Rapport d'activités première année de 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 Wroking DATA&IA) : Fabrice Bourge & Tiphaine Marie

Date début de thèse: Mars 2020

1 Introduction

Cette 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 et/ou sur des serveurs. Ces volumes de données sont amenées à croître d'avantage.

Qui n'est pas régulièrement confronté à la difficulté de retrouver un mail ou un document sur son poste de travail, même en utilisant les outils de recherche avancée ? Le travailleur du savoir perd un temps considérable à retrouver l'information dont il a besoin. Différentes études indiquent qu'il passe entre 20 et 30% de son temps à chercher l'information [2].

Ces données stockées et accessibles à tout moment représentent une source considérable de connaissance potentielle, souvent ignorée et inexploitée. On estime que 80% de ces données sont peu ou pas structurées [3], c'est l'une des causes principales de leur non exploitation. Comment donc exploiter ces connaissances, certes explicites, mais souvent peu structurées, contextuelles et noyées dans un océan d'information ? Et parfois même hétérogènes à cause de la multiplication des outils utilisés. Ma thèse s'intéresse à ce problème, mais sur le plan de la constitution de fils de discussions de ces données issues de conversations en entreprise. L'objectif de la thèse est de pouvoir automatiser en partie la compréhension fine et rapide des échanges professionnels pour les collaborateurs d'Orange. Ceci par le biais d'assistants de rappel d'engagement, de graphes de leurs actions sur un projet, de recherches faciles et guidées. L'atteinte de cet objectif passe nécessairement par l'analyse discursive des données textuelles. Des recherches ont été menées pour analyser les discours et les discussions pour tenter de déduire les corrélations logiques entre messages, afin de permettre le traitement automatique pour des besoins de fouille de données ou traitement de langage naturel. Ainsi dans les années 1990 des travaux basés sur les méthodes d'apprentissage automatique (*Machine Learning* en anglais) appliqué à la classification de document selon leurs thématiques par [Lewis, 1992] ont été effectués. Quelques années plus tard, des variantes de ces mêmes méthodes sont utilisées pour

1. `www.radicati.com` (2015) ; "Email Statistics Report, 2015-2019"
2. `www.articlecube.com/research-shows-searching-information-work-wastes-time-and-money`
3. `www.datamation.com/big-data/structured-vs-unstructured-data.html`

classer les emails [Cohen, 1996] et détecter les sentiments [Pang et al., 2002]. En s'appuyant sur les travaux de [Searle, 1975] sur les actes de langage (*Speech Acts* en anglais) et sur ceux de [Finke et al., 1998] visant à détecter automatiquement des actes dans des conversations téléphoniques , [Cohen et al., 2004] proposent en 2004 une approche visant à détecter l'intention des auteurs dans les emails grâce à une ontologie d'actes de langage. Leur conclusion était, d'une part, qu'il faudrait tenir compte du contexte d'un email afin de pouvoir détecter des actes de langage implicites et, d'autre part, qu'il est fréquent qu'un message porte sur plusieurs sujets de discussion en même temps. Ceci soulève deux problèmes difficiles : la segmentation des messages et le démêlage des discussions imbriquées. Cependant en amont de ces problématiques se dresse une autre problématique sans laquelle celles-ci ne sauraient trouver un intérêt, il s'agit de la constitution des corpus. Ainsi dans les prochaines sections de ce rapport, nous allons détailler ces problématiques en s'appuyant sur des recherches qui se rapprochent plus ou moins de notre contexte et qui ont fortement marqués nos lectures. Avant de conclure, nous présenterons tout aussi nos perspectives à court, moyen et long terme, ainsi que les formations faites jusqu'ici dans le cadre du plan individuel de formation.

2 Reconstruire des fils de discussions et problématiques connexes

Reconstruire des fils de discussions consiste à partitionner des flux de données issues de différents types de modalité de communication, en conversation et sujets cohérents. En général, c'est un problème lié à des communications médiées par ordinateur (CMO) : forums, liste de diffusion ou emails. Les conversations et sujets cohérents visés par la reconstruction de fils de discussion peuvent être d'ordre purement structurel, temporel ou sémantique, ceci avec différents niveaux de granularité. Plus la granularité est fine, plus on a un fil de discussions distinctes avec des segments d'échanges parfaitement démêlés. D'où une autre problématique connexe : le démêlage de discussions imbriquées. Des fils de discussions avec une granularité extra-fine faciliterait par exemple l'accès à l'information au bon moment (*push notification*), la classification d'emails, une visualisation explicite d'informations dans une boîte de messagerie.

2.1 Travaux connexes

Une étude préliminaire de notre problématique et les différentes lectures que nous avons effectuées cette première année nous ont conduit à identifier les problématiques connexes de la thèse. Nous allons par la suite vous présenter ces problématiques sous le prisme d'articles qui nous semblent pertinents.

2.1.1 Identification et reconstruction de fils de discussions

Identifier ou détecter des fils de discussion à partir de données CMO s'avère être une tâche très importante pour des applications telles que le marketing digital, la recherche d'informations, la lutte contre la criminalité numérique, pour ne citer que ceux-ci. Plusieurs mécanismes ont été développés afin d'adresser ce problème de reconstruction de fils de discussions. Plus précisément les travaux de [Lewis and Knowles, 1997] et de [Yeh, 2006], se sont appuyés sur les informations d'en-têtes d'emails (émetteur, sujet et destinataires, *thread index*, dates, etc.) et des méthodes de calcul de similarité de certains mots clés identifiés dans des emails. De telles approches sont fortement liées à des corpus et sont très coûteuses à ré-implémenter sur de nouveaux corpus parce qu'elles nécessitent l'analyse par des experts du domaine pour l'identification de ces *features* ou mots clés importants.
Pour répondre à cette difficulté, [Domeniconi et al., 2016] proposent une approche qui combine des calculs de similarité de huit *features* construits pour chaque email. Dans leur approche un email est représenté dans un espace tri-dimensionnel avec le premier axe porté sur le contenu sémantique, le second sur les interactions sociales (expéditeur/destinataire) et le dernier sur le temps de création d'un message. Ils constituent ainsi la feature sémantique d'un email via une requête sur la plateforme AlchemyAPI [4]. Cette plate-forme prend en entrée un texte non structuré et retourne diverses informations riches en sémantique. Parmi ces informations ils se sont intéressés à trois composantes : les mots-clés thématiques, les entités-nommées et ses concepts inhérents (par exemple pour le texte *"Mes marques préférées sont BMW et Porsche"* AlchemyAPI retourne *Industrie automobile*). AlchemyAPI retourne ces informations avec des valeurs de confidence compris entre 0 et 1. Ces valeurs ont permis de construire trois vecteurs sémantiques pour chaque message. Ces vecteurs ont ensuite été utilisées pour calculer des similarités cosinus entre deux messages. Une autre composante linguistique qu'ils calculent entre deux messages est la similarité cosinus entre les représentations sac-de-mots (*BOW - Bag of words*) de leur contenu. La seconde dimension orientée sur les relations sociales d'un message, est évaluée par deux **similarités de Jaccard** : une qui calcule la similarité $f_{S_U}(m_i, m_j)$ entre deux messages (m_i, m_j), chacun représenté par un vecteur $\mathcal{U}_n = \{u_1, u_2, ...\}$ qui est l'union de son émetteur et de ses destinataires

$$f_{S_U}(m_i, m_j) = \frac{|\mathcal{U}(m_i) \cap \mathcal{U}(m_j)|}{|\mathcal{U}(m_i) \cup \mathcal{U}(m_j)|}$$

4. Rachetée par IBM en mars 2015 et retirée des ses services en 2017

La seconde calcule la similarité de Jaccard du voisinage de deux messages $f_{S_N}(m_i, m_j)$. L'ensemble des voisinages d'un utilisateur $\mathcal{N}(u)$ est l'ensemble des utilisateurs ayant reçu au moins un email de u, cet utilisateur est aussi inclus dans son voisinage.

$$f_{S_N}(m_i, m_j) = \frac{1}{|\mathcal{U}(m_i)||\mathcal{U}(m_j)|} \sum_{\substack{u_i \in \mathcal{U}(m_i) \\ u_j \in \mathcal{U}(m_j)}} \frac{|\mathcal{N}(u_i) \cap \mathcal{N}(u_j)|}{|\mathcal{N}(u_i) \cup \mathcal{N}(u_j)|}$$

