Constitution de fils de discussion cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration

Lionel Tadonfouet Tadjou

Université Sorbonne Université

Ecole Doctorale Informatique, Télécommunications et Electronique - ED130

Inria de Paris / Équipe ALMAnaCH
Orange Innovation Caen

Thèse de doctorat
Discipline : Informatique

Présentée par

Lionel Tadonfouet Tadjou

Dirigée par

Laurent Romary

Co-encadrée par

Eric De La Clergerie, **Fabrice Bourge** et **Tiphaine Marie**

Pour obtenir le grade universitaire de
Docteur de l'Université Sorbonne Université

Constitution de fils de discussion cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration

---

Présentée et soutenue publiquement le 19 octobre 2023 devant le jury composé de :

Sylvain Lamprier	Université d'Anger	Examinateur
Chloé Clavel	Telecom Paris	Rapporteur
Iris Eshkol-Taravella	Université Paris-Nanterre	Examinateur
Laurent Romary	Inria Paris	Directeur
Éric De La Clergerie	Inria Paris	Co-Encadrant
Frédéric Bechet	Université Aix-Marseille	Rapporteur

À CELLES ET CEUX
QUI M'ONT PRÉCÉDÉ ET QUI
ONT CONTRIBUÉ À MA PRÉSENCE ICI!!!

Remerciements

Je tiens à remercier les membres de mon jury de thèse - Sylvain Lamprier, Cholé Clavel, Frédéric Bechet, et Iris Eshkol-Taravella qui ont généreusement offert de leur temps pour réviser ce manuscrit et évaluer mon travail dans la soutenance à venir.

Je remercie sincèrement Laurant Romary, Éric de la Clergerie mes directeurs de thèse. Je leur suis très reconnaissant pour le temps qu'ils ont consacré à me guider, pour leur bienveillance tout au long de ma thèse et leur accompagnement dans mes recherches. Leur approche pédagogique, leurs encouragements et leur profond intérêt pour ma réussite ont constitué un soutien exceptionnel qui m'a aidé à surmonter toutes les épreuves pendant ces 38 mois de travail. Je tiens particulièrement à remercier Éric de la Clergerie pour son accompagnement, sa bonne humeur et sa patience, qui ont été une source de motivation tout au long de cette thèse. Je remercie aussi Frédéric Landragin, qui avec Sylvain Lamprier ont été les membres de jury de mes comités de suivi de thèse.

Merci ensuite à Fabrice Bourge et Tiphaine Marie mes co-encadrants chez Orange Innovation. Ils m'ont appris à surmonter des difficultés et à réaménager mon énergie sur des problèmes bien précis. Fabrice s'est toujours consacré avec dévouement à mon encadrement sur les aspects méthodologiques et pédagogiques de ma thèse. Je vous remercie sincèrement pour votre soutien dans tous mes projets, y compris ceux qui allaient au-delà du cadre de ma recherche. Votre tolérance envers mes idées les plus audacieuses de sur-ingénierie a été d'une aide précieuse. Vos encouragements récurrents ont été des facteurs essentiels dans ma croissance professionnelle et personnelle.

Un merci spécial à mon encadrant de Master, Maurice Tchoupé pour la flamme de la recherche qu'il a attisée en moi et pour l'esprit méthodologique qu'il a su m'inculquer.

Merci ensuite à tous mes collègues, notamment ceux d'Orange Innovation Caen : Maryline, Cédric, Olivier, Christian, Christophe, Alpha-Oumar, Laurence, Marc, Jean-François, les deux Valentin, Paul, Baptiste, sans oublier ceux de Lanion. Une mention spéciale à Fabrice J. pour sa bonne humeur, ses blagues et ses questions détournées sur les publications d'articles. Merci à Safa, doctorante dans une autre équipe, pour ses encouragements. Merci à tous les collaborateurs qui ont donné leur consentement pour la constitution d'un corpus d'emails chez Orange. Je remercie les membres de l'équipe ALMAnaCH d'Inria Paris avec qui j'ai partagé de bons moments quand je passais au bureau.

