Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Constitution de sous-fils de conversations d'emails

Lionel Tadonfouet Tadjou[1, 2, 3] Eric De La Clergerie[1]
Fabrice Bourge[2] Tiphaine Marie[2]
(1) Inria, Paris, France
(2) Orange Innovation, Caen, France
(3) Sorbonne Université, Paris, France
{lionel.tadonfouet,Eric.De_La_Clergerie}@inria.fr
{fabrice.bourge, tiphaine.marie}@orange.com

RÉSUMÉ

Les conversations d'emails en entreprise sont parfois difficiles à suivre par les collaborateurs car elles peuvent traiter de plusieurs sujets à la fois et impliquer de nombreux interlocuteurs. Pour faciliter la compréhension des messages clés, il est utile de créer des sous-fils de conversations. Dans notre étude, nous proposons un pipeline en deux étapes pour reconnaître les actes de dialogue dans les segments de texte d'une conversation et les relier pour améliorer l'accessibilité de l'information. Ce pipeline construit ainsi des paires de segments de texte transverses sur les emails d'une conversation facilitant ainsi la compréhension des messages clés inhérents à celle-ci. A notre connaissance, c'est la première fois que cette problématique de constitution de fils de conversations est abordée sur les conversations d'emails. Nous avons annoté le corpus d'emails BC3 en actes de dialogues et mis en relation les segments de texte de conversation d'emails de BC3.

ABSTRACT

Email conversations in the workplace are sometimes difficult to follow by collaborators because they can deal with multiple topics and involve many interlocutors. To improve understanding of key messages, it's helpful to create subthreads within the conversation. In our study, we propose a two-stage pipeline to recognize dialogue acts in email text segments and link them to improve information accessibility. This pipeline creates pairs of text segments across the conversation, making it easier to understand the key messages. To our knowledge, this is the first time this issue of creating conversation threads has been addressed in email conversations. We annotated the BC3 corpus of emails with dialogue acts and linked conversation email text segments.

MOTS-CLÉS : fils de conversations, emails, acte de dialogues, appariement d'énoncés, corpus, SetFit.

KEYWORDS: Conversation threads, emails, dialogue acts, utterances pairing, corpus, SetFit.

1 Introduction

Depuis quelques décennies, avec l'évolution d'internet et l'avènement de l'intelligence artificielle, les conversations Médiées par Ordinateur (CMO, *Computer Mediated Communication* – CMC en anglais) sont d'intérêt pour des recherches en linguistique, psychologie et dans bien d'autres domaines. Les

emails, les chats et les échanges dans des forums font partie de ces conversations et sont des canaux d'échanges utilisés dans des entreprises via des outils de communications et de collaboration. Les contenus issus de ces outils regorgent d'importantes connaissances et le fait qu'ils soient peu ou pas structurés limite leur exploitation pour en extraire leur quintessence. Une problématique générale induite par le besoin d'une meilleure compréhension ou l'extraction de connaissances de ces contenus dans le cadre des emails est la reconstruction de fils de conversations d'emails.

Un fil de conversation dans un corpus d'e-mails est formellement défini comme un ensemble d'emails échangés sur le même sujet entre le même groupe de personnes via des actions de réponse ou de transfert (Erera & Carmel, 2008). Pour (Dehghani *et al.*, 2012) il existe deux types de structure de conversation d'emails :

— Linéaire : les emails appartenant à la même conversation sont détectés et disposés dans l'ordre chronologique, formant une structure à une seule branche.
— Arborescente : Dans une conversation, les utilisateurs peuvent choisir de répondre à un email précis déjà existant dans la conversation produisant ainsi une structure en arbre avec une racine et ses branches.

Reconstruire un fil de conversation d'emails consiste ainsi à produire soit la structure linéaire ou arborescente permettant ainsi une meilleure compréhension du contenu de ladite conversation. Plusieurs travaux ont approché la problématique de reconstruction de fils de conversation d'emails sous des trois prismes différents. Tout d'abord, l'algorithme de Zawinski [1] aborde le problème en s'appuyant uniquement sur les méta-données pour la construction de fils de conversation. Ensuite il y a des approches qui se basent sur les contenus afin de regrouper les emails en conversations avec des structures linéaires ou arborescentes. Enfin l'identification des thématiques dans les conversations d'emails sert aussi de base pour une reconstruction de fils de conversations d'emails.

Ces travaux reconstruisent les structures de fils de conversation d'emails permettant une meilleure lisibilité des contenus desdites conversations et une identification des relations parent/enfant entre les emails d'une même conversation. Cependant ils ne permettent pas d'avoir un accès à l'essence des informations contenues dans une conversation. Aussi ces approches ne permettent pas de facilement suivre l'évolution d'une conversation, ni de savoir quelles sont les principales actions menées par les interlocuteurs dans de telles conversations d'emails. Ces actions fortement lieés aux actes de dialogues exprimés dans les messages des interlocuteurs, permettraient de cartographier la progression d'un projet avec en plus les différentes contributions des collaborateurs. Une conversation en plus de permettre des échanges sur des thématiques, est avant tout une communication entre des interlocuteurs, d'où l'existence des actes de dialogue. L'évolution d'une conversation peut par exemple répondre aux questions suivantes : est-ce que les questions posées en amont dans la conversation ont été répondues ou non ; est-ce que des approbations ou désaccords ont été émis en retour à des suggestions exprimées.

Les valeurs ajoutées qui résulteront de la remédiation des insuffisances susmentionnées, constituent les éléments de motivation des travaux décrits dans ce papier. Nous y proposons une approche de constitution de sous-fils de conversation d'emails qui s'appuie sur les métadonnées, principalement la relation **reply-to** entre deux emails, les actes de dialogue de segments de texte extraits d'emails, la similarité sémantique entre ces segments et la production de paires transverses de ces segments de texte. La figure 1 met en avant une conversation du corpus BC3 avec ses emails, ainsi que les paires transverses de segments de texte construites via annotation. Un exemple de relation transverse est la relation entre le segment de texte *"Those who so wish could attend both weeks, and other people could attend only one week"* de l'email 1 qui est une **suggestion** et le **désaccord** *"Jacob, No*

1. message threading

way. Taking one week out of our calender 3 times...". Notre approche de constitution de sous-fils de discussion permet non seulement de démêler de façon fine une conversation d'emails mais aussi surtout de connaître l'état d'évolution de ladite conversation.

