Rapport d'activités 2022 de la thèse: *Constitution de fils de discussion cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration.*

Thèse convention CIFRE **Orange**/**Inria**

Doctorant: Lionel Tadjou (lionel.tadjou@orange.com, lioneltadjou@gmail.com)

Encadrants Inria (équipe ALMAnaCH) : Laurent Romary & Éric de La Clergerie

Encadrants Orange (équipe Smart Working DATA&IA) : Fabrice Bourge & Tiphaine Marie

Date début de thèse: 18 Mars 2020

1 Introduction

Ce rapport d'activité de thèse s'inscrit dans la continuité du premier rapport et présente les explorations et avancements qui ont été menés principalement sur l'année 2022. Notre thèse s'intéresse à la constitution de fils de discussion cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration (C&C). Cet intérêt a été motivé par des observations sur les dernières décennies au cours desquelles le travail collaboratif en entreprise s'est considérablement développé, notamment au travers de l'effort d'équipement des collaborateurs avec des infrastructures et des outils de C&C facilitant l'échange et le partage permanent d'information : e-mail, messagerie instantanée, audio et vidéo conférence, les réseaux sociaux d'entreprise, etc. [Maier and Hädrich, 2011]. Selon les estimations de *The Radicati Group* pour la messagerie professionnelle, le nombre d'e-mails échangés quotidiennement au niveau mondial était estimé à plus de 128 milliards en 2019. Au niveau individuel, ces estimations se déclinent de la manière suivante : en 2019 chaque collaborateur a envoyé 30 e-mails par jour en moyenne et en a reçu 96, dont 19 étaient du spam [1]. Ces usages génèrent d'énormes quantités de données, dont une grande partie est souvent stockée sur le poste de travail ou sur des serveurs. Ces volumes de données sont amenés à croître davantage.

Différentes études indiquent qu'un travailleur du savoir passe entre 20 et 30% de son temps à chercher l'information [2]. C'est pour réduire ce temps de recherche d'informations que nous nous intéressons aux démêlage de conversations sur des données en français et plus précisément sur des emails. Cependant vu la rareté de corpus d'emails d'entreprise en français, nous nous sommes intéressés à d'autres types de modalités de conversations comme les discussions des pages Wikipedia [3] en français, mais aussi pour tester les approches que nous allons mettre sur pied nous avons aussi décidé d'utiliser des corpus d'emails en anglais : BC3 [4] et Enron. Le corpus MRDA [5] qui est un corpus de transcriptions d'enregistrements de réunions a aussi été exploité parce que annoté en actes de dialogue. Une fois ces corpus identifiés, nous avons procédé à des analyses statistiques sous les prismes de conversations, de posts/messages, d'interlocuteurs, de thématiques et termes récurrents. Ces analyses statistiques ne permettant que d'avoir un aperçu global de ces corpus. La reconnaissance d'acte de dialogue et l'identification de thématique dans des conversations sont deux prismes dont la combinaison aiderait à la résolution de la problématique de démêlage de conversation.

Dans la suite de ce rapport nous allons présenter les angles par lesquels nous espérons atteindre nos objectifs ainsi que les différentes analyses qui ont été menées sur les corpus identifiés : Discussions Wikipedia, BC3, Enron, MRDA

1. `www.radicati.com` (2015) ; "Email Statistics Report, 2015-2019"
2. `www.articlecube.com/research-shows-searching-information-work-wastes-time-and-money`
3. `https://www.ortolang.fr/market/corpora/wikidisc`
4. `https://github.com/dailykirt/ML_Enron_email_summary/tree/master/data/BC3_Email_Corpus`
5. `https://github.com/NathanDuran/MRDA-Corpus`

et OrangeCorpus. Une section est réservée à la mise en place d'un référentiel d'actes de dialogue en s'appuyant sur la norme ISO 24617-2. Ce référentiel sera ensuite utilisé pour l'annotation de segments de texte dans les emails. Certaines approches qui ont été explorées dont le zero-shot et fine-tuning pour l'identification d'actes de dialogue et BERTopic [6] pour l'identification des thématiques dans conversations d'emails seront détaillées dans les dernières sections de ce rapport.

2 Comment atteindre notre objectif ?

Pour atteindre notre objectif de démêlage de conversations, plusieurs angles sont à considérer pour une meilleure compréhension desdites conversations. Entre autres de ces angles on a : les métadonnées, la proximité thématiques entre les emails d'une conversation, les actes de dialogue inhérents aux segments de texte de ces conversations.

Les métadonnées de conversations et plus précisément d'emails permettent de savoir l'ordre des emails dans une conversation, les interlocuteurs d'une conversation mais aussi une thématique principale déduite des sujets d'emails. Ces métadonnées à elles seules ne permettent qu'une séparation atomique d'une conversation en différents emails, or notre objectif est de séparer de façon fine les contenus de chaque email et de mettre en relation ces contenus fins pour la reconstitution de fils de discussions.