Je souhaite également exprimer ma gratitude envers mes proches. À Carcille mon épouse, un immense merci pour ta présence et ton soutien quotidien, pour les différentes stratégies que tu as employées pour faire renaître ma motivation pendant mes moments les plus difficiles, pour toutes les fois où tu as dû abréger ton sommeil pour me réveiller afin que je puisse avancer dans mes travaux de recherche, pour la relecture de mes documents. À mes sœurs Christiane et Myrande, son mari Xavier et à mon frère Franklin, merci d'être ces points d'ancrage sur lesquelles je peux à chaque fois venir retrouver mon équilibre. À mes cousines Mary et Fabiola. À mes amis très proches que je considère comme des membres de ma famille, je pense ainsi à Kenfack Thierry Loïc,

Casimir, Vincent, Kendzo, Thierry, Raoul, Daniel, Carnot, Yves, Borel, aux Jumelles Taquefouet; je vous remercie énormément pour vos encouragements, pour les moments de joie qu'on a partagés ensemble et qui m'ont permis de me changer les idées au cours de ces années de thèse. À mon beau père et ma tante Maman Grâce, pour leurs conseils réguliers. À la grande Famille Tonfack et à la descendance de Ma'ah Zo'oh.

Et enfin, à mon père Tadonfouet Pierre-Marie, qui m'a transmis la discipline, le sérieux et la tolérance, un grand merci pour tes précieux conseils qui font de moi celui que je suis aujourd'hui. Un grand merci à ma mère, Guloung Célestine, pour m'avoir transmis la résilience, le dynamisme, l'humilité, la solidarité et l'amour. Merci d'avoir toujours cru en moi. Je vous dédie ce manuscrit avec tout mon amour et ma reconnaissance.

Résumé

Constituer des fils de conversations cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration est un processus de transformation d'une conversation écrite et asynchrone en sous-conversations. Chacune de ces sous-conversations traitant d'un sujet spécifique tout en gardant l'ordre d'arrivée des messages émis par les interlocuteurs dans la conversation originale. Ces sous-conversations donnent ainsi lieu à des structures de conversations linéaires ou arborescentes. Ce processus peut s'appliquer sur les discussions de forum mais aussi sur des conversations d'emails, ces deux exemples étant plus généralement des représentants de Contenus Médiés Par Ordinateur (CMO).

Pour constituer ces sous-fils de conversations d'emails, il est nécessaire de s'appuyer sur les métadonnées de ceux-ci et leurs contenus. Néanmoins, ces éléments ne nous semblent pas suffisants en pratique. En effet, une conversation par email est en fait un dialogue avec une structure discursive potentiellement utile pour suivre l'évolution de la discussion. Il faut cependant noter que ce dialogue est asynchrone, ce qui introduit des spécificités. Dans les dialogues synchrones, il ressort très souvent des relations très fortes entre des énoncés consécutifs qui dans un long échange peuvent ainsi constituer des clusters de sous-conversations. Pour constituer des sous-fils de conversations à partir de conversations originales d'emails, nous nous appuyons sur ce type de relations entre les phrases d'emails successifs : ces relations sont dites **transverses**. Contrairement aux dialogues où ces relations peuvent facilement être identifiées, ceci est une tâche très complexe pour ce qui est des conversations d'emails et constitue la principale sous-problématique nommée **appariement d'énoncés** pour laquelle nous proposons des approches de résolution. Les conversations regorgent généralement beaucoup d'informations linguistiques et paralinguistiques, les **actes de dialogue** en font partie, ils aident très souvent à mieux cerner le contenu d'un échange et pourrait fortement contribuer à constituer des sous-fils de conversations via une meilleure identification des relations entre des énoncés. Ceci est l'hypothèse que nous posons dans le cadre de la résolution du problème d'**appariement d'énoncés**, s'appuyant sur une première phase de **classification d'énoncés de dialogues**.

Dans le manuscrit, nous présentons les travaux connexes à notre problématique de base, ainsi que les sous-problématiques mentionnées ci-dessus. Autour de cet axe de travail principal, nous abordons divers aspects connexes mais importants, nécessaires ou utiles. Ainsi, nous abordons de façon approfondie ce que sont les CMO, l'analyse discursive et son historicité ainsi que les corpus disponibles pour approcher de tels problèmes. Ensuite nous proposons différentes approches de résolution de nos sous-problématiques avec des expériences bien détaillées et des évaluations de nos approches. Enfin, notre manuscrit se clôture sur des propositions telles que : l'application des approches proposées à d'autres types de CMO comme les forums et d'autres pistes à explorer pour résoudre la problématique de constitution de sous-fils de conversation.