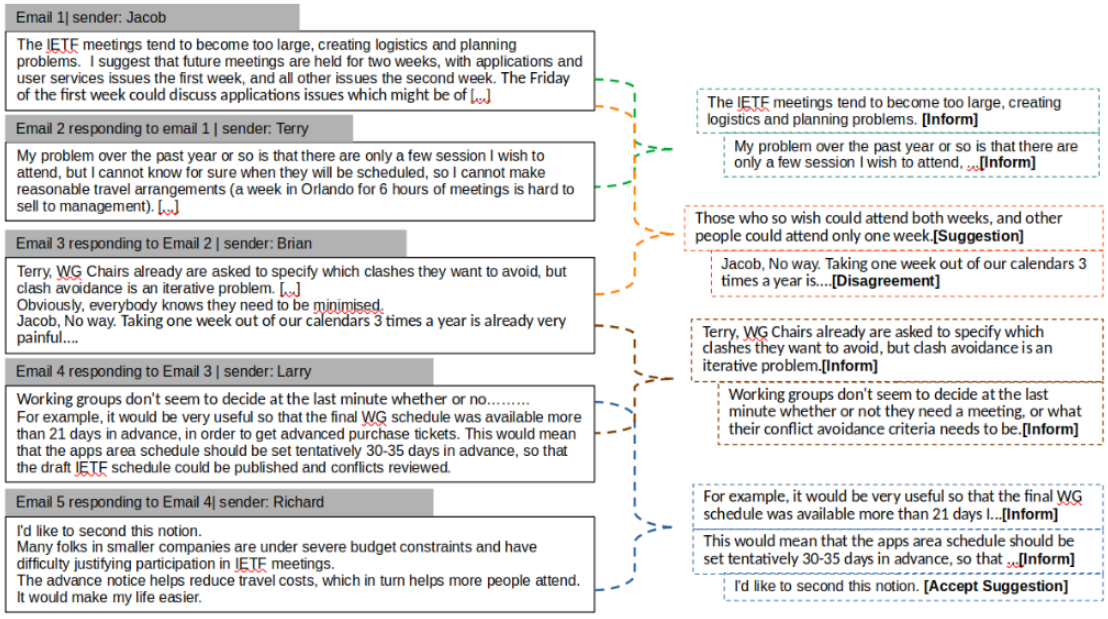

FIGURE 1 – Extrait d'une conversation du corpus BC3 avec des paires de segments de texte appariées

La figure 1 illustre la problématique de constitution de sous-fils de conversations d'emails. Les principales contributions dans ce papier pour la constitution des sous-fils de conversations d'emails sont :

— la production d'un référentiel d'actes de dialogue qui s'appuie sur la norme ISO 24617-2 [2]
— l'utilisation de ce référentiel pour annoter en actes de dialogues les segments de texte du corpus BC3
— La mise en œuvre de deux modèles dont le premier pour la tâche de reconnaissance d'actes de dialogue sur des segments de texte d'emails la seconde pour l'appariement de segments de texte de façon transverse. Le second modèle s'appuie sur les prédictions du premier

2 Travaux connexes

Les emails sont un outil de communication et de collaboration largement utilisé en entreprise et ils contiennent en plus des informations sur l'état d'avancement dans le cadre des projets, des données très riches et difficilement accessibles parce que entremêlées dans des fils de conversations d'emails qui sont peu ou pas structurés. La reconstruction de fils de conversations d'emails et l'identification des thématiques abordés dans ces conversations sont les principales problématiques liées à cette modalité de communication. Pour répondre à ces problématiques, il existe plusieurs approches.

— **Approches basées sur les méta-données pour la construction de fils de conversation**

2. Language resource management -Semantic annotation framework (SemAF)-Part 2: Dialogue acts

L'algorithme de Zawinski est l'un des algorithmes les plus populaires pour la construction de fils de conversation d'emails. Il est basé sur des informations provenant de métadonnées. Cependant, étant donné que ces champs d'en-tête sont facultatifs pour les clients de messagerie, ils ne sont pas toujours disponibles. De plus, ces données ne permettent pas de reconstruire toutes les conversations d'emails avec précision.

— **Approches s'appuyant sur les contenus d'emails afin de regrouper les emails en conversations sans tenir compte de la structure de celles-ci**
(Wu & Oard, 2005) utilisent les objets d'emails pour détecter les fils de conversation. Plus précisément, ils regroupent les emails dans le même fil de conversation si les emails ont le même objet et ont au moins un participant commun. (Wang *et al.*, 2008) extraient les fils de conversation d'emails à l'aide de l'algorithme de Zawinski, puis fusionnent ou décomposent les fils extraits en fonction de leurs objets afin de reconstruire les conversations. De même (Erera & Carmel, 2008) regroupent les emails en conversations cohérentes exploitant une fonction de similarité qui prend en compte tous les attributs d'email pertinents, tels que l'objet de l'email, les participants, la date de soumission et le contenu de l'email.

— **Reconstruction des fils de conversations en structures arborescente**
(Lewis & Knowles, 1997) traitent la reconstruction de fils de conversations d'emails comme un problème d'extraction d'informations. Ils ont étudié cinq approches d'extraction pour déterminer si un email est une réponse à un autre, et leurs résultats indiquent que l'utilisation du texte cité dans un email comme requête et du contenu non cité d'autres e-mails comme documents est la stratégie la plus efficace. Dans leurs recherches, Lewis et Knowles se sont concentrés uniquement sur le corps du texte de l'email et n'ont utilisé aucune autre information disponible dans les emails.
Pour constituer un corpus d'emails pseudo-anonymisés sous des structures arborescentes, (Tadonfouet Tadjou *et al.*, 2021) utilisent les méta-données, les contenus d'emails ainsi que les messages cités dans ceux-ci.
L'approche de (Yeh, 2006) suppose que tous les emails d'une conversation ont le même objet et que la durée de la conversation est généralement plus courte qu'une période fixe. Par conséquent, ils divisent les e-mails en plusieurs groupes, où tous les messages du même groupe ont des objets identiques et la différence de temps maximale entre deux e-mails du groupe est inférieure à un seuil fixe. Ils tentent ensuite de reconstruire l'arborescence des fils de discussion des emails en identifiant les relations parent-enfant entre les emails au sein du même groupe. Bien que leur méthode soit efficace pour détecter les structures arborescentes, les hypothèses qu'ils ont formulées ne sont pas toujours valables, comme en témoignent leurs expériences.
(Joshi *et al.*, 2011) ont utilisé la segmentation et la détection d'emails quasi identiques pour trouver et organiser des messages qui devraient être regroupés en fonction de leurs relations de réponse et de transfert. Ils supposent qu'un email répondant à un autre email contient en tant que texte cité dans un segment séparé, le texte de l'email auquel il répond. Ainsi, ils reconstruisent les fils de conversation tout en tenant compte de ces modèles de segmentation.