L'identification de différentes thématiques (topic modeling) dans un email seul permet d'ores et déjà de distinguer des sous-conversations au niveau thématique. Sur deux ou plusieurs emails, on pourrait ainsi extraire des segments de texte transverses qui ne portent que sur des thématiques bien distinctes et notre problématique serait en partie résolue. Cependant vu la petite taille de texte dans les emails, il s'avère difficile de distinguer aisément des sous thématiques dans un email, d'où d'autres approches de reconstitution de fils de discussions ou démêlage de conversation qui consistent à identifier les différents actes de dialogue intra-emails, transverses sur les emails de conversations et leur mise en relation. Mais aussi le fait de savoir qu'un segment de texte d'un email à l'instant $t-1$ suit un autre de l'instant t contribuerait à approcher notre problématique. Ces deux dernières approches trouvent leur solution respectivement avec un classifieur de segment de texte en actes de dialogue et un second (ou régresseur) indiquant si un mail est potentiellement la continuation d'un autre mail (indépendamment des métadonnées)

Ces différentes analyses pour l'atteinte de notre objectif s'effectueraient aisément si des corpus d'emails adaptés ou annotés pour répondre à notre problématique existaient, ce qui n'est malheureusement pas le cas. C'est pour cette raison que nous avons identifié et analysé un certain nombre de corpus dont les détails sont fournis dans les prochaines sections.

3 Corpus et analyses

Comme mentionné dans l'introduction, afin d'approcher notre problématique de reconstitution ou démêlage de conversation, nous nous sommes intéressés à un certain nombre de corpus, ceci à cause de l'absence de corpus d'emails en français sur lesquels nous pourrions effectuer nos expériences. C'est dans ce sillage que nous avons exploité une partie d'un corpus de discussions en français de Wikipédia. Les autres corpus que nous avons exploités sont des corpus en langue anglaise, entre autres un petit corpus d'emails BC3 annotés en actes de dialogue, un corpus de transcriptions d'enregistrements de réunion MRDA et enfin une partie du corpus d'Enron ne contenant que des conversations. Enfin nous savons aussi exploité de façon succincte un corpus d'emails au sein d'Orange (non public) extrait des postes de certains collaborateurs. Le nombre de threads, les tokens et n-grams fréquents, l'analyse en actes de dialogue et le nombre d'interlocuteurs par conversations sont les prismes sous lesquels ces corpus ont été analysés. Dans les sections à venir nous allons détailler pour chacun de ces corpus les analyses qui ont été menées.

3.1 Discussions Wikipédia

Le corpus de discussions des pages de Wikipédia est un corpus téléchargé de la plateforme ORTOLANG [7] qui est une plateforme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française. Ce corpus a été constitué et rendu disponible au format XML, encodé selon la norme TEI-P5. dans le cadre des travaux de Ho-Dac and Laippala [2017] pour la caractérisation des discussions en ligne. Ce corpus WikiDiscussion contient 65 612 pages de discussion « article » contenant au minimum 2 mots sur les 3,5 millions de pages qui ont été extraits du dump des pages de discussions en 2015, Les analyses que nous avons effectuées sur ce corpus de discussions Wikipédia n'ont porté que sur 5000 pages de discussions, soit environ 7,7% du corpus total. De ces 5000 pages de discussions, 2865 conversations distinctes ont été extraites. Plusieurs caractéristiques ont ensuite été extraites du corpus, et d'autres pont été calculées. Parmi celles extraites, on a entre autres : le titre de la page de discussion et la liste des interlocuteurs de la dite page. Le nombre de posts ou messages par page de

6. https://maartengr.github.io/BERTopic/
7. www.ortolang.fr

discussion (en moyenne 4,45 par page), les tokens les plus fréquents (article, source, y, non, faire, etc.), le nombre d'interlocuteurs par page, le nombre d'interlocuteurs anonymes (1.28 en moyenne par page), le nombre de messages envoyés par des bots, etc. sont les caractéristiques qui ont été calculées. Les différentes caractéristiques peuvent être consultées sur cette page github [8]. De ces données extraites, il ressort majoritairement les titres de page suivants : *Discussions, Avis, Supprimer, Avis non décomptés, Conserver, Fichier proposé à la suppression, liens externes modifiés, Votes, Neutre, Avis divers non décomptés, etc.* Ces titres laissent apparaître que les discussions autour des pages Wikipédia sont fortement à caractère de vote afin de valider ou pas la publication, la suppression ou conservation d'une page (ou article). Ce vote est fait par les interlocuteurs intervenant dans la discussion de ladite page. Une analyse des fréquences de mots, bi-grams et tri-grams que vous trouverez ici [9] montrent les expressions fréquentes suivantes : *article, supprimer, source, @url(représentant les liens), mettre, référence, section, conserver, etc.* en début et fin de discussion. Ce qui laisse transparaître que les discussions sont autour du référencement de certains contenus, de la conservation ou suppression de la page. Cependant dans les bi-grams on retrouve *critères_ admissibilité, utiliser_ modèle, existence_ source, déplacer_ supprimer, conserver_ fusionner, accentuer_ idée, avis_ admissibilité, etc.* qui poussent encore à déduire que les discussions tournent autour des publications/conservations/suppressions/mises à jour d'articles ; ce qui fait sens parce que les publications d'articles sur Wikipedia doivent avoir des sources fiables, des contenus bien rédigés respectant des modèles prédéfinis. Il est donc tout à fait normal que dans les discussions, il ressort des votes pour validation ou pas d'articles à publier. À part ces aspects de vote et suite à une analyse manuelle via lecture de quelques articles, nous avons identifié des *questions/réponses, demande d'actions, suggestions, acquiescement* mais à très faible fréquence qui sont pourtant très intéressants pour approcher notre problématique. Le caractère conversationnel autour du vote des discussions Wikipedia étant distant de ce que l'on peut retrouver dans les conversations d'emails d'entreprise qui sont en général autour du suivi d'un projet, de demande d'action/d'informations, de partage d'informations etc. nous a emmené à ne pas creuser davantage sur le corpus Wikipedia et à s'orienter directement sur des corpus de conversations d'emails dont BC3, Enron et le corpus d'Orange.