Enfin la dernière dimension portée sur le temps se calcule par une similarité qui est égale au logarithme de l'inverse de la distance entre deux messages, cette distance exprimée en jours :

$$f_{S_T}(m_i, m_j) = \log_2(1 + \frac{1}{1 + |t_{m_i} - t_{m_j}|})$$

L'inverse de la norme de distances est utilisé ici pour avoir des valeurs entre 0 et 1. La figure 1 montre deux messages

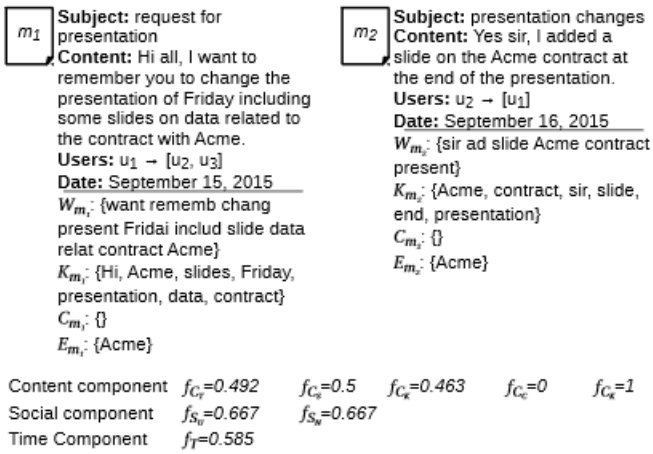

FIGURE 1 – Exemple de calcul de similarité des *features* entre deux messages

avec leurs *features* inhérentes et les valeurs des différentes similarités calculées. Après le calcul de similarité entre les messages, ils définissent une distance appelée ***SIM*** entre deux points (emails) qui se calcule à partir des différents valeurs des similarités $\mathcal{F} = \{f_{C_T}, f_{C_S}, f_{C_K}, f_{C_E}, f_{C_C}, f_{S_U}, f_{S_N}, f_T\}$ présentées plus haut.

$$SIM(m_i, m_j) = \prod_{f_i n \mathcal{F}} (1 + f(m_i, m_j))$$

Ils ont ensuite calculé des matrices $N \times N$ avec des similarités entre chaque paire de messages (m_i, m_j) et utilisent deux algorithmes de clustering qui sont les plus connus en terme d'approche de clustering basée sur les distances. Il s'agit de l'algorithme basé sur la densité DBSCAN d'[Ester et al., 1996] qui crée des clusters en fonction d'un seuil d'évaluation entre deux points (messages dans ce papier) et la seconde approche de clustering hiérarchique agglomérative de [Bouguettaya et al., 2015] qui consiste à créer pour chaque point son cluster (avec le point en question comme premier élément) et à les fusionner progressivement à partir d'une liaison moyenne entre eux.
Dans ce papier les auteurs proposent une autre méthode de clustering de messages ou construction de fils de discussions. Cette approche est un classifieur binaire qu'ils ont nommé SVC et qui prend en entrée des combinaisons de sous-composantes de la mesure de similarité ***SIM*** entre deux messages avec comme label 1 si les deux messages appartiennent à la même conversations et 0 sinon. Les probabilités prédites par ce classifieur sont utilisées comme distance entre des paires de messages afin de construire le cluster. L'avantage d'une telle approche est qu'elle trouve de façon autonome les caractéristiques appropriées pour chaque corpus, ce qui nécessitait l'intervention d'experts dans les anciennes approches. Pour cette méthode d'apprentissage supervisée, ils ont expérimenté plusieurs classifieurs : les arbres de décisions (J48, Random Forest), les machines à vecteurs de support (LibSVM) et la régression logistique.
Ils ont entraîné et testé leurs méthodes de clustering d'apprentissage supervisé sur sept corpus :

— Le corpus BC3 [Ulrich and Murray, 2008]
— Un sous ensemble de liste de diffusion du corpus Apache d'août 2011 à Mars 2012 [5]
— Des listes de diffusion du corpus Fedrora Redhat Projet collectés pendant 6 mois en 2009 [6]

5. `tomcat.apache.org/mail/dev`
6. `www.redhat.com/archives/fedora-devel-list`

— le contenu de deux pages Facebook publiques nommées *Healthy Choice*[7] et *World Health Organizations*[8]
— Le contenu de groupes publiques de Facebook au nom de *Healthcare Advice*[9] et *Ireland Support Android*[10]

Enfin ils montrent les meilleures performances de leurs différentes approches comparées à celles de l'état de l'art dont [Wu and Oard, 2005], [Erera and Carmel, 2008] et [Dehghani et al., 2013] sur les différents corpus sus-cités. De la même façon ils comparent leurs différentes approches sur ces mêmes corpus en utilisant qu'un certain nombre de similarités de *features* comme le sujet des messages, les informations retournées par AlchemyAPI, etc.) et montrent tout aussi des détails des performances de leurs différentes méthodes supervisées sur ces corpus.

Les détails de l'article qu'on vient de présenter montrent que les auteurs ont construit ou enrichi le contexte de chaque message par exemple via des requêtes sur AlchemyAPI et les différentes *features* qu'ils ont définies. En effet pour mieux reconstruire des fils de discussions, des conversations d'emails ou de liste de diffusion, il faut bien les contextualiser, une tâche peu évidente pour des machines.

C'est dans cette esprit de contextualisation de messages ou d'emails que des chercheurs de *Yahoo* et *Amazon* [Avigdor-Elgrabli et al., 2018] étudient comment évaluer automatiquement la relation sémantique entre des messages dans une boîte de messagerie. Donner la possibilité à un utilisateur d'avoir accès à une liste de messages sémantiquement liés à un message qu'il lit ou a sélectionné dans sa boîte est leur objectif. Un défi pour eux est de proposer à l'utilisateur des messages qui soient spécifiques aux besoins de celui-ci. Ces chercheurs présentent leur travaux comme une généralisation du problème de reconstruction de fils de discussions, parce qu'ils vont au delà des cette problématique et proposent un aperçu contextuel plus large qu'un fil de discussions à partir d'un message. Ils postulent que leur objectif peut être vu comme un mécanisme de recherche implicite dans lequel le message sélectionné ou lu par l'utilisateur constitue une *requête*. Ce qui permettrait à l'utilisateur d'avoir de l'information de sa messagerie sans avoir à formuler une requête explicite.

Afin d'atteindre leur objectif, ils ont utilisé un large corpus de boîtes de messagerie d'un grand fournisseur de services de messagerie. Pour évaluer la relation sémantique entre deux messages, ils cherchent à produire un score basé sur la relation des messages telle que perçue par les humains. Ils approchent cette tâche comme un problème de classification dans laquelle ils mesurent la corrélation entre le score produit par leur modèle et un ensemble de labels positifs ou négatifs. Ils considèrent aussi leur tâche comme un problème de ranking parce qu'ils doivent fournir à l'utilisateur une liste restreinte de messages. Et donc ils ordonnent d'abord les messages candidats en fonction du score produit par leur modèle et mesure l'efficacité de leur modèle sur les k-meilleurs résultats. Pour entraîner leur modèle, ils utilisent les *features* suivantes pour chaque message :

— La différence de temps entre les dates d'envois et de réception des emails,
— Les contacts ou interlocuteurs des emails,
— Les sujets des emails représentés avec le modèle de sac de mots continus Word2Vec de [Mikolov et al., 2013] (sur deux corpus différents dont Wikipedia et un du domaine d'emails) et calcule un poids nommé ***CEN*** pour chaque sujet

$$CEN(t_i) = \frac{\sum_{w_i \in t_i} W2V(w_i).IDF(w_i)}{\sum_{w_i \in t_i} IDF(w_i)}$$

et la similarité cosinus entre deux sujets est égale à : $SIM(t_i, t_j) = COS(CEN(t_i), CEN(t_j))$
— les contenus d'emails représentés par des points communs comme les noms, les mots rares et ceux hors du vocabulaire, etc.