(Dehghani *et al.*, 2012, 2013) en s'appuyant sur le corpus BC3 proposent dans le premier papier une approche qui considère la recherche de fils de conversation d'email comme un problème d'optimisation, et exploite la programmation génétique pour rechercher intelligemment dans l'espace des solutions possibles. Dans le second, ils explorent deux nouvelles approches d'apprentissage, LExLinC et LExTreC, qui essayent d'extraire les structures linéaires et arborescentes des conversations, respectivement. LExLinC apprend à extraire les relations

entre fils de conversations d'emails et partitionne l'ensemble des données en clusters d'emails de sorte que chaque cluster représente un thread de conversation. D'autre part, LExTreC essaie d'apprendre des relations parents-enfants parmi les emails et extrait la structure arborescente des conversations.

— **Identification de thématiques dans les conversations d'emails**
(Joty *et al.*, 2010) ont annoté et rendu disponible le corpus BC3 annoté manuellement avec des sujets. Ils évaluent la fiabilité des annotateurs, montrent comment les modèles de segmentation de sujets existants (LCSeg et LDA) peuvent être appliqués aux emails et proposent deux nouvelles extensions de ces modèles qui utilisent non seulement des informations lexicales mais exploitent également une structure de conversation à un niveau plus fin de manière cohérente. Ils capturent la structure de conversation des emails au niveau du fragment (citation) sous la forme d'un graphe de fragment de citations (FQG – *Fragment Quotation Graph*). Un FQG capture la relation de réponse, l'utilisation des citations et d'autres fonctionnalités de conversation. LCSeg proposé pour la première fois par (Galley *et al.*, 2003) est un modèle de segmentation basé sur des chaînes lexicales qui suppose que les changements de sujet sont susceptibles de se produire là où des répétitions fortes de termes commencent et se terminent. Il commence par calculer des chaînes lexicales pour chaque mot qui ne fait pas parti des stop-words basé sur les répétitions de mots. Il classe ensuite les chaînes selon deux mesures : le nombre de mots dans la chaîne et la compacité de la chaîne. Il calcule ensuite la similarité cosinus (ou fonction de cohésion lexicale) à la transition entre les deux fenêtres d'analyse. Une faible similarité indique une faible cohésion lexicale et un changement net signale une forte probabilité d'une frontière de sujet réelle.

3 Méthodologie et Formalisation

3.1 Hypothèse et méthode

Une conversation est constituée d'au moins deux emails, avec au moins deux interlocuteurs, chacun de ses emails aborde au moins un sujet et contient au moins un segment de texte qui est une phrase courte ou une combinaison de plusieurs phrases. Pour mettre en relation certaines phrases d'une conversation d'un email B avec ceux d'une email A, on peut tout simplement s'appuyer sur les métadonnées de la conversation comme la relation *reply-to* entre deux emails, mais aussi sur leur similarité sémantique. Cependant certaines phrases d'emails sont souvent très courtes (moins de 4 mots) et ainsi dépourvues de contexte pour un meilleur score de similarité sémantique avec les phrases d'un email précédent. Une courte phrase pourrait par exemple être *"Ça me va"* : un **accord** ou une *appréciation* qui répond à une *suggestion* dans un précédent email. En général, dans une conversation, certains contenus ou segments de texte à dans un email de continuation sont des réponses, élaborations, suggestions ou d'autres types d'actes de dialogue qui sont en relation avec des segments de texte d'emails précédents. Ces relations entre deux segments de de texte d'emails sont dites **transverses**. On peut aussi dire que deux segments de texte sont adjacents du fait de l'existence d'une relation entre eux. Notre objectif dans ce papier est d'identifier des paires de segments de texte qui sont reliés de manière transverse au sein d'une même conversation. Notre hypothèse est de s'appuyer non seulement sur les métadonnées et la similarité sémantique entre segments de texte mais de les compléter avec les actes de dialogues de ces segments afin d'avoir un système robuste d'appariement de segments de texte entre emails.

Après consolidation des différentes paires entre elles, on obtient des groupes de segments de texte qui

représentent la ou les parties essentielles de la conversation, ce qui vise à en faciliter la lecture et la compréhension. L'extraction de paires transverses d'une conversation se déroule en trois étapes :

— La segmentation des contenus d'emails, par défaut en phrases dans ce papier
— L'assignation d'actes de dialogue aux segments (**Dialogue Acts Recognition - DAR**) : différents modèles sont entraînés basés sur des réseaux de neurones avec sur la dernière couche une fonction SoftMax pour la classification des différents actes de dialogue décrits dans la section 5.1. Le segment de texte précédent celui à prédire est utilisé en contexte afin d'améliorer les performances des modèles.
— L'appariement des segments de texte ou d'énoncés(**AE**) ou des emails d'une conversation : un classifieur binaire est entraîné pour cette tâche en s'appuyant sur le framework SetFit (**?**). Comme entrée à ces modèles, des paires de segments positives et négatives sont construits à partir des corpus existants BC3 et Reddit. Les paires positives et négatives appartiennent à une même conversation. Chaque paire est constituée de deux segments de texte extraits respectivement de deux emails E_i et E_j, i et j des entiers tels que $i < j$. Les deux segments (S_a, S_b) d'une paire positive ont chacun leurs actes de dialogue respectifs (DaS_a, DaS_b) et sont en relation du fait que DaS_b est répond au sens large du terme à DaS_a. Par exemple, DaS_a peut être une suggestion et DaS_b une appréciation. Une paire négative quant à elle a ses segments de texte qui ne sont pas en relation et ont été choisis de façon aléatoire dans une conversation.

