors le traitement automatique du contenu textuel compliqué. Des derniers travaux entrepris au LIA ont notamment mon-

tré, 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 [32] que la modélisation des échanges entre utilisateurs (*i.e.* qui parle à qui), sans tenir compte de leur contenu, permet 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 [8].

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.

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 qui vient de démarrer.

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 [46], fondé sur les CRF (*conditional random fields*), des algorithmes classiques d'ingénierie documentaire et un système de règles.

Résumé automatique

Soit dans sa forme textuelle [43], soit multimédia [19, 40], 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, 45, 42].

Le résumé automatique multimédia et multilingue a fait l'objet du projet CHIST-ERA/ANR-AMIS [33, 34]. Ce projet a donné l'opportunité de traiter la langue arabe standard, l'anglais et le français en développant des techniques de résumé automatique basées sur du texte et la parole [16].

Le résumé automatique de texte est très lié à la compression de phrases [28, 29] 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 [17, 18].

L'évaluation de la qualité des résumés, sujet controversé dû à son subjectivité, a également été étudiée au LIA [44].

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. un 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] Xavier Bost, Serigne Gueye, Vincent Labatut, Martha Larson, Georges Linarès, Damien Malinas, and Raphaël Roth. Remembering winter was coming : Character-oriented video summaries of tv series. *Multimedia Tools and Applications*, in press, 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] 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.

[7] Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Camelin, and Yannick Estève. Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *Interspeech*, 2019.

[8] Noé Cecillon, Vincent Labatut, Richard Dufour, and Georges Linarès. Abusive language detection in online conversations by combining content-and graph-based features. *Frontiers in Big Data*, 2 :8, 2019.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Imed Laaridh, Christine Meunier, and Corinne Fredouille. Perceptual evaluation for automa-

tic anomaly detection in disordered speech : Focus on ambiguous cases. *Speech Communication*, 105 :23–33, 2018.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linarès. Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems*, 6(1) :38–55, 2019.

[33] 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.

[34] 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.

[35] Mathias Quillot, Cassandre Ollivier, Richard Dufour, and Vincent Labatut. Exploring temporal analysis of tweet content from cultural events. In *International Conference on Statistical Language and Speech Processing*, pages 82–93. Springer, 2017.

[36] 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.

[37] 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.

[38] 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.

[39] Mickael Rouvier and Benoit Favre. Senseilif 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.

[40] 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.

[41] Natalia Tomashenko, Antoine Caubrière, and Yannick Estève. Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech. In *Interspeech*, 2019.

[42] Juan-Manuel Torres-Moreno. Artex is another text summarizer. *arXiv preprint arXiv :1210.3312*, 2012.

[43] Juan-Manuel Torres-Moreno. *Automatic Text Summarization*, volume 1. John Wiley & Sons, 2014.

[44] 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.

[45] Juan-Manuel Torres-Moreno, Patricia Velázquez-Morales, and Jean-Guy Meunier. Cortex : un algorithme pour la condensation automatique de textes. *ARCo*, 2 :365, 2001.

[46] 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

Agata SAVARY
agata.savary@univ-tours.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 syn-

taxiques, 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 deux premiers systèmes francophones de résolution de la coréférence, entraîné sur le corpus ANCOR. Ce travail se poursuit désormais dans le cadre du projet ANR DEMOCRAT. De même, 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 coréférence.

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 laboratoire 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

LIMSI : Sciences et Technologies de la Langue

LIMSI UPR 3251 / ILES
CNRS, Université Paris-Saclay
https://www.limsi.fr/fr/recherche/iles

Pierre ZWEIGENBAUM
Responsable du groupe ILES
pz@limsi.fr

LIMSI UPR 3251 / TLP
CNRS, Université Paris-Saclay
https://www.limsi.fr/fr/recherche/tlp

Jean-Luc GAUVAIN
Responsable du groupe TLP
gauvain@limsi.fr

Membres au 1/1/2020

Permanents (ILES)

