



RAPPORT D'ACTIVITES A L'ISSUE
DE LA 1^{ère} ANNEE

**Si ce n'est déjà fait, pensez à joindre
la dernière attestation d'inscription en doctorat (2^{ème} ou 3^{ème} année)**

Merci de présenter succinctement les travaux réalisés au cours de l'année écoulée. Préciser l'avancement des recherches par rapport au planning initial, l'évolution par rapport au projet initial et les éventuels problèmes rencontrés (intégration à l'équipe de recherche, adaptation aux contraintes du milieu industriel, ...).

Ma thèse CIFRE portant sur la « **Constitution de fils de discussion cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration.** »

Cette première année de thèse s'est articulée autour de l'état de l'art, de constitution de corpus de travail constitué de mails et de premières analyses de ce corpus.

Au cours de la première année, j'ai effectué une revue précise de la littérature scientifique autour des problématiques plus ou moins proches de celles identifiées dans la cadre des objectifs généraux de ma thèse. Ces lectures d'environ une trentaine d'articles portaient essentiellement sur l'analyse de corpus conversationnels (emails, chats, forum, etc.) avec un accent particulier sur ce qui touche à l'identification et la classification d'actes de dialogues, de segmentation de texte, de démêlage de fils de conversations et d'analyse de discours. Ces problématiques ont été abordées par différentes méthodes partant d'algorithmes classiques de machine Learning jusqu'à des approches plus complexes de deep learning incluant des niveaux contextuels différents. Une autre thématique abordée dans ces lectures est la constitution de corpus avec la prise en compte du respect du RGPD (Règlement Général sur la Protection des Données) et le droit au secret de la correspondance. Ces différentes lectures ont donné lieu à mon état de l'art en cours de rédaction avec un plan détaillé et complet et constituent, pour nombre de ces articles, des bases solides pour l'avancement de mes travaux. De ces lectures aussi, j'ai pu tirer l'organisation générale de rédaction d'articles scientifiques.

J'ai eu des séances de travail en laboratoire avec mes encadrants pour des premières ébauches d'analyses de données sur un corpus appelé Webis Gmane Email Corpus 2019. Ces premières analyses de corpus ont consisté à construire des relations entre les emails en langue française extraits du corpus. Il a été prévu d'indexer des données pour des recherches textuelles et de concevoir une time-line de présentation des emails avec des clusters de conversations. Des premières expérimentations ont été faites par la suite pour la pseudo-anonymisation des données.

Afin de travailler sur des données réelles d'entreprise, j'ai développé un outil qui permet l'identification et l'optimisation du nombre d'interlocuteurs ainsi que les emails et conversations dans lesquels ces interlocuteurs sont impliqués. Ce même outil permet l'extraction des données

réelles qui vont tour à tour être pseudo-anonymisées, analysées afin de détecter les caractéristiques de corpus qui aideront plus tard à une conception de modèle d'intelligence artificielle pour répondre à notre problématique de constitution de fils de discussion cohérents.

Merci de préciser la répartition du temps passé en entreprise et au laboratoire.

Cette répartition vous convient-elle ? Va-t-elle évoluer ?

Rappeler les interruptions longues de travail (ex : congé maternité, accident, ...)

Au cours de cette première année, à cause de la situation sanitaire due au COVID-19, j'ai passé environ 5% du temps en laboratoire, 30% en présentiel au sein de l'entreprise, et le reste en télétravail. Dans ce contexte très particulier, j'organise avec mes encadrants des séances de travail par visio-conférence d'environ 1h30 toutes les deux ou trois semaines. Cette répartition me convient parce que je parviens à garder un contact régulier avec mes encadrants en laboratoire. Elle évoluera à coup sûr lorsqu'il sera question de faire des analyses plus poussées sur le corpus d'entreprise en cours de constitution à cette période de février-mars 2021.

Avez-vous bénéficié de formation(s) au cours de l'année écoulée ? Merci d'en préciser le thème et l'organisateur.

Au cours de cette première année, j'ai effectué plusieurs formations proposées par Orange pour un volume horaire d'environ 24 heures :

- 1- Gagnez en impact dans vos écrits ! Développez vos compétences – Ateliers Écriture Efficace
- 2- S'initier à l'utilisation de GitLab et aux usages DevOps par la pratique
- 3- S'initier à l'IA avec Python n°4A : apprentissage automatique - partie ½
- 4- S'initier à l'IA avec Python n°5 : traitement automatique du langage naturel

Dans le cadre de mon plan individuel de formation, je me suis inscrit à des formations ci-dessous proposées par Sorbonne Université qui se dérouleront en mars 2021 :

1. [OPEN DATA : ASPECTS GÉNÉRAUX DES DONNÉES DE LA RECHERCHE STM \(S20210304\)](#) le 04 Mars de 17h00 à 20h00 (UTC+1)
2. [DIFFUSION AND IMPACT OF ACADEMIC RESEARCH, WHAT IS AT STAKE? \(S20210310A\)](#) le 10 Mars de 14h00 à 17:00 (UTC+01).
3. [COMMUNIQUEZ EFFICACEMENT A L'ECRIT \(S20210322B\)](#) du 22 Mars à 09h30 au 23 Mars à 18:00 (UTC+01)
4. [LE CIRCUIT DE PUBLICATION ACADÉMIQUE SCIENCES \(S20210325A\)](#) le 25 Mars de 09h30 à 12:30 (UTC+01).
5. [RÉALISER UN POSTER SCIENTIFIQUE \(S20210311D\)](#) du 11 Mars à 14h00 au 25 Mars à 16:00 (UTC+01).
6. [UNDERSTANDING AND APPLYING THE ENTREPRENEURIAL METHOD \(S20210331A\)](#) le 31 Mars de 09h00 à 13:00 (UTC+02).