Leur corpus d'expérimentation contient environ 5 millions de messages qui ont été lus par des utilisateurs pendant trois mois. Sur ce corpus ils ont collecté des ensembles de mails candidats (sémantiquement proches) de chaque message source. Cela s'est fait par comparaison des mesures de différence de temps, de la similarité de Jaccard entre les contenus et sujets, et entre les contacts partagés entre deux messages. Ces scores ont permis d'agréger pour chaque message source ses 30 meilleurs candidats. Après suppression de certains messages non lus et non candidats, le corpus d'expérimentation a été réduit à environ 2 millions de messages contenant 33 000 messages sources avec chacun ses candidats. Ces messages candidats ont été labelisés positif ou négatif par une stratégie d'annotation automatique. Cette stratégie se base sur des logs de recherche d'informations et sur les dossiers crées pas les utilisateurs. A partir de ces informations, deux messages sont considérés comme en relation s'ils respectent les conditions suivantes :

— ils apparaissent dans les 20 premiers résultats d'au moins trois requêtes différentes
— Ils sont présents dans deux sessions et sont vus par un utilisateur dans un intervalle de temps de 5 minutes
— ils sont dans un même dossier (qui contient moins de 40 messages) créé par l'utilisateur

Avec ce corpus annoté, les auteurs effectuent différentes évaluations avec la mesure AUC - Aire sous la courbe ROC (*Area Under the ROC - Receiver Operating Characteristic*) qui est une mesure robuste de performance de classifieur binaire parce que son calcul repose sur la courbe ROC complète et qui implique tous les seuils de classification. Les performances des différentes représentations word2Vec des sujets des messages sur les modèles de mots de wikipedia sont respectivement de 64,4% et 66%. Ceci montre l'avantage des représentations word2Vec sur des

7. `www.facebook.com/healthychoice`
8. `www.facebook.com/WHO`
9. `www.facebook.com/groups/533592236741787`
10. `www.facebook.com/groups/848992498510493`

corpus de domaine spécifique. Par la suite avec l'algorithme de régression logistique, ils présentent les valeurs AUC sur les différentes combinaisons de *features* :

— *temps* → **0.527**
— *temps+contact* → **0.707**
— *temps+contact+sujet* → **0.755**
— *temps+contact+sujet+contenus* → **0.776**

Ils montrent aussi des évaluations de quelques méthodes de classification : arbre de décisions, naïve bayésienne, régression logistique et *Random Forest* ; entraînées et testées avec tout les *features*. il en ressort que les algorithmes Random Forest et de régression logistique sont les plus performants avec des scores AUC respectifs de **0.790** et **0.776**.

Ce second article ne détaille pas comment sont constitués et calculés ses *features* temps, contacts et contenus des messages comme le premier article de [Domeniconi et al., 2016] que nous avons présenté plus haut. Les auteurs de ce second article ne testent pas leur approche sur des corpus connus tels que Enron ou ceux utilisés dans les expériences du premier article présenté. Aussi ils ne comparent pas leur approches avec celles de l'état parce qu'ils postulent que c'est la première fois qu'une méthode évalue des relations contextuelles sémantiques entre deux emails. Et pourtant les travaux de [Domeniconi et al., 2016] détaillés plus haut tentent tout aussi d'enrichir le contexte des messages qu'ils comparent. Le second article classe les k-meilleurs résultats d'un messages, ceci peut être considéré comme un cluster de messages comme c'est le cas dans [Domeniconi et al., 2016]. Cependant, On peut retenir l'utilisation d'une représentation word2Vec sur un modèle de domaine d'emails pour le calcul de certaines similarités.

De ces deux articles, il ressort que la construction de fils de discussions ou l'établissement de relation contextuelle sémantique passent par l'exploitation de toutes les informations d'un message ou emails, de l'en-tête au corps en passant par les relations sociales entre les différents interlocuteurs. De même leurs expérimentations montrent que la méthode supervisée ***Random Forest*** est meilleure pour ce type de classification binaire. Néanmoins depuis leurs travaux, le domaine de représentation de langage a fortement évolué avec le modèle **BERT** (***Bidirectional Encoder Representations from Transformers***) basé sur les **Transformers** qui eux s'appuient sur les **mécanismes d'attention**. Au vu des performances de ces nouveaux modèles, **CAMEMBERT** [11] qui est le modèle basé sur une architecture proche de celle de BERT mais entraîné sur des corpus en français a été développé par des chercheurs de l'équipe *ALMAnaCH* [12] de chez Inria et il devrait être utilisé pour nos travaux. Mais nous serions contraints par la taille de ces modèles et les ressources matérielles qu'il demande parce que notre objectif est de trouver une solution qui fonctionnerait sur un poste de travail simple d'un collaborateur. Construire des fils de discussions simples et portés sur des thèmes bien précis nécessitent comme annoncé plus haut un démêlage de conversations imbriquées.

2.1.2 Démêlage de fils de discussions

Nous avons mentionné dans l'introduction l'importance des connaissances dissimulées dans les données conversationnelles d'outils de C&C en entreprise ou sur des plate-formes publiques forums, groupes et pages sur des réseaux sociaux. Cependant, extraire ces connaissances n'est pas une tâche simple en raison : des idées et concepts très peu ou pas explicites exprimés dans ces messages, de l'utilisation du vocabulaire spécifique en fonction du canal de communication, et des réponses aux messages dispersés au fil du temps. Dans la littérature, une autre problématique principale liée aux fils de discussion est le **démêlage de conversations**. ***Démêler des fils de discussions*** revient à identifier et séparer les messages entremêlées de conversations.

C'est dans cette optique que [Jiang et al., 2018] tirent parti de l'apprentissage de représentations de langage pour démêler les conversations. Ils procèdent en deux étapes : tout d'abord ils abordent le problème de similarité avec un algorithme d'apprentissage profond basé sur les réseaux de neurones convolutifs et ensuite ils utilisent un algorithme basé sur des scores de confiance élevés entre des paires de messages, ceci pour l'identification des conversations. La figure 2 illustre ce processus en deux étapes.
Leur approche pour le calcul de similarité entre des messages ne nécessite pas d'annoter des données. Ils formulent une hypothèse selon laquelle le temps écoulé entre deux messages ne doit pas dépasser une heure. Ainsi pour évaluer les similarités entre les messages, ils proposent leur méthode nommée ***Siamese Hierarchical Convolutional Neural Network (SHCNN)*** qui est un réseau convolutif hierachique siamois. SHCNN capture à la fois des représentations sémantiques de bas et de haut niveau d'un message. Leur architecture prend simultanément deux messages respectivement sur deux réseaux convolutifs hiérarchiques (HCNN) identiques qui créent chacun un vecteur de dimension 128 pour chacun des messages pris en *input*. Le vecteur produit est une concaténation de deux vecteurs de dimension 64 qui représentent respectivement les *features* sémantiques de bas et de haut niveau. La figure 3 montre les différentes convolutions faites par un HCNN pour l'obtention du vecteur de dimension 128 représentatifs des deux niveaux sémantiques. Une fois ces vecteurs $\widehat{m}_i$ et $\widehat{m}_j$ générés pour estimer la similarité entre eux, les auteurs exploitent l'affinité entre ces deux vecteurs dans un même espace. Ils calculent la différence absolue de k-élément des deux vecteurs $|\widehat{m}_i(k) - \widehat{m}_j(k)|$. Enfin ils utilisent la fonction **Sigmoïde** pour obtenir le

11. `https://camembert-model.fr/`
12. `http://almanach.inria.fr/index-en.html`

résultat final $\widehat{y}(m_i, m_j)$ de SHCNN qui est la probabilité de similitude entre les messages m_i et m_j. Ces calculs sont illustrés par la figure 4.

La deuxième étape de leurs travaux consiste à regrouper les conversations en fonction des probabilités de similitude calculées entre les messages. Les auteurs proposent un algorithme appelé *Conversation Identification by Similarity Ranking (CISIR)* qui s'appuient sur les meilleurs scores de similarité. CISIR construit un graphe avec des paires de messages dont le score de similarité est supérieur au seuil inférieur des scores de similarité, de cette manière des sous-graphes sont créés représentant différentes conversations d'un fil de discussion. Les auteurs évaluent leurs méthodes SHCNN et CISIR avec six approches de l'état de l'art sur 4 corpus et les résultats montrent que leur approche surpassent celles l'état de l'art.

Ces travaux de [Jiang et al., 2018] bien présentés dans leur papier, fait ressortir tout de même un problème avec leur hypothèse d'un temps écoulé entre deux messages. De cette hypothèse, résulterait dans certains cas une perte d'information très importante. Un autre point à noter dans leurs travaux est l'utilisation des messages de conversation de plus de 10 mots alors que dans certaines conversations les messages de tailles plus petites peuvent être porteuse d'informations importantes, comme par exemple des actes de dialogues. Les différentes approches présentées jusqu'ici nous montrent que les problématiques d'identification, de reconstruction de fils de discussions et celles de démêlage de conversation sont plus ou moins identiques de part les différents approches similaires mises en oeuvre pour leur résolution.

La granularité très fine ciblée dans nos objectifs pour reconstruire les fils de discussions ne pourrait être atteinte sans des études contextuelles plus fines. Ces études s'orientent notamment sur les identifications de segments de texte dans les messages de conversations qui portent généralement des actes de langage très variés. De l'identification de ces actes pourrait résulter des couches de fils de discussions représentées en acte de dialogue.