Abstract

Constituting coherent threads of conversation from professional communication and collaboration tools is a process of transforming a written, asynchronous conversation into sub-conversations, each dealing with a specific topic while maintining the order of arrival of the messages sent by interlocutors in the original conversation. These sub-conversations thus result in linear or tree-like conversation structures. This process can be applied to forum discussions but also to e-mail conversations, both examples being more generally representative of Computer Mediated Content (CMC).

To build up these sub-threads of e-mail conversations, we need to rely on their metadata and content. In practice, however, these elements do not seem sufficient. An e-mail conversation is, in fact, a dialogue with a discursive structure that is potentially useful for tracking the evolution of the discussion. It should be noted, however, that this dialogue is asynchronous, which emphasises specificities. In synchronous dialogues, very strong relationships often emerge between consecutive utterances, which in a long discussion can form clusters of sub-conversations. The constitution of conversation sub-threads from main conversations is based in this type of relationships between the sentences of successive emails in a conversation : this type of relationship is refered to as **transverse**. Unlike dialogues, where such relations can easily be identified, this is a very complex task in email conversations and constitutes the main sub-problem called **statement matching** for which we suggest several resolution methods.

Conversations generally abound in linguistic and paralinguistic information, among which are **dialogue acts**. They very often help to better identify the content of a dialogue and could strongly contribute to constituting conversation sub-threads via a better identification of relations between utterances. This is the hypothesis we state in the context of solving the **statement matching** problem, based on an initial phase of **classification of dialogue statements**.

This manuscript decribes the work related to our core problem, as well as the sub-problems mentioned above. Around this main focus, we address various related but important, necessary or useful aspects. Thus, we take an in-depth look at CMOs, discourse analysis and its historicity, as well as the available corpus to approach such problems. Then we offer different resolution methods for our sub-problems, with well-detailed experiments and evaluations of said methods. Finally, our manuscript concludes with the following propositions : the application of the proposed methods to other types of CMO, such as forums, and other possibilities to be explored to solve the problem of constituting conversational sub-threads.

Table des matières

Acronymes

AD	Acte de dialogue
ADs	Actes de dialogue
AE	Appariement d'énoncés ou de segments de texte
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short Term Memory
CLEADs	CLassification d'Énoncés en Actes de Dialogue
CMO	Communication Médiée par Ordinateur
CRF	Conditional Random Fields (Champs markoviens conditionnels)
CSV	Comma-Separated Values
CTD	Contrôleur technique des données
DAR	Dialogue Acts Recognition (Reconnaissance d'actes de dialogue)
DiAML	Dialog Act Markup Language
ETL	Extraction, Transformation, Load/Chargement
GPU	Graphics processing unit(unité de traitement graphique)
IAA	Inter-Annotators Agreement (Accord inter-annotateurs)
JSON	JavaScript Object Notation
MRDA	Meeting Recorder Dialogue Act Corpus
NER	Named entities recognition (Reconnaissance d'entités nommées)
PCA	Principal component analysis
RGPD	Règlement général sur la protection des données
SMS	Short Message Service
TALN	Traitement automatique de langage naturel

1 Introduction

Constituer des fils de conversations cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration est un processus de transformation d'une conversation écrite et asynchrone en sous-conversations, chacune traitant d'un sujet spécifique tout en gardant l'ordre d'arrivée des messages émis par les interlocuteurs dans la conversation originale. Ces sous-conversations donnent ainsi lieu à des structures de conversations linéaires ou arborescentes. Ce processus peut s'appliquer sur les discussions de forum, mais aussi sur des conversations d'emails. Pour les collaborateurs d'une entreprise, dans leur boîte d'emails respective, un tel processus lorsqu'aboutit facilitera l'accès à la quintessence d'information, permettra pour une conversation : de savoir les différentes sous-thématiques abordées dans celle-ci, d'obtenir des réponses claires et sans ambiguïté aux problèmes et questions posés, de suivre son évolution et ainsi savoir si une conversation est close ou en attente de compléments d'informations. Il permettra aussi à ces collaborateurs de facilement identifier leurs engagements, contributions et actions courantes sur un ou plusieurs projets, facilitant la compréhension des sujets abordés dans des projets et ainsi une meilleure gestion de ceux-ci au sein d'une entreprise.