3.2 Corpus de conversations d'emails : Enron, BC3, Orange

Approcher la problématique de démêlage de conversations au niveau des emails requiert de posséder des corpus de conversations d'emails. Nous avons ainsi identifié le corpus BC3, des conversations extraites du corpus Enron, en plus des emails collectés chez Orange. Sur ces corpus, nous avons effectués différentes analyses dont la classification de segment de texte en actes de dialogues, la recherche sémantique, l'analyse d'arbre de dépendances syntaxiques et la prédiction d'emails suivants aka *Next Email Prediction.*

3.2.1 Corpus BC3

Le corpus BC3 Ulrich et al. [2009] contient des fils de discussion provenant de la liste de diffusion du World Wide Web Consortium (W3C). Les discussions portent sur une variété de sujets tels que l'accessibilité du Web et la planification de réunions. La partie annotée de la liste de diffusion se compose de 40 fils de discussion pour 269 emails. Les threads ont été annotés par trois annotateurs humains par thread qui ont tout d'abord écrit des résumés, puis ont lié les phrases des threads aux phrases des résumés. Le corpus contient également des annotations d'actes de la parole. Nous sommes intéressés aux annotations en actes de parole faites sur ce corpus qui regroupent les actes suivants : *proposition, meeting, request, subjective, commitment* parce que les actes de dialogues sont un facteur important de compréhension automatique de conversations. Certains de ces actes de paroles ne reflètent pas ce que l'on peut retrouver dans la norme ISO 24617-2 qui fournit un ensemble de concepts empiriquement et théoriquement bien motivés pour l'annotation de dialogue qui peuvent s'appliquer aux conversations écrites. L'acte *request* dans les annotations de BC3 fait référence à une demande d'informations, et pourtant dans la norme, cet acte s'assimile plutôt à un acte commissif demandant à un interlocuteur d'effectuer une action. De même, les actes *meeting* et *subjective* ne font pas partir de la norme ISO sur laquelle nous devons nous appuyer parce que c'est un référentiel très utilisé notamment dans avec les corpus de parole comme MRDA et SwitchBoard. Dans leurs travaux de reconnaissance d'actes de dialogues dans des contenus d'emails et forums avec des approches semi-supervisées,Jeong et al. [2009] vont faire ré-annoter les phrases de BC3 (avec douze actes de dialogues listés dans le tableau ci-dessous) par deux annotateurs avec un accord inter-annotateur égale a 0,79.
Le fort pourcentage de l'acte de dialogue *Statement* nous a emmené a scruter ce corpus BC3 et à constater que cette classe peut être décomposée en d'autres actes de dialogues de la norme. Nous avons ainsi construit un premier référentiel d'acte de dialogue extrait des fonctions communicatives (2) de la norme ISO, que nous avons ensuite utilisé pour affiner les annotations du corpus BC3. La mise en relation des phrases transverses sur les conversations est une seconde couche d'annotations qui a été faite, ceci nous aidera plus tard à extraire des sous-conversations : ceci est une des approches de démêlage de conversations sous le prisme seul d'actes de dialogue portés par des segments de textes et les relations entre ces segments. Les détails sur l'affinage des actes de dialogues de BC3 et les différents graphes de transitions construits sont disponibles ici.

8. https://ltadjou.github.io/2022PhDReportFiles/Wikidisc2018_Threads_Report/
9. https://ltadjou.github.io/2022PhDReportFiles/Wikidisc2018_Threads_Report/OthersStats/

Tag	Description	BC3
S	Statement	69.56%
P	Polite mechanism	6.97%
QY	Yes-no question	6.75%
AM	Action motivator	6.09%
QW	Wh-question	2.29%
A	Accept response	2.07%
QO	Open-ended question	1.32%
AA	Acknowledge and appreciate	1.24%
QR	Or/or-clause question	1.10%
R	Reject response	1.06%
U	Uncertain response	0.79%
QH	Rhetorical question	0.75%

FIGURE 1 – Actes de dialogues du corpus BC3, extrait de Joty and Mohiuddin [2018]