2.1.3 Segmentation de texte et Actes de dialogues

La segmentation de texte, comme son nom l'indique consiste à segmenter ou découper un contenu textuel en unités de texte de petite taille et porteuse d'un contexte sémantique bien précis. C'est une phase nécessaire dans le traitement automatique de langage pour la résolution de certains problèmes tels que l'alignement de corpus, la contraction et/ou simplification de texte, l'identification des actes de dialogue dans le cadre de l'analyse de discours, etc.

L'identification d'actes de dialogues pourrait être perçu comme la détermination des intentions des interlocuteurs dans une conversation. C'est dans ce sillage que [Wang et al., 2019] étudient l'identification d'intentions dans les courriels en situation de travail en entreprise. Le corpus **Avocado** [13] qui est un grand corpus d'emails d'une ancienne entreprise de technologie de l'information qui portait le même nom "Avocado". Ce corpus est une collection de 938 035 emails et de pièces jointes pour 279 comptes majoritairement des comptes employé de la dite entreprise. [Wang et al., 2019] utilisent ce corpus pour étudier les caractéristiques des intentions dans les emails d'entreprise et proposent une meilleure méthode d'identification de ces intentions. Ils se focalisent sur les intentions au niveau des phrases et montrent comment enrichir le contexte d'une phrase avec le contenu complet d'un email et ses méta-données. Cet enrichissement de contexte améliore les performances des modèles d'identification des intentions.

Pour caractériser les intentions d'emails, ils se basent sur certains travaux de la littérature pour définir quatre catégories générales et abstraites : échanges d'informations, gestion des tâches, planification et communication sociale. Et à chacune de ces catégories ils ont associé plusieurs intentions comme ci-dessous :

— **Échanges d'information** : partage, demande d'information
— **Gestion des tâches** : demande, promesse d'action
— **Planification** : Planifier une réunion, rappel
— **Communication sociale** : messages de salutations, notes de remerciement, etc.

13. `https://catalog.ldc.upenn.edu/LDC2015T03`

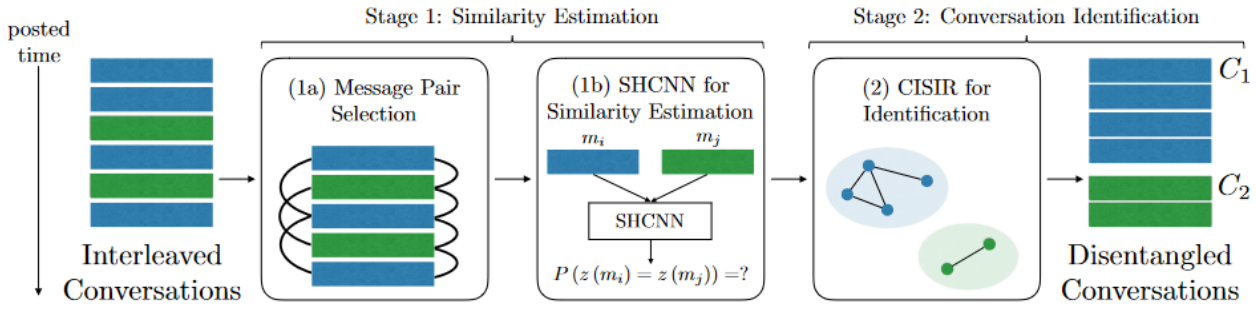

FIGURE 2 – Illustration des étapes de la méthode de [Jiang et al., 2018]

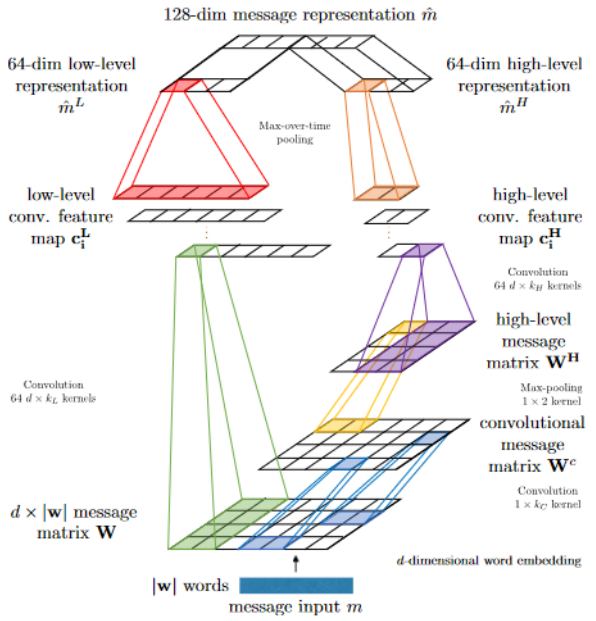

FIGURE 3 – Représentation hiérarchique d'un message via HCNN ; Les étiquettes avec une taille de police plus grande indiquent les tenseurs correspondants, et les étiquettes avec une taille de police plus petite expliquent les opérations entre les tenseurs.

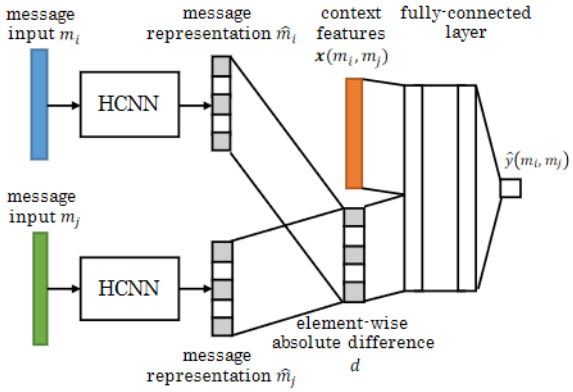

FIGURE 4 – Le CNN hiérarchique siamois (SHCNN) pour l'estimation de similitude.

Pour mieux comprendre les caractéristiques des intentions des utilisateurs, ils sélectionnent 1300 fils discussions de façon aléatoire. Ils font annoter les messages de ces fils de discussions par trois annotateurs, ceci en fonction des intentions et sous-intentions listées précédemment. Pour chaque message les annotateurs peuvent choisir plusieurs intentions et le jugement final est fait par une stratégie de vote majoritaire. Les intentions retenues pour chaque email sont celles sélectionnées par deux annotateurs. Pour l'accord inter-annotateur ils ont eu un score Kappa de **0.694**. La figure 5 montre la distribution des intentions sur le corpus sélectionné et annoté. Cette même figure montre les différentes intentions et leur fréquence d'apparition dans l'échantillon choisi. Elle montre par exemple que l'*échange d'information* et la *gestion des tâches* sont les plus fréquentes dans un corpus d'emails d'entreprise. D'autres analyses de leur échantillon de données montrent que : **55,2%** de messages contiennent une intention unique, **35,8%** contiennent deux intentions et enfin **9%** contiennent au minimum 3 intentions. Ils observent que certaines intentions sont très corrélées. La figure 6 montre la co-occurrence des différentes intentions et les relations entre chaque paire d'intentions. Sur cette même figure, on observe que le *partage d'information* et la *demande d'information* sont très susceptibles d'apparaître dans un même email, tandis que le *social* et le *rappel/planification de meeting* sont peu susceptibles de se retrouvés dans un même email.

Dans l'article, ils font une étude sur l'effet de levier du contexte pour la détection d'intentions dans un email. Pour cela, ils utilisent une seule intention : la *demande d'information* et étudient l'effet du contexte sur les performances humaines pour l'identification de la présence ou non de cette intention dans une phrase. Ils utilisent le contenu complet d'un email comme contexte. Ils sélectionnent 540 emails à parti de la vérité de terrain précédemment constitué, de telle sorte que la moitié d'entre eux possèdent des labels positifs avec l'intention *demande d'information*. Et donc cette échantillon a été envoyé à deux groupes, le premier groupe avait les phrases cibles (celles annotées par l'intention dans la première expérience) et le contenu complet du message, alors que le second groupe n'avait accès qu'aux phrases cibles. Chaque instance a été annoté par trois annotateurs et la majorité des annotations pour chacune des instances représentent sa prédiction par un humain.

Le tableau 1 montre que les phrases vrais positives bénéficient significativement du contexte (contenu entier d'un email). Toutes ces expérimentations montrent que les annotateurs humains identifient beaucoup mieux les intentions dans les phrases quand elles sont fournies avec un contexte. Ces résultats démontrent l'intérêt d'utiliser le contexte pour entraîner des modèles d'identification d'intentions dans les emails.