Vos travaux ont-ils donné lieu à publication, présentation à un congrès, poster ou encore à dépôt de brevet au cours de l'année écoulée ? Merci de préciser les auteurs, le titre, les

références complètes du journal, des actes (...), la date, les pages et notamment le caractère international des communications.

Jusqu'ici mes travaux n'ont pas encore donné lieu à publication, mais cela est à venir dans les prochains mois notamment pour des conférences, workshops nationaux et internationaux qui portent sur plusieurs thématiques qui ont trait au Traitement Automatique des Langues Naturelles. Entre autres LaTeCH-CLfL 2021, CLEF 2021, TALN-RECITAL 2022, LREC 2022, etc. sont des exemples de ces conférences et/ou workshops dans lesquels mes travaux pourraient être publiés. Ces publications me permettront de m'ancrer dans la tâche de rédaction d'articles scientifiques et aussi de leur présentation.

Quelles sont vos perspectives pour l'année à venir ?

Pour l'année à venir, la finalisation de l'état de l'art est une priorité et par la suite des analyses approfondies des corpus seront faites et donneront lieu à la création de modèles d'IA spécifiques à nos problématiques. Les résultats de ces analyses seront présentés au travers d'articles scientifiques donnant lieu à des publications. Un prototype d'application intégrant ces modèles d'IA sera développé pour un cas d'utilisation concret.

Dans l'état actuel d'avancement de vos travaux, envisagez-vous de soutenir votre thèse dans les délais impartis ?

L'état actuel d'avancement de mes travaux me permet d'affirmer que je devrais pouvoir soutenir ma thèse dans les délais, à moins que je fasse face à des imprévus ou incidents qui m'obligeraient à interrompre mes travaux pour une de longue durée.

Commentaires et appréciations du responsable scientifique dans l'entreprise

Lionel Tadjou est un doctorant motivé qui s'est intégré très facilement au sein de l'équipe d'accueil à Orange, en dépit des difficultés liées au contexte sanitaire de cette première année. Il fait preuve d'autonomie dans ses travaux ; ceci s'est manifesté en particulier lors de la phase d'état de l'art (lectures et début de rédaction d'une synthèse) ainsi que lors de la phase de spécification et de développement de l'outil d'extraction des conversations sur les postes de travail des utilisateurs.




Lorsqu'il se sent en difficulté, il n'hésite pas à en faire part à ses responsables scientifiques en entreprise afin de trouver une solution pour aller de l'avant. En outre, il sait travailler en transverse afin, par exemple, de bénéficier de l'expérience de collègues ou de capitaliser sur des outils logiciels existants.

Lionel est organisé dans son travail et dispose de bonnes capacités à communiquer, tant à l'oral qu'à l'écrit. Il possède déjà une bonne vision d'ensemble de l'état de l'art des travaux liés à sa problématique de recherche. Son bon sens, son ouverture d'esprit et son expérience professionnelle précédente seront des atouts précieux pour le reste de la thèse. La deuxième année qui débute sera l'occasion de jauger ses capacités de créativité et d'innovation afin d'apporter de la valeur au travers d'un ou plusieurs use-cases dans un environnement d'entreprise.

Commentaires et appréciations du directeur de thèse

Lionel Tadjou s'est parfaitement approprié son sujet de thèse au cours de cette première année alliant les contraintes industrielles inhérentes à la nature du projet à l'identification des pistes scientifiques les plus pertinentes pour obtenir des résultats qui puissent aller au-delà de l'état de l'art. Les échanges réguliers que nous avons eu, notamment en présence d'Éric de la Clergerie qui co-encadre sa thèse l'ont conduit à prévoir un plan d'expérimentation précis intégrant notamment l'ensemble de l'ingénierie nécessaire à la bonne gestion des données primaires (échanges d'emails) sur lesquelles il devra travailler.

Je suis particulièrement confiant sur sa capacité à la fois d'obtenir des résultats probants sur le problème qu'il doit résoudre et de démontrer la pertinence de ces résultats dans un cadre scientifique plus large.

Nom du doctorant	Nom et fonction du responsable scientifique en entreprise	Nom et titre du directeur de thèse
Lionel Tadjou Taddonfouet	Fabrice Bourge Ingénieur de recherche	Laurent Romary Directeur de Recherche, Inria
E-mail personnel* : lioneltadjou@gmail.com	E-mail* : fabrice.bourge@orange.com	E-mail* : laurent.romary@inria.fr
Date et signature 23/02/2021 	Date, cachet et signature ¹ 16/03/2021 	Date, cachet et signature ³ 18/3/2021 

* Pour actualisation de nos fichiers.

Ce document est à retourner à: ANRT – Service CIFRE, 41 bd des Capucines, 75002 Paris.

¹ Conformément à l'article 6 des conditions générales d'octroi d'une CIFRE, si le rapport n'est pas signé par les deux responsables, il ne sera pas valide. En conséquence, le paiement de la subvention sera interrompu.