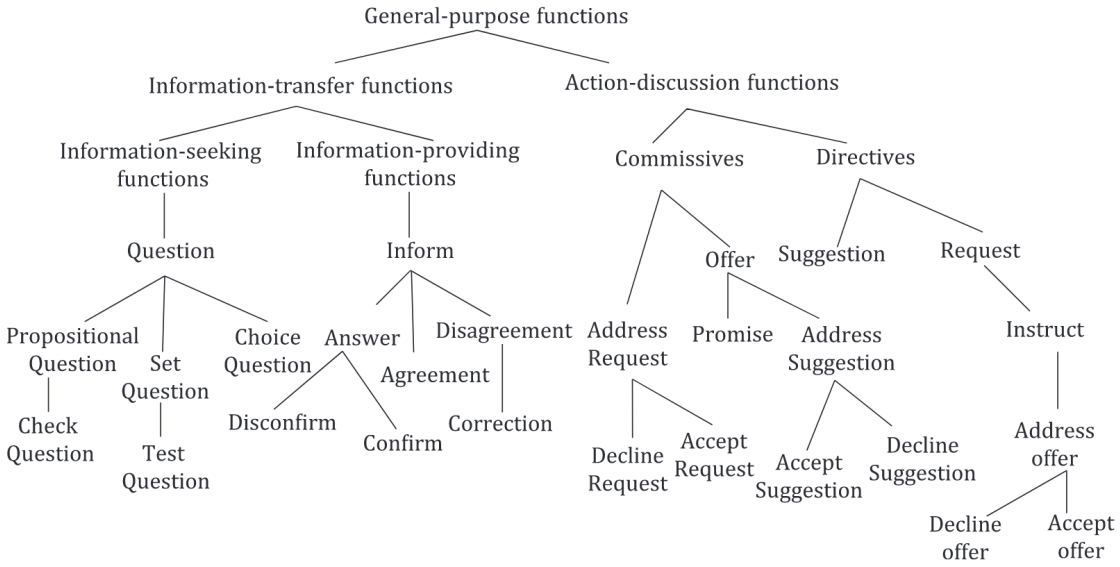

FIGURE 2 – Fonctions de communications à usage général : extrait de la norme ISO 24617-2

La petite taille du corpus BC3 (environ mille phrases pour une dizaine d'acte de dialogues) ne nous permet pas de fine-tuner un modèle de classification des phrases en actes de dialogue. Au moment de la rédaction de ce rapport, le framework SetFit (*SetFit - Efficient Few-shot Learning with Sentence Transformers*) de Karimi Mahabadi et al. [2022] a vu le jour et permet de faire du few-shot learning de façon rapide (même sur CPU) avec des résultats satisfaisants. Cette approche sera testé très prochainement et le modèle obtenu permettra un transfert sur de plus grandes quantités de données comme Enron. Toutefois le corpus BC3 est en cours de finalisation d'annotation, sera rendu public. Il nous aidera à valider nos approches pour le démêlage de conversations.

3.2.2 Corpus Enron

Le corpus d'emails Enron est l'un des corpus les plus utilisés lorsqu'il est question d'analyse d'emails et de conversations asynchrones dans le monde de la recherche. Cependant pour la problématique de reconnaissance d'actes de dialogues dans des conversations écrites, il n'a pas souvent été utilisé. Les travaux de Taniguchi et al. [2020] qui constituent un corpus annoté (6 000 emails pour 35 000 phrases dans plus de 2 000 fils de discussion) en actes de dialogues sur deux niveau (email et phrases d'emails) est l'un des plus récents. Celui-ci nous aurait permis de gagner beaucoup de temps dans nos travaux, mais malheureusement ce corpus n'est pas disponible. Nous avons

cependant identifié les travaux de Jamison and Gurevych [2013] qui ont extrait 70178 threads du corpus Enron et les ont rendu disponibles. Notons que ce sous-corpus n'est pas annoté. Dans leur papier, il dresse un tableau ci-dessous 3 de la taille des conversations dans leur corpus.

Thread Size	Num threads
2	40,492
3	15,337
4	6,934
5	3,176
6	1,639
7	845
8	503
9	318
10	186
11-20	567
21+	181

FIGURE 3 – Nombre des conversations dans le Corpus de Threads Enron.

Nous avons utilisé cette version du corpus Enron pour fine-tuner BERT avec une approche contrastive sur la tâche de *Next Email Prediction* dont les différents résultats sont disponibles sur cette page. Actuellement nous explorons la possibilité de transférer un modèle de reconnaissance d'actes de dialogue entraîné sur les données de transcriptions de réunion (MRDA) pour classifier en actes de dialogues les phrases de ce sous-ensemble du corpus Enron et effectuer un second transfert sur les conversations d'Orange. La piste d'entraîner des modèles multilingues sera fortement considérée. Nous envisageons aussi de faire annoter une partie de ce corpus comme les annotations effectuées par Taniguchi et al. [2020] par un prestataire afin d'avoir une bonne base en termes de données pour valider nos approches.