Ils proposent un framework pour cette tâche d'identification d'intentions. Le framework en question s'appuie principalement sur le contexte d'un email pour identifier les intentions dans une phrase. Et donc il possède trois composants principaux : un encodeur de phrase, de contexte et un couche de fusion de ces deux encodeurs ; comme sur la figure 7. Les représentations d'occurrence de mots en **n-grams** (tri-grammes) tant pour les composants *Sentence encoder* et *context encoder* sont utilisées. La couche de fusion des *features* consistent en une concaténation des *features* des contextes de l'encodeur de la phrase et du contexte. Ensuite les *features* concaténées sont

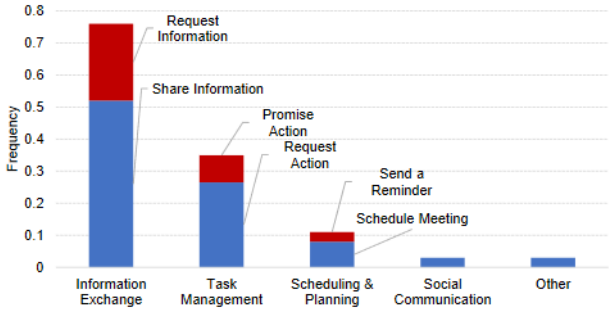

FIGURE 5 – Fréquence des intentions sur un sous-ensemble du corpus Avocado

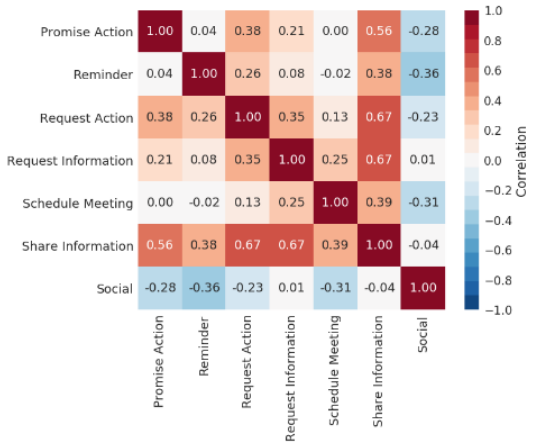

FIGURE 6 – Distribution de paire de sous-intentions dans un même email

Predictions	True Positive	True Negative
Positive	175(%32.4)	14(%2.6)
Negative	95(%17.6)	256(%47.4)

TABLE 1 – Matrice de confusion des prédictions humaines

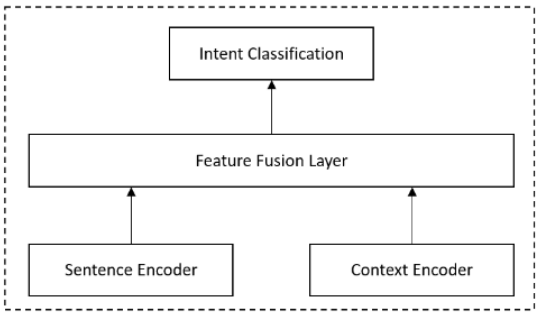

FIGURE 7 – Augmentation de contexte d'une phrase pour y identifier les intentions

passées dans un classifieur. Ils ont utilisé deux méthodes traditionnelles d'apprentissage automatique (**régression logistique - LR** et **machine à vecteurs de support - SVM**).

Ils ont aussi expérimenté des approches d'apprentissage profond. Pour la composante qui encode la phrase cible, ils transforment chaque mots de la dite phrase, avec la technique de words embeddings en une matrice $W \in R^{d \times |V|}$. Ici $|V|$ est la taille du vocabulaire et d est la dimension du plongement de mot. Par la suite ils font passer les plongements de mots de la phrase dans un réseaux de neurones récurrents bidirectionnel avec des cellule GRU (*Gated Recurrent Unit*). Ce réseau bidirectionnel produit pour chaque mot w_i de la phrase cible un vecteur $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$ avec $\overrightarrow{h_i}$ et $\overleftarrow{h_i}$ respectivement l'état caché avant et arrière du mot w_i. Ces états permettent d'avoir des informations avant et à après chaque mot. Pour encoder le contexte (message complet de l'émail) d'une phrase s , la composante d'encodage du contexte encode chaque phrase du contexte en un vecteur de taille fixe. Ils utilisent une opération d'attention sur des mots spécifiques tels que *réunion, point, discussion, se réunir* par exemple dans le cadre d'une intention de planification de de réunion. Cette opération d'attention se calcule avec un vecteur α en fonction des couches H (créées avec le processus définis pour la phrase cible), un hyper-paramètre u_s, une matrice de poids de la phrase W_s et un vecteur biais b_s.

$$\alpha = Softmax(u_s Tanh(W_s H^T + b_s))$$

La représentation d'une phrase du contexte s'obtient par $r_s = \alpha H$. La couche de fusion de ces représentations de phrase et de contexte quant à elle produit un vecteur augmenté de la phrase cible s dont on veut identifier les intentions inhérentes. Ils calculent ainsi une matrice $A = F(H_s, R_c) \in R^{L \times N}$, avec H_s les états cachés de la phrase cible, R_c la représentation du contexte, L la longueur de la phrase cible, N le nombre de phrases dans le contexte. Chaque élément A_{ij} se calcule comme suit :

$$A_{ij} = w_c^T [H_s^i; R_c^j; H_s^j \circ R_c^j]$$

où w_c est un vecteur de poids entraînable, [;] une concaténation de vecteurs et $\circ$ la multiplication par éléments. Dans cette matrice A, chaque ligne représente un token de la phrase cible et chaque élément spécifie la pertinence de chaque phrase du contexte sur ce token. Avec des opérations de normalisation de cette matrice comme dans les formules de calcul de α et r_s présentées plus haut, [Wang et al., 2019] génèrent un vecteur contextuel v_s

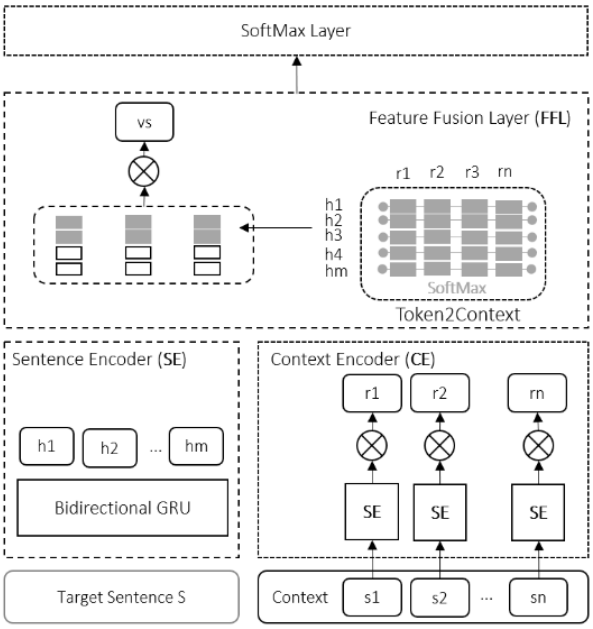

FIGURE 8 – Aperçu de DCRNN

	SM	PA	RI
LR : Sent	66.86	74.73	74.71
SVM : Sent	66.24	73.56	75.45
CNN : Sent	67.12	75.21	73.15
LR : Sent + Cont	66.302	74.75	74.76
SVM : Sent + Cont	64.24	71.53	75.20
DCRNN	73.48‡	80.42‡	78.37‡

FIGURE 9 – Aperçu de DCRNN

représentatif de la phrase cible et de son contexte. Ce vecteur est ensuite passé dans une couche *softmax* afin d'effectuer la prédiction. Cette dernière couche calcule une probabilité p avec W_p, b_p des paramètres de cette couche.

$$p = Softmax(W_p v_s + b_p)$$

Ils utilisent la fonction de perte d'*entropie croisée* pour entraîner le modèle. Ils illustrent leur modèle complet par l'image de la figure 8. Ils nomment leur modèle **Dynamic-Context Recurrent Neural Network- DCRNN** qui peut se traduire par *Réseau de neurones récurrents à contexte dynamique*.

Enfin ils expérimentent les différents méthodes présentées ci-dessus pour trois intentions : *demande d'information (RI)*, *planification de réunion (SM)* et *promesse d'action (PA)* avec respectivement des corpus de 7080, 47914 et 9076 d'emails. Le tableau de la figure 9 présente les différents scores $F1$ avec et sans contexte, obtenus à partir des approches traditionnelles d'apprentissage automatique et celles basées sur les réseaux de neurone récurrent (RNN). On y observe que ces RNN sont meilleurs que les anciennes approches tant sur les prédictions avec ou sans contexte.