3.2 Formalisation du problème

Étant donné une conversation C comportant n emails $\{E_1, E_2, ..., E_n\}$, chacun de ces emails E_i contient m segments de texte $\{s_1^i, s_2^i, ..., s_m^i\}$. La construction des paires de segments de texte

$P = \{(s_a^i, s_b^j), (s_d^k, s_c^l), \cdots (s_m^{n-i}, s_m^n)\}$ transverses sur les emails de la conversation C telle que

$(i < j < k < l \cdots < n)$ et $(a, b, c, d \cdots, m) \in \mathbb{N}$. Les deux éléments d'une paire :

— appartiennent à deux emails distincts E_i et E_j de C avec $i < j$
— ont une relation basée sur les actes de dialogue de type question-réponse, suggestion appréciation, etc.
— peuvent avoir une similarité sémantique

La figure 2 illustre les deux étapes de notre pipeline dont l'objectif est de faire ressortir les paires de segments positives (annotées "yes" dans le schéma)

4 Corpus et annotations

Nous avons utilisé le corpus d'emails BC3 (Ulrich *et al.*, 2008) et le corpus "*Coarse Discourse Sequence Corpus*" (CDSC) de (Zhang *et al.*, 2017) extrait du forum Reddit. CDSC a été annoté en acte de dialogues par trois annotateurs, qui ont aussi mis en relation les différents posts (post B répond au post A) d'une conversation, ce qui a guidé notre choix pour compléter les données du corpus BC3 qui est constitué de seulement 40 conversations pour 261 emails et 1127 phrases. Le corpus BC3 a été construit à la base pour une tâche de résumé de conversations d'emails mais (Jeong *et al.*, 2009) l'ont utilisé dans leurs travaux de classification en actes de dialogues des phrases d'emails et de forums avec des approches semi-supervisées. Ils ont fait réannoter les phrases de BC3 avec douze

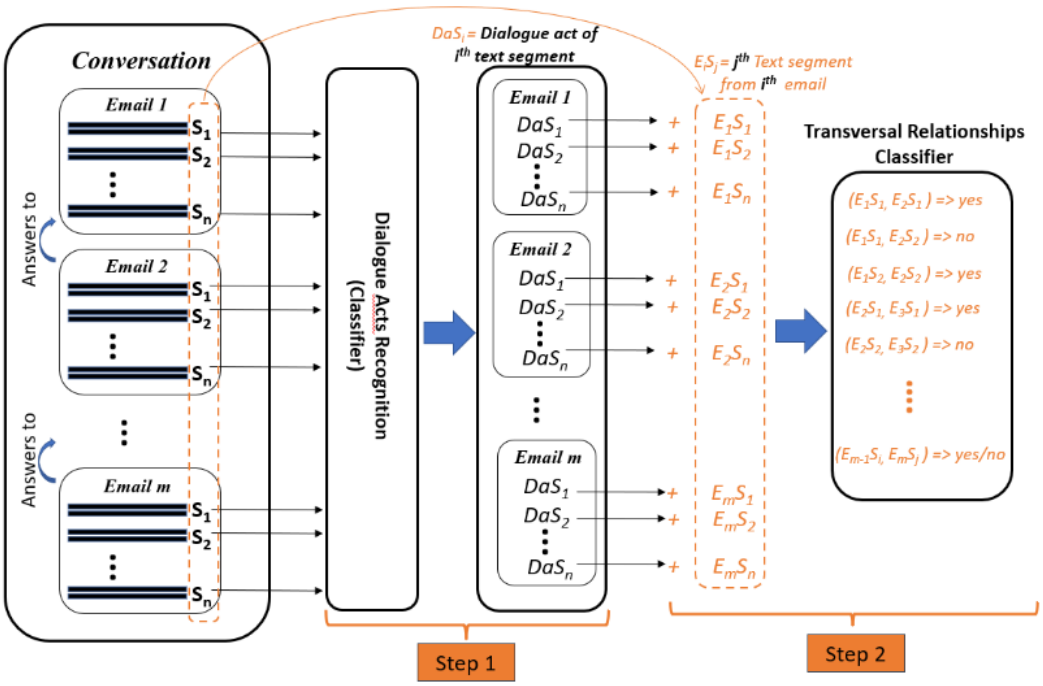

FIGURE 2 – Processus en deux étapes pour l'appariement de segments de texte

actes de dialogues par deux annotateurs avec un accord inter-annotateur égal à 0,79. Cependant nous avons constaté que ces actes de dialogue ne répondaient pas à notre besoin car imprécis : plusieurs phrases d'emails annotées comme *"statement"* étaient pourtant principalement des suggestions ou des élaborations. Pour cette raison, nous avons décidé de réannoter le corpus BC3 en actes de dialogues en relation entre phrases d'emails dans une conversation. Les actes de dialogues utilisés proviennent d'un référentiel que nous avons spécialement établi pour notre besoin sur les conversations d'emails et aussi pour des raisons de clarté.

4.1 Mise en œuvre d'un référentiel d'annotations

La classification de segments de texte en actes de dialogue est une problématique abordée avec d'autres corpus, par exemple avec MRDA [3] un corpus de transcriptions de réunion (Joty & Mohiuddin, 2018; Mohiuddin *et al.*, 2019). Dans la section 5.1 nous décrivons comment nous avons utilisé MRDA pour entraîner des modèles de classification d'énoncés de réunion en actes de dialogue. Afin de pouvoir transférer les modèles de reconnaissance d'actes de dialogues entraînés sur un corpus d'emails, nous avons opté pour une adaptation des actes de dialogue définis dans MRDA parce que ceux-ci sont très diversifiés et certains d'entre-eux peuvent difficilement se retrouver dans les contenus d'emails. L'adaptation que nous avons effectuée s'est déroulé en trois étapes :

— premièrement nous avons identifié sur les différentes fonctions communicatives, les actes de gestion de tâches et les feedbacks définis dans la norme ISO 24617-2 qui répondaient le mieux à l'identification des actes de dialogues dans les emails et aussi aideraient à une structuration dialogique de conversation d'emails

3. Meeting Recorder Dialogue Act Corpus

— ensuite nous avons défini des acronymes que nous avons définis qui permettent de facilement identifier chaque acte de dialogue sans ambiguïté
— enfin nous avons fait correspondre les acronymes d'actes de dialogue du MRDA avec les nôtres en consultant de façon collégiale (deux personnes) une centaine d'énoncés de MRDA de différents types d'actes pour s'assurer de bien tous les prendre en compte.

La table 1 récapitule le référentiel que nous avons défini. Dans la colonne la plus à gauche, les actes de dialogue de MRDA mappés. En orange les fonctions communicatives de la norme ISO 24617-2 et enfin en bleu les acronymes et actes de dialogues que nous avons définis. Les acronymes s'interprètent facilement avec leurs actes de dialogue respectifs, ce qui n'est pas le cas pour ceux de MRDA et de Switchboard Dialogue [4].