3.2.3 Corpus MRDA

Le corpus MRDA (*Meeting Recorder Dialogue Act Corpus*) est un corpus de transcriptions d'enregistrements de réunion annotées en actes dialogues avec 3 niveaux d'étiquettes : basique, général et complet. Nous utilisons ce corpus afin d'entraîner un modèle pour la classification ou reconnaissance d'actes de dialogue sur des énoncés de conversations. MRDA est la base un corpus audio, d'où la présence de multiples marqueurs de conversations orales tels que *umh, umhumh, you know, so, hummm, etc.* qui sont quasi absents dans les conversations écrites surtout dans des emails d'entreprise. Plusieurs fois ces marqueurs seuls constituent des énoncés de conversations qui ont été annotés. Nous sommes partis de l'hypothèse que ces marqueurs vont créer du bruit dans les modèles que nous allons entraîner et donc nous avons filtré le corpus MRDA en supprimant les énoncés constitués seulement de ces marqueurs ainsi ces derniers apparaissant dans les énoncés. Après avoir filtré le corpus on a obtenu 36722, 7985, 7918 énoncés respectivement pour les données d'entraînement, de validation et de test.

Plus haut, nous avons mentionné le fait d'utiliser des actes de dialogue de la norme ISO 24617-2. Dans la continuité de cette action, nous avons entrepris le mapping des actes de dialogue de MRDA avec ceux de la norme ISO tout en rajoutant certains actes que nous estimons importants et ne faisant pas partie des fonctions communicatives de la norme. Ce sont principalement : *explication, reformulation, politesse, appréciation/évaluation* qui doivent probablement appartenir à la grande classe *feedback* dont nous allons mieux cerner le sens pour une meilleure ré-affectation de certains actes de dialogue de MRDA. Le choix de mapper les actes de dialogues de MRDA tient du fait qu'il est l'un des corpus de conversations annotées en actes de dialogue et sur lequel nous allons nous appuyer pour approcher notre problématique. Afin de mieux mapper certains actes de dialogues, nous avons analysé certains énoncés de MRDA de façon collégiale (deux personnes) avant de décider du mapping à effectuer. Ce tableau récapitule les différents mapping que nous avons effectués. Dans ce tableau les actes de dialogue de la colonne "A" sont ceux de MRDA, le sous-tableau avec les bordures en gras représente les fonctions communicatives 2 de la norme et enfin la colonne "B" contient les abréviations des actes de dialogue. Une mise à jour de ce tableau va permettre d'avoir des actes de dialogues non ambigus, répondant à notre besoin et indexés sur la norme. La répartition des nouveaux actes de dialogues sur le corpus filtré est accessible ici.

Nous avons aussi entraîné des modèles avec les plongements de mots simples de Keras (KE), de Glove(GLV) et fine-tuné le modèle BERT pour la tâche de classification d'énoncé en acte de dialogue. Comme inputs à ces modèles, nous avons utilisé dans un premier temps les énoncés à classifier pris indépendemment les uns des autres et dans un second temps, nous avons considéré un contexte qui était l'énoncé précédant celui à classifier. Nous avons aussi rajouter des couches d'auto-attention afin de voir si cela aurait un impact sur les performances. L'image ci-dessous montre des résultats obtenus sur le corpus MRDA filtré pour cette tâche de classification d'énoncés en actes de dialogue.

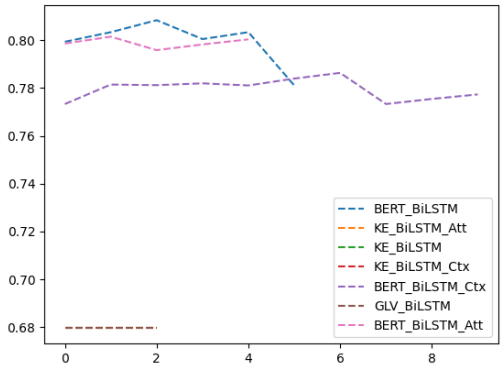

FIGURE 4 – Performances des différentse configurations pour la classification en actes de dialogues ; (les courbes qui ne sont pas visibles ont une performance inférieure à 0.68)

D'autres tests ont été effectués en rajoutant une couche CRF, ceci s'inspirant des travaux effectués dans ce projet [10] qui utilisent des plongements spécifiques construits à partir des concepts extraits du corpus MRDA. Dans cette approche avec CRF, le modèle prend en entrée un input constitué de plusieurs énoncés avec chacun son acte de dialogue respectif, la couche CRF essaie de trouver le meilleur chemin de prédiction d'acte de dialogue pour chaque énoncé. La performance du modèle est d'autant meilleure que le nombre d'énoncés dans un input est grand. Cette même architecture a été utilisée en remplaçant les plongements par ceux de BERT. L'image et le tableau ci-dessous récapitulent respectivement l'architecture du modèle et les différents résultats obtenus.