Cependant, il n'existe pas de travaux de par nos recherches qui portent sur la constitution de sous-fils de conversations pour les emails tel que nous l'avons décrit, c'est pour cette raison que ladite constitution de sous-fils de conversations est la principale problématique des travaux de cette thèse. En plus, il n'existe de pas de corpus de conversations d'emails en langue française disponible. Dans ces travaux, nous essayons de répondre aux questions suivantes :

— Comment constituer un corpus de conversations d'emails en entreprise en respectant le RGPD et le secret de la correspondance.

— Comment détermine-t-on qu'un fil de conversation d'emails peut être transformé en sous-fils de conversations.

— Quelles approches adoptées pour constituer des sous-fils de conversations cohérentes à partir d'une conversation originale d'emails

1.1 Contexte général

Dans le monde de la technologie et plus précisément dans le domaine de la communication, on assiste de plus en plus à une évolution marquée tant sur le plan matériel que logiciel. On peut ainsi noter la sortie ces dernières années de téléphones très puissants avec des capacités pouvant supporter de multiples applications, telles que les natives (appels, SMS, calendrier, etc.) améliorées avec des services intelligents comme l'autocomplétion des textes lors de l'envoi d'un SMS, des assistants vocaux pour appeler des contacts. Les jeux vidéo et les applications de communications ont tout aussi connu des avancées considérables avec le plus qu'a apporté l'intelligence artificielle

dans tous les domaines. Concernant les applications de communications, on a assisté à un boom d'applications de chat telles que WhatsApp, Messenger, Viber, WeChat, Line, etc.; des plates-formes de communication collaborative telles que Slack, Discord, Fleep, Workplace, etc. On ne peut parler d'outil de communication sans mentionner le principal d'entre eux utilisé en entreprise à savoir l'email au travers des applications et services comme Outlook, Gmail, Yahoo, etc.

Selon les estimations de The Radicati Group concernant la messagerie professionnelle, le nombre d'emails 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 emails par jour en moyenne et en a reçu 96, dont 19 étaient du spam. Cette explosion d'outils numériques de communication et de collaboration a conduit à la génération d'énormes quantités de données notamment textuelles qui sont en général sont stockées sur des serveurs ou des postes personnels sous formes d'archives et sont au fur et à mesure supprimées pour des raisons de limitation d'espace de stockage ou de non exploitation. En général ces données ne sont pas exploitées du fait qu'elles sont peu ou pas structurées.
Comme exploitation primaire, on a la recherche d'informations faite par des cadres en entreprise. Ces derniers passent beaucoup de temps de travail sur cette tâche de recherche d'informations, dans leur messagerie par exemple. Des études estiment ente 20 et 30% ce temps passé à chercher de l'information enfouie dans les emails.

Concernant les conversations d'emails, elles regorgent d'informations très importantes surtout dans le contexte professionnel où plusieurs collaborateurs échangent entre eux via emails dans le but d'atteindre un objectif commun comme, par exemple l'aboutissement d'un projet. Ces conversations échangées se présentent en général sous plusieurs formes (réponses, transferts, mailing list, ...) avec ou sans pièces jointes. Les emails vont de simples envois à des transferts, en passant par des réponses. Ces dernières sont parfois faites de façon imbriquée, c'est-à-dire que les phrases sont insérées directement au niveau de celles de l'email auquel on répond. On retrouve aussi des emails transférés qui donnent lieu à de nouvelles conversations. La non-structuration des emails est en partie due à ces formats d'édition d'emails non standardisés, entraînant ainsi un accès difficile aux informations véhiculées dans ces emails. Cet accès difficile ou incompréhension de l'information n'est pas seulement une conséquence de l'absence ou d'une mauvaise structuration des emails, mais elle est aussi fonction des aspects linguistiques très variés utilisés par les participants dans une conversation par emails. Ces incompréhensions sont très souvent dues aussi à l'absence de connaissance d'un contexte commun par les participants.