- Annelies BRAFFORT (CNRS)
- Michael FILHOL (CNRS)
- Sahar GHANNAY (Université Paris-Sud)
- Cyril GROUIN (CNRS)
- Thierry HAMON (Université Paris-Nord)
- Gabriel ILLOUZ(Université Paris-Sud)
- Thomas LAVERGNE (Université Paris-Sud)
- Anne-Laure LIGOZAT (ENSIIE)
- Aurélie NÉVÉOL (CNRS)
- Patrick PAROUBEK (CNRS)
- Sophie ROSSET (CNRS)
- Anne VILNAT (Université Paris-Sud)
- Pierre ZWEIGENBAUM (CNRS)

Permanents (TLP)

- Gilles ADDA (CNRS)
- Philippe BOULA DE MAREÜIL (CNRS)
- Hervé BREDIN (CNRS)
- Caio CORRO (Université Paris-Sud)
- Laurence DEVILLERS (Univ. Paris-Sorbonne)
- Jean-Luc GAUVAIN (CNRS)
- Camille GUINAUDEAU (Université Paris-Sud)
- Lori LAMEL (CNRS)
- Jean-Sylvain LIÉNARD (CNRS)
- Joseph-Jean MARIANI (CNRS)
- Hélène MAYNARD (Université Paris-Sud)
- Albert RILLIARD (CNRS)
- Ioana VASILESCU (CNRS)
- François YVON (CNRS)

Permanents (CPU)

- Elise PRIGENT (Université Paris-Sud)

Permanents (AMI)

- Michèle GOUIFFÈS (Université Paris-Sud)

Doctorants, Université Paris-Saclay : *ED STIC* : ILES : Alexandra BENAMAR, Hugo BOULANGER, Oralie CATTAN, Juan Manuel CORIA, Hicham EL BOUKKOURI, Marion KACZMAREK, Corentin MASSON, Nicolas PARIS, Tsanta RANDRIATSITOHAINA, Léon-Paul SCHAUB, Mathilde VERON, Yuming ZHAI, TLP : Hugues Ali MEHENNI, Marc BENZAHRA, Aman Zaid BERHE, François BUET, Caroline ÉTIENNE, Aina Gari SOLER, Margot LACOUR, Benjamin MAURICE, Syrielle MONTARIOL, Anh Khoa NGO HO, Minh Quang PHAM, José Carlos ROSALES, Jitao XU, AMI et ILES : Hannah BULL, Hussein CHAABAN, ILES et AMI : Valentin BELISSEN, ILES et CPU : Félix BIGAND, ILES et TLP : Léo GALMANT, *Doctorants, Sorbonne Université*, *ED Santé publique* : ILES : Antoine NEURAZ.
Contractuels et post-doctorants : ILES : Sadaf Abdul RAUF, Somnath BANERJEE, TLP : Mathilde HUTIN, Paul LERNER. .

Introduction

Le département Sciences et Technologies de la Langue (STL) regroupe les activités du LIMSI en *traitement de la langue* au sein des groupes ILES et TLP. Le spectre des recherches conduit dans ces deux groupes est très large et s'étend à tous les aspects de la langue, *orale, écrite ou signée*.

Le groupe ILES (Information, Langue Écrite et Signée) se centre sur l'étude de la langue écrite ou signée, aussi bien dans ses fonctions de communication 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 le groupe 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 technolo- giques 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-Sud) et en perception visuelle (en collaboration avec le groupe CPU) ; (2) l'élaboration de ressources linguistiques et d'outils permettant de manipuler ces ressources (*i.e.* aide à l'annota- tion par traitement d'images) ; et (3) les principaux thèmes de recherche en traitement automatique de la LSF : la reconnaissance (en collaboration avec le groupe AMI), 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 présente 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'exploi-

tation 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 (HDR 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 récentes 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 l'utilisation de ces systèmes dans les interactions avec des robots. Le deuxième axe porte sur l'interaction affective avec des machines 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 bots ou des robots.