		Communicative functions from ISO-24617-2				
MRDA Dialogue Acts (Full Labels)	Ours Labels	0	1	2	3	4
br	Q	Info-seeking functions	Question			
|Q|fh|fh; |Q|q?|e; qy; qo	PrQ			Propositional Question		
bu, d, g	CkQ				Check Question	
qw	SQ			Set Question		
	TQ				Test Question	
qr; qrr|qrr	ChQ			Choice Question		
S; |S|fh|fh	I	Info-providing functions	Inform			
no	An			Answer		
na	Cf				Confirm	
nd; ng	Dcf				Disconfirm	
aa	Ag			Agreement		
ar	Dag			Disagreement		
bc	Cr				Correction	
	AdR	Commissives	Offer	Address Request		
	AcR				Accept Request	
	DeR				Decline Request	
cs	O					
	Pr			Promise		
	AdS			Address Suggestion		
	AcS				Accept Suggestion	
	DeS				Decline Suggestion	
	S	Directives	Suggestion			
co; qy|cs	R		Request			
	Is			Instruct		
	AdO				Address Offer	
	AcO					Accept Offer
	DeO					Decline Offer
fw;by;ft;fa	P		Politeness			
ba	AA		Appreciation/Assessment			
bd	M		Miscellaneous			
df, s|f;bs; arp;bsc; aap; |t	Ex		Elaboration			
am	Hy		Hypothesis, Asumption			

TABLE 1 – Liste d'acronymes définis s'appuyant sur la norme ISO 24617-2 et leur correspondance d'actes de dialogues dans MRDA

4.2 Annotation du corpus BC3

Il existe quelques travaux d'identification d'actes de dialogues dans les conversations d'emails comme ceux de (Taniguchi *et al.*, 2020) qui ont annoté en actes de dialogue plus de 2k fils de conversations du corpus Enron avec deux granularités différentes : 35k phrases avec une granularité fine et 6k emails annotés avec une granularité moins fine. Cependant ce corpus d'emails annotés n'est pas disponible et de par nos travaux, excepté les travaux de (Jeong *et al.*, 2009) pour la tâche de classification des phrases d'emails en actes de dialogues ; il n'existe pas de corpus d'emails finement annoté en actes de dialogue et en relation de messages, ce que nous avons effectué avec le corpus BC3.

Deux personnes ont ainsi annoté ce corpus BC3 en s'appuyant sur le référentiel mis en œuvre. Nous avons annoté 20 conversations, soit 662 segments de texte. La valeur de Kappa (Viera & Garrett, 2005) est de 0.47 pour les annotations en actes de dialogue, cette valeur s'interprète comme un accord modéré entre les deux annotateurs.

Concernant les appariements de segments de texte sur ces 20 conversations, les deux annotateurs ont

4. Switchboard Dialogue Act Corpus

respectivement identifié 289 et 237 relations entre les segments dans lesdites conversations avec une intersection de 107 relations. Ces disparités mettent en exergue la difficulté de la tâche d'appariement de segments de texte transverses même pour des humains. Cependant pour entraîner nos modèles d'appariement de segments de texte, nous avons utilisé l'union des paires positives annotées issues des deux annotateurs, soit 418 au lieu de l'intersection qui est de très petite taille. Nous avons constitué 196 paires négatives avec chaque paire constituée de segments de texte de la même conversation d'emails. Dans la suite de ce papier, BC3 est utilisé pour référencer le total de 614 paires.

4.3 Corpus Reddit

Nous utilisons une version du corpus Reddit [5] (Zhang *et al.*, 2017) finement annoté par trois personnes en actes de dialogue et en relation REPLY-TO entre les messages de chaque conversation. Ces annotations portent sur environ 10k fils de conversation de Reddit. Ce corpus est l'un des rares corpus de conversations asynchrone (forum) annotés sur ces deux aspects.

La table 2 fournit les tailles des différentes paires constituées ainsi que leur distribution pour l'entraînement et les tests de notre modèle. Les actes de dialogues de BC3 et Reddit sont cependant différents dans nos expériences et nous avons donc établi une correspondance des actes de dialogue de BC3 vers ceux de Reddit, obtenant un corpus que nous avons nommé $BC3_{map}$.

DataSet	Train	Validation	Test	Total
BC3	229 (156 PP + 73 PN)	105 (71 PP + 34 PN)	280 (191 PP + 89 PN)	418 PP + 196 PN
Coarse Reddit	7536 (2998 PP + 4538 PN)	932 (374 PP + 558 PN)	942 (375 PP + 567 PN)	3747 PP + 5663 PN
Total	7648 (3110 PP + 4538 PN)	984 (400 PP + 584 PN)	1080 (444 PP + 636 PN)	

TABLE 2 – Distribution des données utilisés (PP : Paires positives, PN : Paires négatives)

5 Expériences, résultats et analyses

5.1 Reconnaissance d'actes de dialogue

Nous utilisons le corpus MRDA afin d'entraîner un modèle pour la classification ou reconnaissance d'actes de dialogue sur des énoncés de conversations. MRDA est à la base un corpus d'échanges 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 énoncés sont essentiellement constitués de ces marqueurs. 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. Nous avons aussi supprimé ces marqueurs dans les énoncés.

Après avoir filtré le corpus, nous obtenons respectivement 36722, 7985, 7918 énoncés pour les données d'entraînement, de validation et de test. Nous avons fine-tuné le modèle BERT pour la tâche de classification d'énoncés en actes de dialogue en lui rajoutant une couche BiLSTM à chaque fois suivie d'une d'auto-attention ou pas *(BERT_BiLSTM, BERT_BiLSTM_Att)*. Comme entrées à ces modèles, nous utilisons dans un premier temps les énoncés à classifier pris indépendamment les uns des autres et, dans un second temps, nous considérons un contexte qui est l'énoncé précédant

5. Coarse Discourse

immédiatement celui à classifier donnant ainsi lieu au modèle *BERT_BiLSTM_Ctx* par exemple sans la couche d'attention.

Nous avons entraîné certains de ces modèles avec comme entrées des énoncés regroupés, c'est-à-dire que deux ou plusieurs énoncés qui se suivent dans les transcriptions qui sont du même interlocuteur et qui ont le même acte de dialogue sont regroupés en seul énoncé. Ce regroupement donne lieu à des variantes de nos modèles dont les noms se terminent par *GrpFalse* et *GrpTrue* respectivement pour les modèles avec les entrées non groupées et groupées. Nous utilisons les 20 actes de dialogue suivants *['aa', 'adr', 'ads', 'ag', 'an', 'cf', 'cr', 'dag', 'dcf', 'ex', 'hy', 'i', 'is', 'o', 'p', 'pr', 'q', 'r', 's', 'tc']* extraits de la colonne "Our Labels" de la table 1.