Beaucoup de recherches ont émergé ces dernières décennies afin d'exploiter ces grandes quantités d'information issues notamment des emails, mais aussi de tenter de résoudre ces problèmes de structuration logique et d'incompréhension soulevés précédemment. Ceci se fait via des techniques à base de règles ou de traitement automatique du langage naturel. Dès les années 1990 des travaux de recherche portent sur l'apprentissage automatique appliqué à la classification de documents selon leurs thématiques (Lewis, 1992). Quelques années plus tard l'apprentissage automatique est également utilisé pour classer des emails et pour détecter des sentiments (Cohen, 1996b; Pang et al., 2002). En s'appuyant sur les travaux de Searle portant sur les actes de langage (Searle, 1975) et sur ceux de (Finke et al., 2002) visant à détecter automatiquement des actes de langage dans des conversations téléphoniques (Cohen et al., 2004) proposent une approche visant à détecter l'intention des auteurs des emails grâce à une ontologie d'actes de langage. Leurs conclusions étaient, 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**. Ces problématiques sont abordées avec différentes approches comme nous allons le voir dans la partie I. Au-delà de ces deux problématiques identifiées, il en existe bien d'autres qui nécessitent qu'on s'y intéresse particulièrement parce qu'elles sont soit connexes ou imbriquées aux précédentes. Il s'agit de la **reconnaissance d'actes de langage** et de **la reconstruction de fils de conversations** dans les conversations asynchrones. La première est un sujet largement traité comme nous verrons dans le chapitre 3. La seconde elle aussi a été le centre d'intérêt de certains travaux (chapitre 2), elle permet comme son nom l'indique de reconstruire des fils de conversations dans les emails par exemple en s'appuyant sur les métadonnées et les contenus d'emails.

Le démêlage des discussions ou conversations imbriquées permet de regrouper des segments de texte de conversation, chacun des groupes constitués porte sur une thématique ou sous-thématique bien précise. Dans une conversation d'emails par exemple, plusieurs sujets sont souvent abordés et donc le démêlage de conversation permettrait de constituer des clusters des segments de texte extraits de ladite conversation. Pour constituer ces clusters, les approches pour démêler ces conversations s'appuient généralement sur les coefficients de similarité sémantique segments de textes. Ces coefficients qui sont calculés sur la base des représentations vectorielles desdits segments de texte. Cependant certains segments de texte sont parfois dépourvus de mots, bigrammes, trigrammes ou expressions permettant leur rapprochement sémantique à d'autres énoncées dans la même conversation. Ces segments de texte sont généralement constitués de deux ou trois mots et sont très souvent des acquiescements, des appréciations, des formules de politesse, etc. qui sont des actes de dialogue. Ainsi le démêlage de conversation ne saurait être l'approche la mieux adaptée pour constituer des sous-fils de conversation d'emails, ceci du fait de la mauvaise classification d'un certain type d'énoncés.

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 métadonnées 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

1. message threading

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 de nos travaux. 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 1.1 met en avant une conversation d'emails avec ses métadonnées, ainsi que les paires (texte en surbrillance de même couleur entre deux emails distincts) transverses de segments de texte ou d'énoncés de ladite conversation. Notre approche de constitution de sous-fils de conversation permet non seulement de démêler de façon fine une conversation d'emails mais aussi surtout de connaître l'état d'évolution de ladite conversation. Nous avons dans un premier temps, exploré une approche en deux étapes avec en amont la classification d'énoncés de conversation en acte de dialogue et en aval l'appariement des énoncés de la conversation ; et par la suite nous proposons une approche de bout-en-bout d'appariement d'énoncés d'une conversation.

Dans la suite de notre document, nous allons respectivement utiliser les acronymes **AD** et **ADs** pour un acte de dialogue et pour les actes de dialogue. Et **AE** pour appariement d'énoncés ou de segments de texte.

1.2 Cadre de la thèse

Cette thèse s'est effectuée dans le cadre d'une convention CIFRE entre Orange Innovation et Inria.

Orange Innovation : Koen Vermeulen est le responsable d'Orange Innovation, anciennement appelé Orange Labs, une branche de l'entreprise dédiée à la recherche et au développement.