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). 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 langues en danger. Les travaux en apprentissage automatique se développent dans deux directions principales. Elles s'intéressent d'une part à des techniques variées pour aborder des tâches supervisées d'apprentissage structuré « de bas niveau » (normalisation, étiquetage en parties du discours ou en *chunks*, segmentation morphologique, *parsing*), avec comme ambition de développer des méthodes, par exemple en matière d'adaptation au domaine, qui pourront ensuite être transférées à des problèmes de TA. Elles explorent d'autre part des tâches sémantiques principalement non supervisées (détection de sens, repérage de paraphrases, etc), ici encore avec un focus particulier sur le cadre multilingue, soit qu'il fournisse un contexte facilitateur (détection de sens dans des bitextes), soit qu'il corresponde à l'application finale visée (paraphrases multilingues).

Ces activités impliquent des collaborations resserrées avec le thème « Multilinguisme et paraphase », 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 le Web ou dans des entreprises, 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.

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 d'éva-

luations 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] Patrice Bellot, Véronique Moriceau, Josiane Mothe, Eric SAN JUAN, and Xavier Tannier. INEX Tweet Contextualization Task : Evaluation, Results and Lesson Learned. *Information Processing and Management*, 52(5) :801–819, 2016.

[2] 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.

[3] Houda Bouamor, Aurélien Max, and Anne Vilnat. Multitechnique paraphrase alignment : A contribution to pinpointing sub-sentential paraphrases. *ACM TIST*, 4 :44 :1–44 :27, 2013.

[4] Philippe Boula De Mareüil, Albert Rilliard, Fanny Ivent, and V. Kozhevina. A comparative prosodic study of questions in French in contact with Occitan and Catalan. *Journal of Speech Sciences*, 4(2) :59–72, 2014.

[5] Dominique Boutet and Marion Blondel. Les corpus de Langue des Signes Française. In Annelies Braffort, editor, *La Langue des Signes Française (LSF) : modélisations, ressources et applications*, pages 47–85. ISTE, 2016.

[6] Franck Burlot and François Yvon. Learning Morphological Normalization for Translation from and into Morphologically Rich Languages. *The Prague Bulletin of Mathematical Linguistics*, 108 :49–60, 2017.

[7] 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.

[8] 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.

[9] Laurence Devillers, Marie Tahon, Mohamed El Amine Sehili, and Agnes Delaborde. Inference of Human Beings' Emotional States from Speech in Human-Robot Interactions. *International Journal of Social Robotics*, 2015.

[10] Michael Filhol, Mohamed Hadjadj, and Benoît Testu. A rule triggering system for automatic text-to-Sign translation. In *International workshop on Sign Language translation and avatar technology (SLTAT)*, Chicago, United States, 2013.

[11] Karën Fort and Maxime Amblard. Éthique et traitement automatique des langues. In *Journée éthique et intelligence artificielle*, Nancy, France, 2018.

[12] Souhir Gahbiche-Braham, Hélène Maynard, and François Yvon. Traitement automatique des entités nommées en arabe : détection et traduction. *Traitement Automatique des Langues (TAL)*, 54 :101–132, 2014.

[13] 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.

[14] Brigitte Grau, Anne-Laure Ligozat, and Martin Gleize. Recherche d'information précise dans des sources d'information structurées et non structurées : défis, approches et hybridation. *Traitement Automatique des Langues*, 56(3), 2015.

[15] Cyril Grouin, Véronique Moriceau, and Pierre Zweigenbaum. Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records. *Journal of Biomedical Informatics*, 58 :S133–S142, 2015.

[16] Thierry Hamon, Natalia Grabar, and Fleur Mougin. Querying biomedical Linked Data with natural language questions. *Open Journal Of Semantic Web*, 0 :1 – 19, 2016.

[17] 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.

[18] 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.

[19] Thomas Lavergne, Cyril Grouin, and Pierre Zweigenbaum. The contribution of coreference resolution to supervised relation detection between bacteria and biotopes entities. In *BMC Bioinformatics*, 2015.

[20] Hai Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon. Structured output layer neural network language models for speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 21 :197–206, 2013.

[21] Vincent Letard, Sophie Rosset, and Gabriel Illouz. A mapping-based approach for general formal human computer interaction using natural language. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 34–40, June 2014.

[22] Jean-Sylvain Liénard and Claude Barras. Fine-grain voice strength estimation from vowel spectral cues. In *Interspeech 2013*, Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)., Lyon, France, 2013. ISCA.