5.2 Appariement de segments de texte ou d'Énoncés (AE)

Pour la tâche d'appariement de segments de texte transverses sur les emails d'une conversation, nous utilisons le framework SetFit (SetFit - Efficient Few-shot Learning with Sentence Transformers) de(Tunstall *et al.*, 2022). Nous avons constitué des paires positives et négatives de segments de texte, extraites des corpus BC3 et Reddit.

Les paires positives font partie de la même conversation et le second membre de la paire est en relation avec le premier comme défini dans la section 3.1. Les paires négatives quant à elles sont aussi extraites d'une même conversation, cependant il n'existe aucune relation entre les segments de texte de ces paires. La similarité sémantique est prise en compte avec SetFit qui implémente aussi en son sein l'approche contrastive qui améliore davantage les scores de similarités entre contenus.

Une entrée à notre modèle est donc une paire de segments de texte $([Da]S_i, [Da]S_j)$ avec "[Da]" qui sont des tokens spéciaux créés à partir des actes de dialogues et placés devant chaque segment de texte de la paire. Comme exemple de tokens spéciaux, nous avons [EPNT], [QSTI], AGMT] respectivement pour *"explanation", "question", et "agreement"*

Nous avons entraîné les différents modèles ci-dessous avec pour chacun le (ou les) corpus sur lequel (lesquels) il a été entraîné :

- **AE** : Appariement d'Énoncés sans les actes de dialogues, entraîné afin de montrer quel est l'impact des actes de dialogues sur l'appariement de segment de texte. Ce modèle s'appuie uniquement sur la similarité sémantique
- **AE+ADG** : Appariement d'Énoncés avec les Actes de Dialogues Gold, entraîné de façon indépendante avec Reddit, Reddit+BC3 et Reddit+BC3_map
- **AE+ADP** : Appariement d'Énoncés avec les Actes de Dialogues Prédits en utilisant notre modèle de classification de segments de texte en actes de dialogues. Ce modèle est entraîné de façon indépendante avec Reddit, Reddit+BC3. Les actes de dialogue utilisés reflètent la réalité dans laquelle nos modèles seront utilisés, car les annotations en acte de dialogues seront faites de façon automatique et non manuellement (GOLD) comme dans le modèle AE+ADG

Nous effectuons les tests de nos modèles de différentes manières. D'une part ceux entraînés uniquement sur Reddit sont testés sur les données de test de Reddit et sur l'ensemble du corpus BC3, ceci afin de voir s'il a y transfert de connaissance des données de forums sur les emails. D'autre part les modèles entraînés sur Reddit+ BC3/BC3_map sont uniquement testés sur des leurs données de test respectives extrait du découpage en données d'entraînement, de validation et de test. Ces tests sont effectués sur les données avec et sans actes de dialogue, ceci afin de pouvoir identifier l'apport réel de l'utilisation des actes de dialogues pour notre tâche d'appariement.

Ces modèles ont tous été entraînés avec les mêmes hyperparamètres *({learning_rate :4.3879e -06, num_epochs 5, batch_size :32, 'model_id' :'sentence-transformers/bertbase-nli-mean-tokens', 'num_iterations' : 80})* obtenus en amont en entraînant le modèle **AE+ADG** et ce avec une approche de recherche d'hyperparamètres optimum implémentée avec Optuna.

5.3 Résultats et analyses

5.3.1 Reconnaissance d'actes de dialogue

La figure 3 met en avant les performances des différents modèles que nous avons utilisés pour la classification des segments de texte du corpus MRDA en actes de dialogue. Il ressort de cette figure que le modèle BERT finetuné avec une couche BiLSTM plus une couche d'auto-attention(BERT_BiLSTM_Att_GrpFalse) sans le regroupement des inputs a une meilleure performance au bout de 2 epochs contrairement aux autres combinaisons.

Les modèles qui prennent comme entrée les énoncés avec chacun son contexte respectif donnent de moins bons résultats. Le contexte rajouté à chaque énoncé à classifier devrait logiquement améliorer les performances du modèle, ce qui n'est pas le cas ici. Ceci peut s'expliquer par le fait qu'il y a eu une perte d'information lors de notre processus de filtrage mais aussi du fait que le corpus MRDA est un corpus de transcription de réunion et donc est plutôt constitué de conversations synchrones. De plus lors d'une réunion plusieurs sujets peuvent être abordés de façon entremêlée et donc les contextes que nous rajoutons aux énoncés à classifier n'ont probablement aucune similarité avec ceux-ci au vu de la performance de BERT_BiLSTM_CtxGrpFalse.

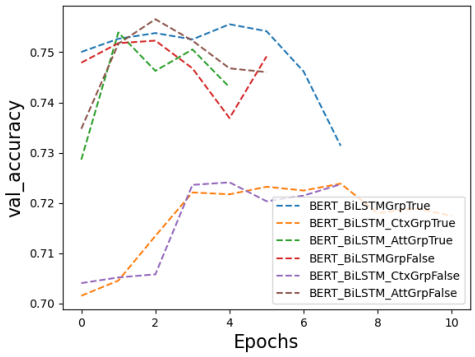

FIGURE 3 – Performances pour la classification en actes de dialogues

5.3.2 Appariement de segments de texte ou d'énoncés (AE)

La table 4 récapitule les scores F1 des différents modèles que nous avons entraînés. Nous avons utilisé la validation croisée k-fold avec k=3 pour les scores F1 lors de l'évaluation de nos modèles sur les données de test BC3, ceci parce que le corpus BC3 est de très petite taille (cf. table 2). Les scores des différents modèles entraînés sur le seul corpus BC3 sont biaisés du fait de la petite de BC3. En plus nous avons relevé de l'overfitting lors de l'entraînement ces modèles.