L'équipe avec laquelle j'ai travaillé dans le cadre de ma thèse est affectée au projet "SmartWorking DataIA". Ce nom de projet, composé des mots clés "smart", "working", "data" et "IA" [2], vise à améliorer la qualité de vie des collaborateurs tout en augmentant leur productivité sur leur poste de travail. Ainsi, l'objectif de l'équipe est de concevoir des capteurs ambiants permettant de savoir :

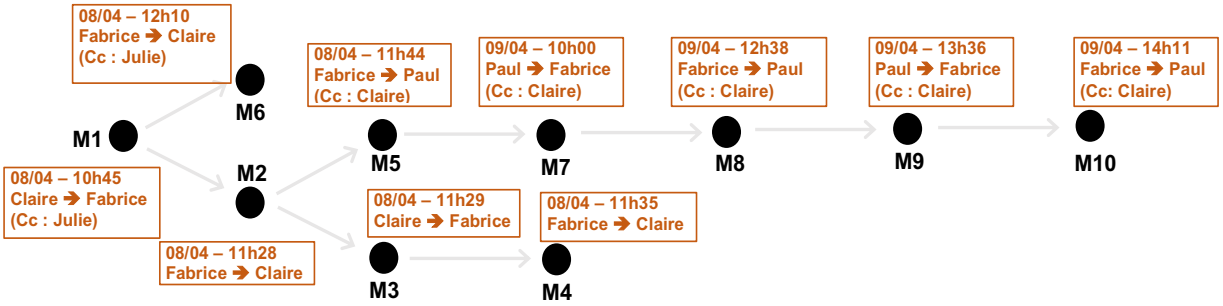
— Quel est le niveau d'attention d'un collaborateur sur une ou plusieurs fenêtres ?

— Quels sont les usages qu'il fait de son poste de travail ?

— Dans quel contexte se trouve-t-il ?

La brique logicielle qui servira à la collecte et l'affichage de ces informations fonctionnerait sur les différents supports utilisés par les collaborateurs d'Orange. Elle est appelée LiveWebble, en référence aux mots "live", "hubble" [3] et "bubble". C'est dans les locaux caennais d'Orange Innovations que j'ai effectué à 80% mes travaux de thèses, co-encadré par Fabrice Bourge et Tiphaine Marie.

2. Acronyme d'intelligence artificielle
3. Hubble est un télescope spatial en orbite depuis 1990



M1	*Voici comme convenu la liste des stopwords supprimés dans le prétraitement avant envoi à SuperOutil.* *Je te laisse regarder pour épurer cette liste. L'idée est de ne plus tomber dans les pièges de type :* • *« pas urgent » transformé en « urgent » ; « reste à faire » transformé en « reste faire » ; …* *Au passage, visiblement pas de mots en anglais dans cette liste. Si on veut être plus logique, il faudrait en avoir…* *Bonne journée,* *Claire*
M2	*Ok, je regarde ça. Néanmoins, on ne peut pas éviter le 2ème problème que tu évoques, sauf à enlever « à » de la liste des stopwords, ce qui me paraît idiot.* *Il faudrait peut-être (urgemment) demander à Paul s'il recommande la suppression des stopwords avec SuperOutil ?*
M3	*C'était juste un exemple. Après je suis d'accord de voir avec Paul pour la notion de stopwords.* *Tu peux lui envoyer un mail ? ou tu préfères que je fasse ?* *Claire*
M4	*Je le fais.*
M5	*Bonjour Paul,* *Avant d'utiliser SuperOutil, nous « nettoyons » les inputs (sujet + contenu de mails) en retirant des stopwords. Mais nous avons constaté que le retrait de certains stopwords nous jouent des tours à une autre étape de notre processus.* *Nous avons donc deux possibilités :* *- soit réduire la liste de nos stopwords pour tenter d'éviter le problème constaté,* *- soit ne pas effectuer du tout de suppression des stopwords.* *Quelle est ta recommandation quant au nettoyage ou non des inputs (via liste de stopwords) avant de les soumettre à SuperOutil ?* *Cdlt,* *Fabrice*
M6	*En attendant la réponse de Paul sur l'intérêt d'enlever les stopwords, voici la liste que je propose (fichier Doc, ainsi que l'Excel à partir duquel je l'ai généré).* *Le raisonnement que j'ai appliqué est le suivant : quand un stopword indique un jugement de valeur (au sens large), je l'ai conservé. C'est subjectif…*
M7	*Bonjour Fabrice,* *SuperOutil n'a normalement pas besoin de nettoyage des stopwords (ils auront des contributions faibles avec la cible). Cela peut avoir des effets négatifs de les supprimer car cela fausse la génération des ngrams (sauf si vous n'utilisez que des ngrams de longueur 1).* *Vous retirez les stopwords en apprentissage et en prédiction ?* *Cordialement,*
M8	*Jusqu'à présent, on retire les stopwords à la fois en apprentissage et en prédiction.* *Actuellement, nous travaillons sur des n-grams de longueur maxi = 5.*
M9	*OK.* *Je déconseille quand même de retirer les stopwords, superOutil est prévu pour pouvoir y résister, et les retirer génère des effets de bords. Vous avez besoin de retirer les stopwords pour une autre fonctionnalité ?*
M10	*On a pris la décision de ne plus les retirer, et on va voir si ça améliore nos résultats.* *Merci.*