Cet article nous montre une fois de plus l'importance du contexte dans la problématique dans nos travaux. Ils démontrent dans ce papier au travers de différentes étapes d'annotation l'importance du contexte dans l'analyse d'un email. Leur approche détaillé d'encodage du contexte semble intéressante, mais nécessite d'être comparée avec les approches de modèle les plus performants comme *BERT* qui donneront probablement de meilleurs résultats. L'identification d'intentions dans les emails semble être une bonne piste pour nous parce qu'on pourrait s'y appuyer pour construire des relations entre intentions dans différents messages d'un même cluster. On pourrait envisager créer des fils de discussions avec ces intentions, ces fils de discussions seront en effet simples et explicites parce qu'un acte de dialogue ou une intention possède une information qui s'interprète facilement. Ces identifications d'intentions peuvent aussi aider au démêlage de conversations, une problématique que nous avons abordé plus haut.

Nous venons ainsi de faire un aperçu de quelques approches utilisées dans la littérature pour la résolution des problématiques de construction de fils de discussions, de démêlage de conversations, de segmentation et d'identification d'actes de langage. Nous avons constaté dans la lecture de ces articles qu'il faut toujours procéder à des expérimentations pour montrer les performances des approches proposées. Mais ces expérimentations utilisent des corpus qui existent déjà (Enron, Avocado, etc.) ou qui sont créés pour l'occasion et serviront pour des futurs travaux. Dans notre cas, vu qu'il est question à la clé de nos travaux de proposer des solutions concrètes qui faciliteraient les tâches des collaborateurs de chez Orange, il va de soit que nous devons effectués des expérimentations sur des données réels d'entreprise et c'est pour cette raison que nous avons entrepris de construire notre corpus. Dans les prochaines sections, nous allons nous intéresser à la constitution de corpus, ainsi que des aspects connexes comme les contraintes juridiques, la collecte de données, leur annotation, etc.

3 Constitution de corpus

Construire un corpus est l'étape primaire dans tout processus d'analyse de données. Cette tâche est difficile à mener suivant le type de données privées ou publiques, orales ou écrites et du support de ces données. Dans le cadre de l'analyse discursive, le développement d'internet et des outils de C&C en entreprise a mis à disposition des chercheurs des données issues principalement de forums et d'emails. Constituer un corpus nécessite une démarche méthodologique spécifique en fonction des objectifs visés. Ces objectifs sont généralement liés à des problématiques comme celles abordées dans les précédentes sections : identification des unités de discours, détection et démêlage de fils de discussions. Des corpus ont été construits dans le cadre la résolution de ces problématiques.

3.1 Corpus existants

Dans le cadre d'un projet de Coordination en équipe, un jeu de gestion simulé par 277 étudiants d'un MBA à l'université Carnegie Mellon a produit le corpus CSpace constitué d'environ 15 000 emails. Ce corpus a été utilisé dans plusieurs travaux d'analyses d'emails. [Cohen et al., 2004] utilisent ce corpus afin de proposer des méthodes de classification supervisée d'emails. Le corpus d'emails le plus utilisé dans le cadre des recherches d'informations, de reconstruction de fils de discussions est le corpus d'entreprise **Enron** construit par [Klimt and Yang, 2004] . Ils l'ont construit et analysé sa pertinence par rapport à la prédiction des dossiers d'emails. Ce corpus reste le corpus de référence pour des analyses de conversations. Ils contient des données brutes représentant environ 1 361 403 messages appartenant à 158 boîtes aux lettres de 149 personnes. Ce corpus sert à de nombreuses analyses comme la détection des différentes parties ou zones d'un email [Jardim et al., 2021] ; l'extraction de connaissance [Zhou et al., 2008], la reconstruction de fils de discussions [Yeh, 2006], [Wang et al., 2008] ; les comportements des interlocuteurs [Joorabchi et al., 2010] et les relations sociales entre eux [Kang et al., 2010] ; la résolution d'entité dans les conversions d'emails [Dakle and Moldovan, 2020], etc.

Plusieurs de ces travaux ont utilisé des approches supervisées, qui nécessitent des données d'entraînement produites par l'annotation des corpus. L'annotation des données est une tâche fastidieuse qui est généralement faites par plusieurs personnes et qui s'évalue par une métrique largement utilisée dans les travaux sus-cités : l'accord inter-annotateurs ou coefficient Kappa de Cohen qui évalue le degré d'accord (ou de concordance) entre des annotateurs quant à l'attribution à des classes des segments de texte de corpus. Les multiples travaux que nous avons mentionnés jusqu'ici ne traitent que des corpus d'emails en langue anglaise et publiquement disponibles. Très peu de travaux utilisent des corpus d'emails d'entreprise en français parce que ces derniers sont rares. Pour des corpus en langue française, certains travaux se sont intéressés aux problématiques liées à l'analyse de conversations avec des corpus d'emails de plateformes de forum et liste de diffusion. [Hernandez et al., 2016], dans le cadre du projet ODISAE ont produit le corpus Ubuntu-fr en français constitué de forums de discussion, de listes de diffusion, de canaux *Internet Relay Channel (IRC)* de la plateforme Ubuntu-fr [14] . Ces travaux se sont inscrits dans le cadre d'une meilleure compréhension de la structure et du fonctionnement de ces différents canaux de communications et de leurs interactions. Dans le même ordre idée pour l'analyse de discours, [Bevendorff et al., 2020] ont introduit le corpus public Webis Gmane Email Corpus 2019, le plus grand corpus en date et complètement traité et segmenté en 15 classes sémantiques de courriels. Il est constitué d'environ 153 million multilingues d'emails (avec 1.8 million d'emails en langue française), le tout extrait de 14 699 listes de diffusion de la plateforme *gmane.io*. Nous avons dans le cadre de nos travaux extrait les emails en langue française de ce corpus afin d'en étudier la consistance en terme de conversations. Des analyses plus approfondies de ce corpus sont à effectuer très prochainement. Collecter et constituer de tels corpus de données dans le cadre de nos travaux est une tâche complexe parce que nous nous intéressons à des contenus privés et à caractère personnels qui sont soumis à des contraintes juridiques.

3.2 Aspects juridiques et collecte de données

Au vu de la rareté de corpus d'emails d'entreprise, nous avons entrepris de constituer dans un premier temps un corpus d'emails d'entreprise et par la suite les contenus d'outils de communication et collaboration telles que Slack, Mattermost, etc. seront tout aussi collectés pour nos analyses futures. Cependant le respect de la loi sur le secret des correspondances, de la confidentialité et du règlement général sur la protection des données (RGPD) a constitué des contraintes dans notre démarche.

Néanmoins, une collaboration avec des juristes de chez Orange, a donné lieu à un processus des demandes d'accord de consentement à des collaborateurs cibles. Ce processus est le résultat d'une analyse de risques que nous avons menée en tenant compte de tous les intervenants et leurs rôles dans la chaîne de manipulation des données. Il se déroule en deux étapes. La première consiste à l'identification de collaborateurs à qui on adresse une demande de consentement et la seconde quant à elle, est axée sur la collecte effective d'emails sur des postes de certains collaborateurs. Les collaborateurs sont identifiés principalement en fonction de leur impact sur le nombre de conversations et d'emails dans la messagerie d'un autre collaborateur. Ils sont identifiées à partir d'un outil que nous avons développé qui se connecte à une messagerie Outlook ou à un fichier d'archive de cette même messagerie.