[23] Anne-Laure Ligozat. Question classification transfer. In *Proceedings of the Association for Computational Linguistics (ACL shorts papers)*, 2013.

[24] Rasa Lileikyté, Lori Lamel, Jean-Luc Gauvain, and Arseniy Gorin. Conversational telephone speech recognition for Lithuanian. *Computer Speech and Language*, 49 :71–82, 2018.

[25] 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.

[26] 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.

[27] Diana McCarthy, Marianna Apidianaki, and Katrin Erk. Word Sense Clustering and Clusterability. *Computational Linguistics*, 42 :245–275, 2016.

[28] Nicolas Pécheux, Guillaume Wisniewski, and François Yvon. Reassessing the value of resources for cross-lingual transfer of POS tagging models. *Language Resources and Evaluation*, 50 :1–34, 2016.

[29] Achintya Sarkar, Cong-Thanh Do, Viet-Bac Le, and Claude Barras. Combination of Cepstral and Phonetically Discriminative Features for Speaker Verification. *IEEE Signal Processing Letters*, 21(9) :1040 – 1044, 2014.

[30] 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.

AfIA
Association française
pour l'Intelligence Artificielle

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 au 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. 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. 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].

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 [8, 11].

En *extraction d'information*, nous nous intéressons 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 » [10], 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'apprentissage (en résumé automatique [9]) et capables de comprendre la hiérarchie des informations structurées pour d'identifier les éléments saillants à retranscrire (en *data-to-text*).

La *propagation d'information* sur les réseaux sociaux est au coeur de nombreuses recherches en apprentissage statistique.

Nous nous focalisons sur l'apprentissage de re-

présentations 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 [7] 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 [12]. 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 [14], 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 [13].

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 *Proceedings of the 7th ACM international conference on Web search and data mining*, 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 *Proceedings of the Ninth ACM international conference on Web Search and Data Mining*, 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] Sylvain Lamprier. A recurrent neural cascade-based model for continuous-time diffusion. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3632–3641, 2019.

[8] 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.

[9] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Answers Unite ! Unsupervised Metrics for Reinforced Summarization Models. In *EMNLP*, 2019.

[10] Etienne Simon, Vincent Guigue, and Benjamin Piwowarski. Unsupervised Information Extraction : Regularizing Discriminative Approaches with Relation Distribution Losses. In *ACL*, 2019.

[11] 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.

[12] Hannes Werthner, Markus Zanker, Jennifer Golbeck, and Giovanni Semeraro, editors. *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 2015.

[13] Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Context-Aware Zero-Shot Learning for Object Recognition. In *ICML*, 2019.

[14] Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Learning Multi-Modal Word Representation Grounded in Visual Context. In *AAAI*, 2018.

AfIA
Association française
pour l'Intelligence Artificielle

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)
- Vincent COLOTTE (MCF)
- Irina ILLINA (MCF)
- Odile MELLA (MCF)
- Slim OUNI (MCF)
- Agnès PIQUARD (MCF)
- Romain SERIZEL (MCF)
- Md SAHIDULLAH (chercheur)
- Élodie GAUTHIER (post-doctorante)
- Manfred PASTÄTTER (post-doctorant)
- Imran SHEIKH (post-doctorant)
- Théo BIASUTTO-LERVAT (doctorant)
- Tulika BOSE (doctorante)
- Guillaume CARBAJAL (doctorant)
- Pierre CHAMPION (doctorant)
- Sara DAHMANI (doctorante)
- Ioannis DOUROS (doctorant)
- Sandipana DOWERAH (doctorante)
- Adrien DUFRAUX (doctorant)
- Raphaël DUROSELLE (doctorant)
- Nicolas FURNON (doctorant)
- Ashwin GEET D'SA (doctorant)
- Ajinkya KULKARNI (doctorant)
- Michel OLVERA ZAMBRANO (doctorant)
- Manuel PARIENTE (doctorant)
- Shakeel Ahmad SHEIKH (doctorant)
- Sunit SIVASANKARAN (doctorant)
- Nicolas TURPAULT (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 mo-

dalités 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 handicap [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