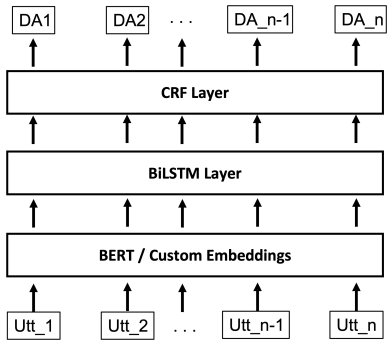

FIGURE 5 – Architecture du modèle

Tag size	Utterances/inputs	Architecture	Accuracy
5*	10	CuEmb+BiLSTM+CRF	88,55%
5	10	BERT+BiLSTM+CRF	83,26%
5	10	BERT+MLP+CRF	83,16%
5	100	CuEmb+BiLSTM+CRF	**90,48**%
10**	100	CuEmb+BiLSTM+CRF	84,34%
10	100	CuEmb+BiLSTM+CRF	76,67%

TABLE 1 – Résultats des modèles entraînés sur le corpus MRDA de base avec différentes configurations, la dernière ligne donne la performance sur le corpus filtré. **CuEmb** fait référence à des plongements spécifiques ; * pour les labels de base ; ** pour les labels généraux

Les résultats dans ce tableau montrent que les plongements spécifiques utilisés dans le projet donnent de meilleurs résultats par rapport au fine-tuning de BERT avec ses différentes couches gelées, ici on utilise juste la représentation vectorielle du token [CLS] placé en début de chaque énoncé lors de la tokenisation avec BERT.

Toujours dans le cadre de pouvoir identifier ou classiffier les actes de dialogues sur des énoncés, la piste de recherche par similarité sémantique a été explorée avec le corpus de base MRDA et la bibliothèque FAISS. Les données d'entraînement ont été utilisées pour construire la base de données de recherche et les données de tests pour effectuer les requêtes. Cette matrice de confusion présente les résultats obtenus, avec les énoncés étiquettés "Statement (s)" sont majoritairement prédits à la place des autres. SetFit a rapidement été testé sur le corpus filtré avec ses actes de dialogues de niveau 1 et différentes configurations sur le nombre d'échantillon (NE) par classes et le nombre de paires (NPT) de textes à générer pour l'apprentissage contrastif. La meilleure performance obtenue a été de 77,94% avec NE =250 et NPT = 250. Ce résultat pourrait être amélioré en procédant à une

10. https://github.com/jonas-scholz123/msci-project

recherche de meilleurs hyper-paramètres avec le framework optuna. Ceci est une tâche que nous allons effectuer d'ici fin novembre 2022.

3.2.4 Corpus Orange

Le corpus Orange est un corpus d'emails collectés sur les postes de 5 collaborateurs, la taille totale du corpus est de 20577 dont 5105 sont des emails uniques et le reste sont des conversations. L'image ci-dessous présente la répartition du nombre d'emails par conversation.

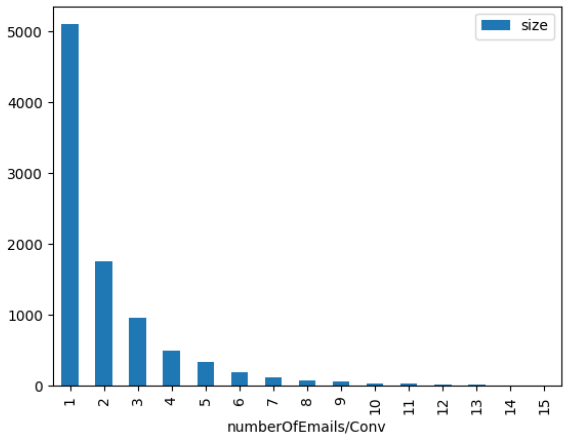

FIGURE 6 – Répartition du nombre d'emails par conversation dans le corpus Orange

Sur ce corpus, le clustering des contenus a été testé dans le cadre d'un prototype qui était en cours de développement lors d'un stage cet été 2022 et dont l'objectif était de mettre sur pied un outil qui exploitera les résultats de nos travaux afin d'afficher les conversations de façon démêlée. L'objectif du clustering était d'extraire de ces clusters des n-grams qui constitueraient des actes de dialogue, intentions ou thématiques liés à des contenus (phrases ou sous-parties) d'email ; l'idée étant de d´obtenir des actes de dialogue comme ceux-ci : *ai_envoyé_documents, réservé_créneau, enverrai_document, candidat_stage, etc.*. BERTopic a été utilisé avec différents modèles pré-entraînés monolingues et multilingues. Cette image montre les thématiques les plus fréquentes avec le modèle pré-entrainé multilingue *distilbert-multilingual-nli-stsb-quora-ranking* . Les résultats de ce clustering n'ont pas été satisfaisants car notre besoin était celui d'extraire des ngrams sémantiquement cohérents. Cependant moyennant la modification de certains paramètres de BERTopic ou l'utilisation des bibliothèques telles que NLTK, Spacy ou textacy, de meilleurs résultats pourraient être obtenus.