Les scores des modèles entraînés sur les données avec les actes de dialogues "GOLD" sont meilleurs que ceux entraînés avec les actes prédits. C'est le cas de AE+ADG entraîné sur Reddit+BC3 qui

		AE			AE + ADP			AE + ADG			
	Données d'entrainements	BC3	Reddit	Reddit+BC3	BC3	Reddit	Reddit + BC3	BC3	Reddit	Reddit + BC3	Reddit+BC3_map
F1-score	BC3	0.72	0.61	0.73	0.72	0.60	0.74	0.74	0.62	0.78	0.77
	Reddit	0.52	0.69	0.68	0.50	0.67	0.68	0.50	0.71	0.75	0.73

TABLE 3 – Résultats des modèles d'appariement de segments de texte

donnent respectivement des scores de 0.78 et 0.75 sur les données de test de BC3 et Reddit. Cependant, AE+ADP entraîné aussi Reddit+BC3 donne de moins bons résultats respectivement 0.74 et 0.69 pour ces mêmes données de test, soit une perte d'environ 4 points. Ceci démontre l'importance de l'utilisation des actes de dialogues les plus fiables pour l'appariement de segments de texte dans les conversations. Ainsi plus performant sera notre modèle de classification de segments de texte en actes de dialogue, meilleurs seront les résultats de nos modèles d'appariement.

Dans nos expériences, nous avons mappé les actes de dialogues de BC3 sur ceux de Reddit formant ainsi BC3_map et nous avons entraîné AE+ADG sur Reddit+BC_map qui donne un score de 0.77 sur les données de test de BC3_map contre 0.78 pour le même modèle mais entraîné sur BC3 avec les actes de dialogues "GOLD", cette différence d'un point peut s'expliquer par le fait de la diversité des actes de dialogues que nous avons mis en œuvre dans notre référentiel (section 4.1 soit 20 actes de dialogues) qui améliore les performances de notre modèle.

AE+ADP entraîné respectivement sur Reddit et Reddit+BC3 donne 0.67 et 0.68 lorsqu'évalué sur Reddit. Cette différence d'un point peut s'interpréter par l'ajout des données BC3 (environ 3% de la taille de Reddit) sur le Corpus Reddit. Et on peut en déduire qu'on pourrait gagner des points sur nos scores avec davantage de données.

Ce même modèle AE+ADP entraîné respectivement sur BC3 et Reddit+BC3 a des scores de 0.72 et 0.74 lorsqu'évalué sur BC3, soit une différence de 2 points. Ce gain est dû à l'augmentation de données (Reddit sur BC3), mais traduit aussi le transfert de connaissance des données de forum vers les données d'emails.

Nous avons analysé un échantillon (28%) extrait des données de test de BC3, la majorité des énoncés de cet échantillon sont prédits comme des questions ou des annonces (*inform*). Ci-dessous quelques vrais négatifs (VN) et faux positifs(FP) prédits par le modèle AE+ADP.

1. VN : *Inform : It is not done but you will get the idea. <–> Assessment- it's a good piece of work.*
2. VN : *Inform : If all web authors felt like this about groups they are not prepared to cater to, its no wonder we need WAI. <–> Question : Jonathan, do you really mean to be insulting to me ?*
3. VN : *Inform : Please take a look at [URL] for a first small attempt at this. <–> Inform : Got a could not connect to remote server from both links at [URL]*
4. FP : *Question : My quesiton is how would a screen reader handle that code.... <–> Inform : He just hadn't run into them in the standard version before trying the version for screen reader users.*
5. FP : *Politness : Thanks for the suggestion. <–> Inform : I would skip IE[PATH] since designers worth 2c can tell you already how things work there by reading the code.*
6. FP : *Question : Can you suggest another venue and possible sponsor ? <–> Inform : I want to go to Venice !*

D'une part ces paires d'énoncés font ressortir que nos modèles ont parfois besoin de contexte pour une

meilleure prédiction : un tel contexte pourrait améliorer la classification des paires 1, 2, 4 et 6. D'autre part l'inexactitude des actes de dialogues de certains énoncés contribue à la mauvaise classification des paires qu'elles constituent. Dans la paire 3, les deux énoncés sont en réalité respectivement une requête et une réponse à celle-ci. Le dernier exemple est un classique du type question/réponse. En plus de la petite taille de nos corpus, cette analyse montre les insuffisances de notre approche et identifie clairement les leviers sur lesquelles s'attaquer pour améliorer nos modèles.

Tous les scores de nos modèles avec actes de dialogue "GOLD" (AE+ADG) sont meilleurs que ceux utilisant les actes de dialogue prédits. Cependant l'utilisation concrète de nos modèles en entreprise se fera avec des actes de dialogue prédits et non "GOLD", vu le coût des processus d'annotations.

Comme baseline, nous avons utilisé les métadonnées (reply-to entre les emails d'une conversation) avec d'une part BM25 (Robertson & Zaragoza, 2009), un algorithme de ranking, souvent utilisé comme baseline ou couplé à d'autres méthodes pour la sélection de réponses à un énoncé dans les dialogues (Yan *et al.*, 2018; Chen *et al.*, 2021; Lin *et al.*, 2020; Henderson *et al.*, 2019) et d'autre part avec un système de similarité sémantique basé sur des modèles neuronaux. Pour ce second système, nous utilisons SentenceTransformers (Reimers & Gurevych, 2019).

	Baseline (avec métadonnées)		Modèle entraîné
	BM25	Sentence-BERT :	AE + ADP (entraîné avec Reddit+BC3)
données de test de BC3	0.58	0.65	**0.74**

TABLE 4 – Comparaison de notre modèle avec nos baselines

6 Conclusion

Dans ce papier, nous avons proposé un pipeline qui s'appuie sur la reconnaissance en actes de dialogue de segments de texte et la mise en relation de ceux-ci pour la reconstruction de fils de discussions dans les conversations d'emails. Les résultats de nos modèles d'appariement de segments de texte de conversation d'emails montrent l'intérêt de l'utilisation des actes de dialogues pour ce problème. L'analyse de ces résultats nous a permis d'identifier les insuffisances de notre approche comme l'absence de contexte dans nos énoncées et l'inexactitude des actes de dialogues prédits dans la première étape de notre pipeline. Pour les futurs travaux nous allons combiner les composantes de notre pipeline dans une architecture bout-en-bout qui prendra en entrée une conversation d'emails et produira en sortie de paires de segments de ladite conversation. Cette architecture enrichira les énoncés avec les contextes des leurs conversations respectives.

Suite à une expérimentation préliminaire favorable effectuée lors de nos travaux qui a consisté à segmenter du texte, labelliser les segments obtenus en actes de dialogues et les appariements tout ceci via le prompting avec les larges modèles de langages existants tels GPT-3 (Brown *et al.*, 2020), BLOOM (Workshop *et al.*, 2022), LLAMA (Touvron *et al.*, 2023), nous allons dans nos prochains travaux constituer à partir d'Enron (Klimt & Yang, 2004), un corpus de taille conséquente finement annoté afin d'améliorer les résultats de nos modèles.