SISO : Système d'Information Spatialisé, Modélisation, Extraction et Diffusion des Données et Connaissances

TETIS UMR 9000 / SISO
AgroParisTech, CIRAD, CNRS, INRAE
https://umr-tetis.fr

Mathieu ROCHE
mathieu.roche@cirad.fr

Membres :

- Vincent BONNAL (CIRAD)
- Rémy DECOUPES (INRAE)
- Hugo DELEGLISE (doctorant)
- Vincent DOUZAL (INRAE)
- Sophie FORTUNO (CIRAD)
- Jean Eudes GBODJO (doctorant)
- Dino IENCO (INRAE)
- Roberto INTERDONATO (CIRAD)
- Rodrique KAFANDO (doctorant)
- Urcel KALENGA (doctorant)
- Martin LENTSCHAT (doctorant)
- Jérôme PASQUET (Univ. Paul Valéry)
- Mathieu ROCHE (CIRAD)
- Lucile SAUTOT (AgroParisTech)
- Maguelonne TEISSEIRE (INRAE)
- Sarah VALENTIN (doctorante)

Introduction

L'objectif de l'équipe SISO 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.

Activités

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 SISO 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, Human-Num, Ortolang).

Extraction d'entités spatiales et thématiques

Une partie des travaux de l'équipe SIO consiste à proposer de nouvelles méthodes d'identification des entités spatiales (absolues et relatives) à partir de corpus peu standardisés [9] dans les domaines de l'agronomie [3] et de l'épidémiologie [1]. Ces informations spatiales peuvent être désambiguisées par des méthodes d'apprentissage supervisé et d'apprentissage actif. 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 [6], le labelling [8] et l'induction d'informations sémantiques [7]. Certaines méthodes mises en œuvre reposent sur la définition de nouveaux descripteurs linguistiques adossés à des méthodes d'apprentissage supervisé.

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.*) [4].

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]. Ainsi, des logiciels de veille en épidémiologie animale (PADI-Web, EpidNews, EpidVis) accompagnés de nouvelles visualisations ont été produits [1, 5].

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] Brett Drury and Mathieu Roche. A survey of the applications of text mining for agriculture. *Computers and Electronics in Agriculture*, 163 :104864, 2019.

[4] Jacques Fize, Mathieu Roche, and Maguelonne Teisseire. Mapping heterogeneous textual data : A multidimensional approach based on spatiality and theme. In *Proc. Internet Science*, pages 310–317, 2019.

[5] 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*, 2019.

[6] 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.

[7] 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.

[8] 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.

[9] 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

SMART : Speech Modelisation and Text, Statistical Machine Translation

LORIA UMR 7503 / SMaRT
CNRS, INRIA et Université de Lorraine
smart.loria.fr

Kamel SMAILI
kamel.smaili@loria.fr

David LANGLOIS
david.langlois@loria.fr

Joseph DI MARTINO
jdm@loria.fr

Membres impliqués

- Kamel SMAÏLI (PR)
- David LANGLOIS (MCF)
- Joseph DI MARTINO (MCF)
- Salima HARRAT (chercheuse associée)
- Karima MEFTOUH (chercheuse associée)
- Mohamed Amine MENACER (doctorant)
- Karima ABIDI (doctorante)
- Nouha OTHMAN (ATER)

Mot-clés

- multilinguisme
- 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

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- : Mohamed Amine MENACER, Traduction automatique de vidéos en dialecte arabe.
- 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'équipe SMaRT cherche à proposer des méthodes pour le traitement automatique des données langagières multilingues aux niveaux textuel et oral. Nos travaux ont entre autres vocation à améliorer les résultats en traduction automatique et en reconnaissance automatique de la parole. Notre approche est celle de l'apprentissage automatique fondée sur des méthodes mathématiques pour identifier, extraire et proposer des associations entre des éléments de deux ou plusieurs langues. Les langues sont étudiées en utilisant des corpus monolingues, et multilingues. Ces corpus peuvent être parallèles ou comparables. Les méthodes utilisées ne sont pas dépendantes de la langue, mais notre équipe se concentre sur le français, l'anglais, et surtout sur

l'arabe, particulièrement sur les dialectes arabes. SMaRT contribue à la recherche dans le domaine du multilingue selon plusieurs axes : construction de corpus, traduction automatique (tables de traduction, moteur de traduction, estimation de qualité et mesures de confiance), 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.