Dans la continuité d'extraction de ngrams, nous avons formulé une approche heuristique d'identification d'intentions ou de thématiques qui consiste à partir des éléments Part-of-Speech (POS) d'une phrase, de constituer des combinaisons (2 ou 3) chaînées en respectant leur ordre d'apparition dans la phrase et en supprimant celles qui sémantiquement n'auraient pas une grande valeur. Ceci afin d'extraire leurs tokens correspondants et en faire des bi-grams ou tri-grams qui seraient des intentions/actions ou des thématiques comme listés dans le précédent paragraphe . Les combinaisons de POS ont été constituées à partir d'observations plusieurs phrases dont les POS ont été calculés. Une combinaison est identifiée comme sémantiquement cohérente et ajoutée à une liste de combinaison lorsque : 2 ou 3 POS dans une phrase dénotent une thématique, une intention ou même un concept ; moyennant une suppression d'un ou de deux POS intermédiaires (dont les valeurs sémantiques ne sont pas importantes) entre ceux qui vont constituer la dite combinaison. Ci-dessous quelques examples :

— Je mets Fabrice en copie étant donné qu'il assume l'intérim.

POS : [('je', 'PRON'), ('mettre', 'V'), ('Fabrice', 'PROPN'), ('en', 'ADP'), ('copie', 'NOUN'), ('être', 'AUX'), ('donner', 'VPP'), ('que', 'SCONJ'), ('il', 'PRON'), ('assume', 'V'), ('le', 'DET'), ('interim', 'NOUN')]

Combinaisons : V_PROPN_NOUN (mettre_Fabrice_copie), AUX_VPP (être_donner), V_NOUN (assumer_interim)

— je vous fais suivre cet échange qui contient un lien vers Plazza ;

POS : [('je', 'PRON'), ('vous', 'PRON'), ('faire', 'V'), ('suivre', 'VINF'), ('ce', 'DET'), ('échange', 'NOUN'), ('qui', 'PRON'), ('contenir', 'V'), ('un', 'DET'), ('lien', 'NOUN'), ('vers', 'ADP'), ('Plazza', 'PROPN')

Combinaisons : V_VINF_NOUN (faire_suivre_échange) , V_NOUN (contenir_lien)

Cette approche a ainsi permis de constituer une liste de 49 combinaisons de POS pour l'extraction d'éventuelles intentions ou thématiques. Les bibliothèques Spacy/textacy auraient pu être utilisées pour l'extraction de ce type de n-grams. Nous avons en quelque sorte réinventer la roue, mais avec des résultats qui répondent mieux à nos besoins. Le tableau de la figure ci-dessous compare les n-grams extraits avec Spacy/textacy et notre approche que nous avons nommée POS_Combination.

	Spacy/textacy	***POS_Combination***
Je mets Fabrice en copie étant donné qu'il assume l'interim	[mets Fabrice, Fabrice en copie, copie étant donné, assume l'interim]	mettre_Fabrice_copie, être_donner, assume_interim
je vous fais suivre cet échange qui contient un lien vers Plazza	[échange qui contient, contient un lien, lien vers Plazza]	['vous_faire', 'faire_suivre_échange', 'échange_qui_contenir', 'contenir_lien_Plazza']
si tu as un peu de temps, peux-tu jeter un coup d'œil aux deux slides que je prévois de présenter (en plus des résultats obtenus jusqu'à présent) sur les travaux à venir, et les commenter ou compléter ?	[résultats obtenus, obtenus jusqu', jeter un coup, coup d'œil, prévois de présenter, résultats obtenus jusqu', jusqu'à présent, travaux à venir, commenter ou compléter]	['QUESTION', 'as_temps_peu', 'peu_tu_jeter', 'jeter_coup', 'jeter_coup_œil', 'coup_œil_slide', 'slide_que', 'je_prévoir_présenter', 'présent_travail', 'présenter_résultat', , 'obtenir_présent', 'travail_venir', 'le_commenter', 'commenter_compléter']

FIGURE 7 – Spacytextacy Vs POSCombination

De ces exemples, il ressort que Spacy/textacy fait ressortir des n-grams sémantiquement corrects et en omet d'autres qui sont par contre produits par l'approche que nous avons développée. Après plusieurs tests sur différentes phrases, nous avons constaté que tous les n-grams extraits jusqu'ici ne s'apparentent pas toujours à des intentions ou thématiques, d'où la nécessité de post-traitements ou d'amélioration de notre approche. Cette approche a finalement été abandonné au profit d'une exploration de classification de segments de texte en acte de dialogue avec du zero-shot Learning.

Notons que ces expériences menées sur le corpus Orange ont été faites bien avant toutes les autres décrites dans les précédentes sections, raison pour laquelle notre liste d'actes de dialogue ne s'appuyait à ce moment-là sur aucun référentiel. La liste en question est la suivante : *question ouverte, déclaration incertaine, politesse, déclaration certaine, s'engager, question précise, déclaration négative, réponse positive,demander un service, acquiescer, ordre, consigne, suggérer, correction, répétition, souhait, subjectivité, argumentation, partage d'information.*

Une trentaine des phrases extraites de deux ou trois conversations ont été utilisées pour cette classification zero-shot. Sur ces résultats, on constate que le modèle Roberta prédit avec environ 75% de précision (pourcentage sur les 27 phrases utilisées) le bon acte de dialogue parmi les trois premiers actes qu'il retourne sur une dizaine d'actes qu'on lui a fournis en entrée. Cependant, pour certaines phrases, sur les trois meilleurs actes de dialogue prédits, aucun d'eux n'est correct, ceci serait dû à une absence de contexte qui n'est rien d'autre que les phrases précédentes dans la conversation. Cette aspect contextuel a été pris en compte avec le corpus MRDA.