FIGURE 1.1 – Conversation d'emails avec ses métadonnées et appariements d'énoncés

Inria (Paris) : Institut national de recherche en sciences et technologies du numérique. Nos travaux ont été menés au sein de l'équipe **ALMAnaCH** qui fait de la recherche en intelligence artificielle associant Traitement du Langage Naturel et Humanités Computationnelles. J'y ai été supervisé par Laurent Romary et co-encadré par Eric de la Clergerie.

1.3 Structure de la thèse

La thèse est structurée en quatre parties principales : Travaux Connexes, Modalités de communication et analyse discursive, Corpus utilisés et méthodes proposées et enfin Conclusion et perspectives.

La première partie, Travaux Connexes, s'intéresse en général à l'état sur les différentes problématiques connexes à la nôtre, elle contient trois chapitres. Le chapitre 2 se concentre sur la reconstruction des fils de conversation, y compris les approches basées sur les métadonnées et le contenu des emails, ainsi que la reconstruction des fils de conversation avec des structures arborescentes. Le chapitre 3 quant à lui se focalise sur les travaux menés dans le cadre de la résolution de la problématique de démêlage de conversations sur des données de forums. Dans le chapitre 4, nous présentons l'état de l'art autour de la problématique d'identification d'actes de dialogue dans les conversations issues de forums pour certains travaux et d'emails pour les autres.

La deuxième partie de ce document contient deux chapitres. Le chapitre 5 traite de la communication médiatisée par ordinateur (CMO), y compris ses avantages et ses inconvénients. Il se penche aussi sur les types de communications : écrites, orales, synchrones et asynchrones avec un accent sur les emails et les forums. Il détaille les caractéristiques respectives de ces derniers, mais aussi donne un aperçu de l'analyse discursive dans le contexte des emails. Le chapitre 6 présente l'analyse du discours, son évolution en se basant sur les différentes taxonomies et schémas d'annotation et enfin il détaille les raisons de la mise en œuvre d'un référentiel d'acte de dialogue que nous avons effectué pour mieux approcher notre problématique.

La troisième partie intitulé « Corpus utilisés et méthodes proposées », possède quatre chapitres. Le chapitre 7 liste les différents corpus utilisés lors de nos travaux avec leurs caractéristiques respectives ; ce chapitre décrit aussi le processus de constitution et pseudo-anonymisation du corpus d'emails d'Orange. Le chapitre 8 détaille les expériences effectuées pour la reconnaissance ou classification en actes de dialogues de segments de textes extraits des fora et des emails. Le chapitre 9 quant à lui porte sur l'approche d'appariement de segments de texte de façon transverse en s'appuyant sur les actes de dialogues, ceci pour la résolution de notre problématique. Enfin le chapitre 10 présente les évaluations et analyses de nos résultats.

Le chapitre 11 de la quatrième partie de ce manuscrit conclut et présente les perspectives de cette thèse de doctorat.

Enfin, la cinquième partie contient toutes les annexes ainsi que des informations supplémentaires pertinentes pour notre travail.

Première partie

Travaux Connexes

2 Reconstruction de fils de conversations d'emails

2.1 Introduction

Les emails sont un outil de communication et de collaboration largement utilisé en entreprise et ontiennent des informations très riches, mais difficilement accessibles parce qu'entremêlées dans des conversations d'emails qui sont peu ou pas structurées. La reconstruction de fils de conversations d'emails et l'identification des thématiques abordées dans ces conversations sont les principales problématiques liées à cette modalité de communication. Pour répondre à ces problématiques, il existe plusieurs approches.