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.

L'équipe s'intéressant donc à la traduction automatique, mais plus globalement au multilinguisme, au niveau du texte et de la parole, le sigle SMaRT peut alors être lu de deux manières : **S***peech* **M***odelisation and* **T***ext* et **S***tatistical* **M***achine* **Tra***nslation* (ou encore **S***peech* **M***achine* **Tra***nslation*).

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) [6] 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) [2], 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 [4].

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 [3].

En ce qui concerne les modèles de traduction, nous avons proposé une méthode originale de construction de table de traduction. L'idée était de ne pas utiliser la notion d'alignement qui était classiquement utilisée, mais d'extraire les relations entre les mots et les séquences via des mesures d'information mutuelle [5].

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 conférence 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 [8]), 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 [4], et nous avons proposé une approche utilisant la théorie de l'*appraisal* [1] 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 [7]. 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 [9] (**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.
- 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, 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 entamons actuellement des travaux sur la détection de *fake news* et sur le *multitask learning* pour apprendre à traduire les dialectes.

Références

[1] K. Abidi, D. Fohr, D. Jouvet, D. Langlois, O. Mella, and K. Smaïli. A Fine-grained Multilingual Analysis Based on the Appraisal Theory : Application to Arabic and English Videos. In *Arabic Language Processing : From Theory to Practice. 7th International Conference, ICALP 2019*, volume Communications in Computer and Information Science book series (CCIS, volume 1108), pages 49–61. 2019.

[2] 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.

[3] A. Douib, D. Langlois, and K. Smaïli. Genetic-based Decoder for Statistical Machine Transla-

tion. In *Springer LNCS series, Lecture Notes in Computer Science*. 2016.

[4] D. Langlois, M. Saad, and K Smaïli. Alignment of comparable documents : comparison of similarity measures on French-English-Arabic data. *Natural Language Engineering*, 2018.

[5] C. Latiri, K. Smaïli, C. Lavecchia, C. Nasri, and D. Langlois. Phrase-based Machine Translation based on Text Mining and Statistical Language Modeling Techniques . *International Journal of Computational Linguistics and Applications*, 2(1-2) :16, 2011.

[6] K. Meftouh, S. Harrat, and K. Smaïli. PADIC : extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey, 2018.

[7] 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.

[8] 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.

[9] 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

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 (DR 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 grande 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.

Depuis plusieurs années l'équipe utilise les outils formels de description du langage naturel, en particulier les grammaires d'arbres adjoints lexicalisées, qui côtoient souvent dans nos recherches

les méthodes statistiques d'apprentissage automatique, en particulier les modèles bayésiens et les 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 travaille depuis plusieurs années 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].

Enfin, 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]. Plus récemment, nous avons développé des approches neuronales de type encodeur-décodeur pour la génération de texte à partir de représentations sémantiques abstraites [12], d'arbres de dépendances [13] 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 représentations 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

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

- Denis BÉCHET (MCF)
- Florian BOUDIN (MCF)
- Mérième BOUHANDI (doctorante)
- Béatrice DAILLE (PR)
- Victor CONNES (doctorant)
- Colin DE LA HIGUERA (PR)
- Chantal ENGUEHARD (MCF)
- Kévin ESPASA (doctorant)
- Ygor GALLINA (doctorant)
- Amir HAZEM (post-doctorant)
- Nicolas HERNANDEZ (MCF)
- Christine JACQUIN (MCF)
- Martin LAVILLE (doctorant)
- Jingshu LIU (doctorant)
- Laura MONCEAUX-CACHARD (MCF)
- Emmanuel MORIN (PR)
- Solen QUINIOU (MCF)

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 deux 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 comme les blogs, les réseaux sociaux, les forums, se couplant à d'autres média ou encore s'exprimant dans des langues différentes. Les travaux en TALN sont par nature multidisciplinaires, au cœur des données, en interaction avec les sciences humaines et sociales (linguistique, terminologie, traduction, sciences de l'éducation et avec d'autres thématiques de l'informatique comme l'apprentissage, la fouille de données, la reconnaissance du signal (parole, geste), la recherche d'informations, *etc.*