14. `www.ubuntu-fr.org`

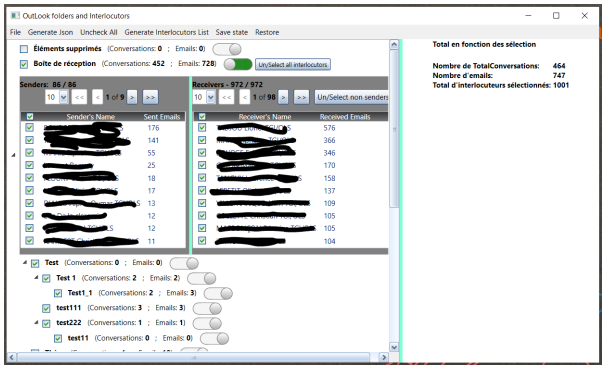

FIGURE 10 – Interface de l'outil développé pour identifier les collaborateurs et extraire les emails

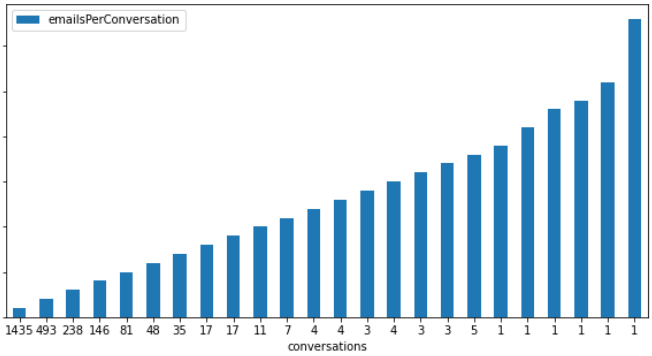

FIGURE 11 – Statistiques d'emails par conversation sur 5670 emails collectés jusqu'ici sur le poste de 4 collaborateurs

L'outil (Figure 10) liste les différents dossiers et sous-dossiers de la boîte de messagerie et donne la possibilité à un collaborateur de dé/sélectionner des dossiers en fonction des emails qu'il aimerait mettre à disposition de nos travaux. Une fois les dossiers sélectionnées, le collaborateur peut dé/sélectionner les interlocuteurs de sa messagerie. A chaque dé/sélection de dossier ou d'interlocuteurs, l'outil calcule le nombre de conversations, d'emails et d'interlocuteurs et permet ainsi de maximiser le nombre d'emails et de conversations à collecter tout en réduisant le nombre de collaborateurs à contacter par la suite. Ainsi de proche en proche entre collaborateurs, ce processus nous a permis jusqu'ici sur 4 postes de travail et pour une quarantaine d'interlocuteurs qui ont donné leur accord de consentement de collecter environ 5670 emails. La figure 11 montre la répartition des conversations en fonction du nombre d'emails qu'ils contiennent.

3.3 Méta-données, annotation et pseudo-anonymisation

Les emails collectés possèdent des méta-données qui correspondent aux élément d'en-têtes. Dans le cadre des emails extraits de Outlook, on distingue :

— ConversationID qui est identique pour des emails d'une même conversation
— ConversationIndex, unique pour un email
— La date d'envoi
— les interlocuteurs : émetteur et destinataires d'un email

Ces méta-données permettent d'avoir un aperçu statistique moyennant quelques manipulations ou traitement de ces méta-données de notre corpus. Dans ces statistiques, on peut extraire des informations comme celles de la figure 11. Il est aussi possible à partir de ces méta-données de construire des fils de discussions avec une granularité faible, c'est-à-dire telle qu'elles sont structurées dans la boîte de messagerie. Elles nous aideront aussi très prochainement à la mise en place d'un outil pour visualiser des clusters de conversations, d'emails sous forme de timeline. Mais aussi elles vont permettre de construire un graphe d'interactions sociales entre les différents interlocuteurs du corpus, avec des arcs pondérés par le nombre d'emails échangés entre ces interlocuteurs. Nous avons mené une telle expérience sur un petit corpus d'emails avec l'outil **Neo4J** qui nous a permis de créer un graphe avec des noeuds de conversations, d'emails et d'interlocuteurs.

Les méta-données sur les interlocuteurs vont permettre ou faciliter le processus de pseudo-anonymisation du corpus. Nous avons développé des scripts qui créent progressivement en parcourant ces méta-données une table de correspondance contenant les noms/adresse mail des interlocuteurs sur une colonne et sur l'autre des chaîne de caractères qui présentent en fait les valeurs pseudo-anonymisées des interlocuteurs.

Pour élargir cette pseudo-annymisation sur les contenus des corpus, il faut identifier dans les corpus des entités nommées. Pour ce faire nous avons testé des outils de reconnaissance d'entité nommées (**REN**) tels que *SEM* [Dupont and Plancq, 2017], l'*EntityRecognizer*[15] de Spacy et un outil basé sur les **Champ aléatoire conditionnel, Conditional Random Fields (CRF)** développé chez Orange lors d'un stage en 2019. Ce dernier outil avec un processus itératif et une interface d'annotation, réduit fortement la tâche d'annotation tout en améliorant son modèle CRF. Il a été entraîné lors du dit stage sur un corpus d'emails certes de petite taille mais identique à celui que nous collectons pour nos travaux. Un processus d'annotation de 1k emails (extrait des 11k emails collectés jusqu'à présent en ce mois d'août 2021) a été effectué par trois annotateurs qui sont des collaborateurs proches de notre équipe chez Orange.

Tout en améliorant la reconnaissance d'entités nommées sur notre corpus d'emails, nous allons mettre à jour notre script de pseudo-anonymisation, afin qu'il puisse utiliser les entités annotées dans le corpus, les croiser avec les correspondances des interlocuteurs et générer des codes de correspondance pour les autres entités (projets, nom d'équipes, application, etc). Notons que des glossaires d'Orange contenant des acronymes par exemple vont être

15. https ://spacy.io/api/entityrecognizer

utilisés pour l'amélioration de notre modèle de REN.
Cette tâche de pseudo-anonymisation nous donne un aperçu de nos objectifs développés dans la prochaine section.

Tout ce processus méthodologique de constitution de corpus d'entreprise incluant une phase de pseudo-anonymisation afin de respecter le RGPD et le secret de la correspondance a donné lieu à rédaction d'un article qui a été accepté en cette fin de mois d'août 2021 au Student Research Workshop de la conférence RANLP 2021 et sera présenté à la dite conférence le 02 ou 03 Septembre 2021.

4 Objectifs à court, moyen et long terme

En termes d'objectifs à court terme, c'est-à-dire dans les semaines à venir, nous nous concentrerons à affiner la chaîne de pseudo-anonymisation à partir du corpus collecté jusqu'ici. Cette chaîne sera déployée sur les données qui seront collectées plus tard pour la complétion de notre corpus. Une autre tâche dans les prochaines semaines consistera à mettre en place un outil de visualisation de nos données, plus précisément les fils de discussions avec une granularité faible, ceci à parti des méta-données de notre corpus. Le type de visualisation en timeline sera implémenté pour distinguer de manière explicite des zones de forte échanges conversationnelles dans notre corpus. La mise à jour progressive du document constituant l'état de l'art de nos travaux s'effectuera progressivement avec l'avancée de nos travaux.

À court et à moyen terme, des analyses approfondies de corpus et expérimentations seront faites sur la base des articles que nous avons présentés plus haut moyennant des adaptations et des améliorations. Ces analyses donneront lieu à la création de modèles d'IA spécifiques à notre problématique. Les résultats de ces analyses seront présentés au travers d'articles scientifiques qui seront publiés et vulgarisés. Un prototype d'application intégrant ces modèles d'IA sera développé pour un cas d'utilisation concret.

Pour la dernière année de thèse, les six premiers mois seront tournés vers des améliorations des modèles qui auront été produits. Lesquelles donneront lieu à d'autres publications scientifiques. Aussi des prototypes seront mis à jour et testés. Les six derniers mois de la thèse serviront à la rédaction du manuscrit de thèse.

4.1 Formations doctorales et autres

Dans le cadre de mon plan individuel (PIF) dans lequel j'ai renseigné mon objectifs après thèse. Cet objectif d'après thèse qui reste partagé entre l'entrepreneuriat et un poste d'ingénieur recherche dans un laboratoire R&D. Dans le cadre d'une préparation à ces objectifs, mais aussi pour mieux avancer sur ma thèse et comprendre certaines aspects de la recherche, j'ai suivi ou je suivrai les cours suivants :

- Open Data : Aspects généraux des données de la recherche : 04/03/2021(3h)
- Réaliser un poster scientifique : 11/25 Mars 2021 (4h)
- Communiquer efficacement à l'écrit : 22/23 Mars 2021 (12h)
- Le circuit de la publication scientifique : 25/03/2021 (3h)
- Undersatnding and applying the entrepreunarial method - 31/03/2021 (4h)
- Impact et diffusion de la recherche, enjeux et limites : 02/04/2021 (3h)
- Lecture rapide : Quadriller l'information - 06, 07 et 08 Avril 202 (12h)
- Décidez pour prioriser, gérez votre temps et votre stress - 14/04/2021 (6h)
- Éthique de la recherche scientifique (pas encore inscrit)

J'ai aussi suivi des cours d'apprentissage automatique et profond orientés sur le traitement de langage naturel sur les plateformes *OrangeLearning*, *Udemy* et *Coursera*.

5 Conclusion

Dans ce rapport, nous avons présenté ce qui a meublé la première année de thèse. De l'état de l'art aux différentes formations suivies, en passant par la collecte de données et les objectifs dans les semaines, mois et année à venir. Des articles qui ont particulièrement attiré notre attention, parce que très proches de nos objectifs, ont été présentés.