Nous venons ainsi de décrire les expériences que nous avons menées sur le corpus Orange depuis en début d'année 2022. Certaines aspects de ces expérimentations comme les combinaisons de Part-Of-Speech seront approfondis et combinés avec d'autres approches que nous avons décrits plus haut notamment pour la reconnaissance d'acte de dialogue dans les emails.

4 Objectifs à court et moyen terme

En termes d'objectifs à court terme, c'est-à-dire dans les semaines à venir, nous nous allons davantage creuser certaines expérimentations présentées plus haut, notamment l'utilisation de SetFit et celle de l'architecture avec CRF dans l'objectif de soumettre un papier à une conférence d'ici le début d'année 2023. Les résultats de ces approfondissements seront transférés sur les conversations d'Orange. La finalisation de notre référentiel d'actes de dialogue sera aussi faite d'ici cette fin d'année 2022 et celui-ci sera utilisé pour éditer un guide d'annotation dans la cadre d'une campagne d'annotation que nous comptons effectuer sur une partie du corpus Enron.

Dès le début d'année 2023, l'accent sera mis sur la rédaction du manuscrit dont un premier plan et un premier jet de l'état de l'art ont déjà été produits. Au cours de cette période de rédaction, certaines expérimentations continueront notamment sous le prisme des thématiques/sous-thématiques dans l'objectif de les combiner avec les

actes de dialogues pour mieux approcher notre problématique de démêlage de conversation. La fin de rédaction du manuscrit se situera en fin juin, début juillet 2023, pour une période de soutenance en septembre 2023.

5 Conclusion

Dans ce rapport, nous avons présenté ce qui a occupé ma deuxième et début de troisième année de thèse. De l'identification de corpus à exploiter, en passant par leurs analyses, les expérimentations menées sur ces corpus principalement pour la tâche de reconnaissance d'actes de dialogue sur des segments de texte ont été présentés, ainsi que le plan à suivre dans les semaines et mois à venir jusqu'à la soutenance.

La reconstruction de fils de discussions ou le démêlage de conversation, l'identification des actes de dialogues et l'extraction de thématiques sont les principales problématiques dans lesquelles nous nous investirons davantage. Nous proposerons ou amélioreront des approches pour les résoudre et mettrons en place un transfert de ces approches sur les emails d'Orange qui ont été collectés jusqu'ici.

En termes d'objectifs à atteindre, des tâches d'approfondissement d'expérimentations sont à mener très prochainement avec objectif de publication d'un papier et la rédaction du manuscrit à moyen terme.

Tous les points abordés dans ce rapport font ainsi état de mes activités de thèse courant ma deuxième année et début de troisième année.

Références

Lydia-Mai Ho-Dac and Veronika Laippala. Le corpus WikiDisc : ressource pour la caractérisation des discussions en ligne. In Ciara R. Wigham and Gudrun Ledegen, editors, *Corpus de communication médiée par les réseaux : construction, structuration, analyse.*, Humanités numériques, pages 107–124. l'Harmattan, March 2017. URL `https://halshs.archives-ouvertes.fr/halshs-01488029`.

Emily Jamison and Iryna Gurevych. Headerless, quoteless, but not hopeless ? using pairwise email classification to disentangle email threads. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 327–335, Hissar, Bulgaria, September 2013. INCOMA Ltd. Shoumen, BULGARIA. URL `https://aclanthology.org/R13-1042`.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1250–1259, Singapore, August 2009. Association for Computational Linguistics. URL `https://aclanthology.org/D09-1130`.

Shafiq Joty and Tasnim Mohiuddin. Speech Act Modeling of Written Asynchronous Conversations : A Neural CRF Approach. *Computational Linguistics (Special Issue on Language in Social Media, Exploiting discourse and other contextual information)*, pages 859–894, 2018. URL `https://www.mitpressjournals.org/doi/full/10.1162/coli_a_00339`.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. Prompt-free and efficient few-shot learning with language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 3638–3652, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi : 10.18653/v1/2022.acl-long.254. URL `https://aclanthology.org/2022.acl-long.254`.

Ronald Maier and Thomas Hädrich. Knowledge management systems. In David G. Schwartz and Dov Te'eni, editors, *Encyclopedia of Knowledge Management, Second Edition*, pages 779–790. IGI Global, 2011. URL `http://www.igi-global.com/Bookstore/Chapter.aspx?TitleId=49027`.

Motoki Taniguchi, Yoshihiro Ueda, Tomoki Taniguchi, and Tomoko Ohkuma. A large-scale corpus of E-mail conversations with standard and two-level dialogue act annotations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4969–4980, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi : 10.18653/v1/2020.coling-main.436. URL `https://aclanthology.org/2020.coling-main.436`.

Jan Ulrich, Giuseppe Carenini, Gabriel Murray, and Raymond T. Ng. Regression-based summarization of email conversations. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*. The AAAI Press, 2009. URL `http://aaai.org/ocs/index.php/ICWSM/09/paper/view/188`.