Thématiques

Les travaux de l'équipe portent sur l'analyse de la langue écrite et relèvent de deux thématiques principales de recherche :

Analyse & Découverte. L'analyse s'intéresse aux modèles formels de la syntaxe et de la sémantique des langues. Nous travaillons sur des grammaires lexicalisées permettant une analyse syntaxique en dépendance et sur des grammaires probabilistes. La Découverte applique diverses méthodes d'analyses sur les corpus de données textuelles pour isoler des éléments remarquables. L'équipe a une forte expertise dans le traitement de documents appartenant à des domaines spécialisés.

Alignement & Multilinguisme. Nous étudions des méthodes de rapprochement de diverses sources de données pour pouvoir bénéficier d'informations complémentaires : les alignements. Nous travaillons sur les alignements de corpus comparables, des textes dans deux langues sans rapport de traduction, des corpus multimodaux, des textes provenant de la transciprion de l'oral ou de l'écriture manuscrite et des textes écrits. Par ailleurs, des ressources lexicales de langues peu dotées ont été créées et mises à disposition sous licence Creative Commons.

Domaines applicatifs

Les domaines applicatifs de l'équipe concernent :

Aide à la traduction. Dans le domaine de l'aide à la traduction, nous cherchons à produire des lexiques bilingues rendant compte des connaissances véhiculées dans les domaines techniques. Ces lexiques fournissent des termes alignés entre

deux langues accompagnés d'exemples d'utilisation. Ils servent notamment aux traducteurs professionnels dans leur travail de révision d'une traduction. Ces mêmes lexiques sont aussi des ressources exploitées dans d'autres applications du TALN reposant sur l'exploitation de lexiques lorsqu'il est nécessaire de déployer une application vers de nouvelles langues comme en fouille d'opinions. Ces lexiques sont acquis à travers des méthodes supervisées ou non, reposant sur une analyse distributionnelle.

Ingénierie des ressources éducatives. En ce qui concerne l'ingénierie des ressources éducatives, nous nous intéressons à la création d'outils permettant de faciliter l'apprentissage des apprenants. Cela passe par de l'aide à la navigation dans des ressources éducatives libres (multilingues, différentes thématiques, différents lieux physiques, différents types de support, *etc.*), de la recherche automatique de ressources complémentaires en lien avec un cours donné par un enseignant ou à la segmentation automatique de ce cours selon les thématiques abordées. Nous nous intéressons également à l'inclusion de publics porteurs de handicap, en particulier les enfants atteints de troubles dys, pour lesquels nous cherchons à identifier les difficultés présentes dans des textes afin de les expliciter. Les enjeux concernent l'évaluation automatique de la qualité, la mesure et l'explicitation de la difficulté, l'antériorité d'un cours, l'identification de concepts ou thématiques.

Analyse sémantique des réseaux sociaux. Dans le cadre de l'analyse sémantique des réseaux sociaux, nos travaux contribuent à la description de plusieurs aspects des communications et des interactions humaines médiées par des réseaux (blog, tchat, forums, *etc.*). Ils concernent aussi bien la mesure de l'engagement de contributeurs (fouille d'opinion, sentiment, émotion), la reconnaissance des intentions de ces derniers (actes du dialogue) ainsi que la modélisation des interactions. Les bibliothèques logicielles que nous avons développées permettent de soutenir des analyses globales au niveau de la conversation (ou d'un corpus de conversation) telles que la détection de situations critiques (par exemple, une conversation qui s'envenime) ou bien l'étude de tendance chez certains publics. Les domaines applicatifs visés sont ceux de la gestion de la relation client et de l'apprentissage en ligne. Nos efforts touchent aussi bien la production de ressources (corpus annotés et lexiques) que le développement de systèmes de reconnaissance automatique fondés sur les méthodes d'apprentissage supervisé à base de réseaux de neurones profonds et d'apprentissage non supervisé exploitant des modélisations à base de graphes.

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/