La reconstruction de fils de discussions ou le démêlage de conversation, l'identification des actes de dialogues et la segmentation de texte sont les principales problématiques dans lesquelles nous nous investirons d'avantages. Nous proposerons des solutions pour les résoudre avec les différentes approches existantes, notamment la représentation contextuelle de langage basée sur les modèles de représentation vectorielle des texte. On a relevé aussi que de tels modèles aidaient d'avantages pour les similarités entre messages quand ils sont entraînés sur des domaines spécifiques de métier. Ce sont par exemple Word2Vec utilisé dans un article, détaillé plus haut. Mais il existe dans la littérature de nouveaux modèles meilleurs que ceux présentés plus haut. Et donc un challenge pour nous serait lors de notre phase d'expérimentation d'utiliser ces nouveaux modèles. Mais surtout en prenant en compte que de tels modèles sont très demandeur de ressources matérielles, et donc nous pourrions nous orienter vers leurs versions compressées comme DistillBERT [Sanh et al., 2019] ou MobileBERT [Sun et al., 2020].

En termes d'objectifs à atteindre, plusieurs tâches seront à effectuer très prochainement : La mise en place d'une chaîne de pseudo-anonymisation de corpus ; la rédaction d'état de l'art et d'articles, des expérimentations à mener sur nos données collectées et du prototypage. Dans ce rapport, les formations dans le cadre du programme individuel de formation et aussi dans le cadre de ma montée en compétences sur les technologies de traitement automatiquement de langage, ont été listées. Certaines seront à faire dans les semaines à venir.

Tous les points abordés dans ce rapport font ainsi état de mes activités de thèse lors de la première années et des objectifs à atteindre.

Références

Noa Avigdor-Elgrabli, Roei Gelbhart, Irena Grabovitch-Zuyev, and Ariel Raviv. More than threads : Identifying related email messages. CIKM '18, page 1711–1714, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi : 10.1145/3269206.3269255. URL `https://doi.org/10.1145/3269206.3269255`.

Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein. Crawling and preprocessing mailing lists at scale for dialog analysis. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1151–1158. Association for Computational Linguistics, 2020. doi : 10.18653/v1/2020.acl-main.108. URL `https://doi.org/10.18653/v1/2020.acl-main.108`.

Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. Efficient agglomerative hierarchical clustering. *Expert Syst. Appl.*, 42(5) :2785–2797, April 2015. ISSN 0957-4174. doi : 10.1016/j.eswa.2014.09.054. URL `https://doi.org/10.1016/j.eswa.2014.09.054`.

William W. Cohen. Learning rules that classify e-mail. In *Proceedings of the 1996 AAAI Spring Symposium on Machine Learning and Information Access*, 1996.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. Learning to classify email into speech acts. In *In Proceedings of Empirical Methods in Natural Language Processing*, 2004.

Parag Dakle and Dan I. Moldovan. Cerec : A corpus for entity resolution in email conversations. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 339–349. International Committee on Computational Linguistics, 2020. ISBN 978-1-952148-27-9. URL `https://www.aclweb.org/anthology/2020.coling-main.30/`.

Mostafa Dehghani, Azadeh Shakery, Masoud Asadpour, and Arash Koushkestani. A learning approach for email conversation thread reconstruction. *J. Inf. Sci.*, 39(6) :846–863, 2013. URL `http://dblp.uni-trier.de/db/journals/jis/jis39.html#DehghaniSAK13`.

Giacomo Domeniconi, Konstantinos Semertzidis, Vanessa Lopez, Elizabeth M. Daly, Spyros Kotoulas, and Gianluca Moro. A novel method for unsupervised and supervised conversational message thread detection. In *Proceedings of the 5th International Conference on Data Management Technologies and Applications*, DATA 2016, page 43–54, Setubal, PRT, 2016. SCITEPRESS - Science and Technology Publications, Lda. ISBN 9789897581939. doi : 10.5220/0006001100430054. URL `https://doi.org/10.5220/0006001100430054`.

Yoann Dupont and Cl'ement Plancq. Un 'etiqueteur en ligne du français. In *24e Conf'erence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 15–16, 2017.

Shai Erera and David Carmel. Conversation detection in email systems. In *ECIR*, 2008.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner. Clarity : Inferring discourse structure from speech. In *STANFORD UNIVERSITY*, pages 25–32. AAAI Press, 1998.

Nicolas Hernandez, Soufian Salim, and Elizaveta Loginova Clouet. Ubuntu-fr : a Large and Open Corpus for Supporting Multi-Modality and Online Written Conversation Studies. In *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1777–1783, Portorož, Slovenia, May 2016. URL `https://hal.archives-ouvertes.fr/hal-01503811`.

Bruno Jardim, Ricardo Rei, and Mariana S. C. Almeida. Multilingual email zoning, 2021.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. Learning to disentangle interleaved conversational threads with a Siamese hierarchical network and similarity ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi : 10.18653/v1/N18-1164. URL https://www.aclweb.org/anthology/N18-1164.

M. E. Joorabchi, J. Yim, M. E. Joorabchi, and C. D. Shaw. Enron case study : Analysis of email behavior using emailtime. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 235–236, 2010. doi : 10.1109/VAST.2010.5649905.

Hyunmo Kang, Catherine Plaisant, Tamer Elsayed, and Douglas Oard. Making sense of archived e-mail : Exploring the enron collection with netlens. *Journal of the American Society for Information Science and Technology*, 61 : 723 – 744, 2010/04/01/ 2010. doi : 10.1002/asi.21275. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.21275/full.

Bryan Klimt and Yiming Yang. The Enron Corpus : A New Dataset for Email Classification Research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Machine Learning : ECML 2004*, pages 217–226, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30115-8.

David D. Lewis and Kimberly A. Knowles. Threading electronic mail : A preliminary study. *Inf. Process. Manage.*, 33(2) :209–217, March 1997. ISSN 0306-4573. doi : 10.1016/S0306-4573(96)00063-5. URL https://doi.org/10.1016/S0306-4573(96)00063-5.

David Dolan Lewis. *Representation and Learning in Information Retrieval*. PhD thesis, USA, 1992. UMI Order No. GAX92-19460.

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.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013. URL http://arxiv.org/abs/1301.3781.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up ? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, July 2002. doi : 10.3115/1118693.1118704. URL https://www.aclweb.org/anthology/W02-1011.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT : smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

John R. Searle. A taxonomy of illocutionary acts. In Keith Gunderson, editor, *Language, Mind and Knowledge*, pages 344–369. University of Minnesota Press, 1975.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert : a compact task-agnostic bert for resource-limited devices, 2020.

Jan Ulrich and Gabriel Murray. A publicly available annotated corpus for supervised email summarization. In *In Proc. of AAAI EMAIL2008 Workshop*, 2008.

Wei Wang, Saghar Hosseini, Ahmed Hassan Awadallah, Paul N. Bennett, and Chris Quirk. Context-aware intent identification in email conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 585–594, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi : 10.1145/3331184.3331260. URL https://doi.org/10.1145/3331184.3331260.

X. Wang, M. Xu, N. Zheng, and M. Chen. Email conversations reconstruction based on messages threading for multi-person. In *2008 International Workshop on Education Technology and Training 2008 International Workshop on Geoscience and Remote Sensing*, volume 1, pages 676–680, 2008. doi : 10.1109/ETTandGRS.2008.321.

Yejun Wu and Douglas W. Oard. Indexing emails and email threads for retrieval. SIGIR '05, page 665–666, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930345. doi : 10.1145/1076034.1076180. URL https://doi.org/10.1145/1076034.1076180.

Jen-Yuan Yeh. Email thread reassembly using similarity matching. In *CEAS 2006 - The Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California, USA*, 2006. URL http://www.ceas.cc/2006/listabs.html#7.pdf.

Yingjie Zhou, Malik Magdon-Ismail, William A. Wallace, and Mark K. Goldberg. A generative model for statistical determination of information content from conversation threads. In Christopher C. Yang, Hsinchun Chen, Michael Chau, Kuiyu Chang, Sheau-Dong Lang, Patrick S. Chen, Raymond Hsieh, Daniel Zeng, Fei-Yue Wang, Kathleen M. Carley, Wenji Mao, and Justin Zhan, editors, *Intelligence and Security Informatics, IEEE ISI 2008 International Workshops : PAISI, PACCF, and SOCO 2008, Taipei, Taiwan, June 17, 2008. Proceedings*, volume 5075 of *Lecture Notes in Computer Science*, pages 331–342. Springer, 2008. doi : 10.1007/978-3-540-69304-8_33. URL https://doi.org/10.1007/978-3-540-69304-8_33.