Références

BROWN T. B., MANN B., RYDER N. *et al.* (2020). Language Models are Few-Shot Learners. DOI : 10.48550/ARXIV.2005.14165.

CHEN W., GONG Y., XU C., HU H., YAO B., WEI Z., FAN Z., HU X., ZHOU B., CHENG B. *et al.* (2021). Contextual fine-to-coarse distillation for coarse-grained response selection in open-domain conversations. *arXiv preprint arXiv :2109.13087*.

DEHGHANI M., ASADPOUR M. & SHAKERY A. (2012). An Evolutionary-Based Method for Reconstructing Conversation Threads in Email Corpora. ASONAM '12, p. 1132–1137, USA : IEEE Computer Society. DOI : 10.1109/ASONAM.2012.195.

DEHGHANI M., SHAKERY A., ASADPOUR M. & KOUSHKESTANI A. (2013). A learning approach for email conversation thread reconstruction. *Journal of Information Science*, **39**, 846 – 863.

ERERA S. & CARMEL D. (2008). Conversation Detection in Email Systems. In C. MACDONALD, I. OUNIS, V. PLACHOURAS, I. RUTHVEN & R. W. WHITE, Éds., *Advances in Information Retrieval*, p. 498–505, Berlin, Heidelberg : Springer Berlin Heidelberg.

GALLEY M., MCKEOWN K. R., FOSLER-LUSSIER E. & JING H. (2003). Discourse Segmentation of Multi-Party Conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, p. 562–569, Sapporo, Japan : Association for Computational Linguistics. DOI : 10.3115/1075096.1075167.

HENDERSON M., VULIĆ I., GERZ D., CASANUEVA I., BUDZIANOWSKI P., COOPE S., SPITHOURAKIS G., WEN T.-H., MRKŠIĆ N. & SU P.-H. (2019). Training neural response selection for task-oriented dialogue systems. *arXiv preprint arXiv :1906.01543*.

JEONG M., LIN C.-Y. & LEE G. G. (2009). Semi-supervised Speech Act Recognition in Emails and Forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, p. 1250–1259, Singapore : Association for Computational Linguistics.

JOSHI S., CONTRACTOR D., NG K., DESHPANDE P. & HAMPP T. (2011). Auto-grouping emails for faster e-discovery. *Proceedings of the VLDB Endowment*, **4**, 1284 – 1294.

JOTY S., CARENINI G., MURRAY G. & NG R. T. (2010). Exploiting Conversation Structure in Unsupervised Topic Segmentation for Emails. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p. 388–398, Cambridge, MA : Association for Computational Linguistics.

JOTY S. & MOHIUDDIN T. (2018). Modeling Speech Acts in Asynchronous Conversations : A Neural-CRF Approach. *Computational Linguistics*, **44**(4), 859–894. DOI : 10.1162/coli_a_00339.

KLIMT B. & YANG Y. (2004). The Enron Corpus : A New Dataset for Email Classification Research. In J.-F. BOULICAUT, F. ESPOSITO, F. GIANNOTTI & D. PEDRESCHI, Éds., *Machine Learning : ECML 2004*, p. 217–226, Berlin, Heidelberg : Springer Berlin Heidelberg.

LEWIS D. D. & KNOWLES K. A. (1997). Threading electronic mail : A preliminary study. *Information Processing Management*, **33**(2), 209–217. Methods and Tools for the Automatic Construction of Hypertext, DOI : https ://doi.org/10.1016/S0306-4573(96)00063-5.

LIN Z., CAI D., WANG Y., LIU X., ZHENG H.-T. & SHI S. (2020). The world is not binary : Learning to rank with grayscale data for dialogue response selection. *arXiv preprint arXiv :2004.02421*.

MOHIUDDIN T., NGUYEN T.-T. & JOTY S. (2019). Adaptation of Hierarchical Structured Models for Speech Act Recognition in Asynchronous Conversation. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 1326–1336, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1134.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. *CoRR*, **abs/1908.10084**.

ROBERTSON S. & ZARAGOZA H. (2009). The Probabilistic Relevance Framework : BM25 and Beyond. *Found. Trends Inf. Retr.*, **3**(4), 333–389. DOI : 10.1561/1500000019.

TADONFOUET TADJOU L., BOURGE F., MARIE T., ROMARY L. & DE LA CLERGERIE É. (2021). Building A Corporate Corpus For Threads Constitution. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, p. 193–202, Online : INCOMA Ltd.

TANIGUCHI M., UEDA Y., TANIGUCHI T. & OHKUMA T. (2020). 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*, p. 4969–4980, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.436.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). LLaMA : Open and Efficient Foundation Language Models. DOI : 10.48550/ARXIV.2302.13971.

TUNSTALL L., REIMERS N., JO U. E. S., BATES L., KORAT D., WASSERBLAT M. & PEREG O. (2022). Efficient Few-Shot Learning Without Prompts. DOI : 10.48550/ARXIV.2209.11055.

ULRICH J., MURRAY G. & CARENINI G. (2008). A Publicly Available Annotated Corpus for Supervised Email Summarization.

VIERA A. J. & GARRETT J. M. (2005). Understanding interobserver agreement : the kappa statistic. *Family medicine*, **37**(5), 360—363.

WANG X., XU M., ZHENG N. & CHEN M. (2008). Email Conversations Reconstruction Based on Messages Threading for Multi-person. *2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing*, **1**, 676–680.

WORKSHOP B., :, SCAO T. L., FAN A., AKIKI C. *et al.* (2022). BLOOM : A 176B-Parameter Open-Access Multilingual Language Model. DOI : 10.48550/ARXIV.2211.05100.

WU Y. & OARD D. W. (2005). Indexing emails and email threads for retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

YAN Z., DUAN N., BAO J., CHEN P., ZHOU M. & LI Z. (2018). Response selection from unstructured documents for human-computer conversation systems. *Knowledge-Based Systems*, **142**, 149–159.

YEH J.-Y. (2006). Email Thread Reassembly Using Similarity Matching. In *International Conference on Email and Anti-Spam*.

ZHANG A., CULBERTSON B. & PARITOSH P. (2017). Characterizing Online Discussion Using Coarse Discourse Sequences.