2.2 Reconstruction de fils de conversations : Métadonnées, Contenus et structure arborescente

2.2.1 Jamie Zawinski et l'enfilage des messages (**message threading**)

Dans le contexte des systèmes de messagerie électronique, l'enfilage des messages (**message threading**) fait référence à l'organisation et au regroupement des messages associés dans une conversation ou un fil de conversation. Il permet aux utilisateurs de visualiser les messages de manière structurée, ce qui facilite le suivi du déroulement d'une discussion. Jamie Zawinski [1] a été un pionnier sur cette thématique d'enfilage des messages; il a développé un algorithme qui a révolutionné la façon dont les clients de messagerie organisent et présentent les conversations. Cet algorithme innovant regroupe intelligemment les messages connexes, créant ainsi un fil de conversation cohérent. L'algorithme de threading (enfilage) de Zawinski analyse les en-têtes et le contenu des messages électroniques pour identifier les relations entre eux. Il recherche les points communs tels que les sujets des messages, les adresses email et les références dans les en-têtes des emails. En identifiant ces modèles, l'algorithme peut déterminer quels messages appartiennent au même fil de conversation. L'algorithme prend en compte différentes variations dans les clients et protocoles de messagerie, en tenant compte des diverses manières dont les systèmes de messagerie gèrent les informations de thread. Cela garantit que les messages provenant de différents clients de messagerie, même ceux qui n'indiquent pas explicitement les relations de thread, peuvent toujours être efficacement liés entre eux.
L'implémentation de threading de Zawinski, initialement développée pour *Netscape Messenger* (plus tard incorporée dans *Mozilla Thunderbird*), offrait aux utilisateurs un moyen pratique de visualiser et de naviguer dans les conversations de courrier électronique. Au lieu d'avoir des messages

1. https ://www.jwz.org/doc/threading.html

individuels encombrant la boîte de réception, les messages associés ont été regroupés, offrant une structure claire et logique. Son travail sur le threading des messages a considérablement amélioré l'expérience utilisateur des clients de messagerie, facilitant le suivi des discussions, la réponse en contexte et la gestion de la surcharge des emails. La vue filetée permettait aux utilisateurs de voir l'ordre chronologique des messages au sein d'une conversation, facilitant une communication efficace et réduisant la surcharge d'informations.

De plus, les contributions de Zawinski au threading () des messages allaient au-delà de *Netscape Messenger* et *Thunderbird*. Son algorithme de threading a influencé le développement d'autres clients de messagerie et systèmes de messagerie, à la fois open source et propriétaires, qui ont adopté des approches similaires pour organiser les conversations. Le travail de Zawinski sur l'enfilage (threading) des messages met en valeur son expertise dans le développement de solutions pratiques pour gérer et présenter de gros volumes de messages électroniques. L'impact de son algorithme est encore visible aujourd'hui, car le threading est devenu une fonctionnalité standard dans de nombreux clients de messagerie et plates-formes de communication, améliorant la productivité et améliorant l'expérience utilisateur dans la communication par courrier électronique.

2.2.2 Approches s'appuyant sur les métadonnées et les contenus d'emails

(Wu and Oard, 2005) utilisent les sujets d'emails pour l'indexation et la récupération des emails dans les fils de discussion. Les auteurs ont mené des expériences en utilisant une collection de listes de diffusion publiques et ont constaté que le regroupement automatique des sujets et la suppression du texte en double avaient peu d'impact sur l'efficacité de la récupération. Ils ont également développé une typologie des questions pour la récupération des emails. Dans leur expérience, ils ont utilisé les listes de diffusion archivées du World Wide Web Consortium (W3C) comme ensemble de données pour l'analyse. La collection comprenait 161 645 messages d'une taille totale de 474 Mo. Les auteurs ont exploré le concept de granularité des documents et si l'enfilage de messages avec des lignes d'objet similaires améliorait l'efficacité de la récupération. Ils ont constaté que la "bonne" granularité du document dépendait de la nature de la question posée.

Pour mener les expériences, les auteurs ont créé trois index à l'aide de *Lucene*, une bibliothèque de moteur de recherche de texte. Les index comprenaient des messages individuels, des discussions avec du texte inclus conservé et des discussions avec du texte inclus supprimé. Ils ont formulé trois questions de justification de la conception et évalué le nombre de messages pertinents dans les 10 principaux documents pour chaque question. Les résultats ont montré peu de différence entre les trois indices, suggérant que le threading et le texte inclus peuvent ne pas être initialement bénéfiques pour la récupération.

Cependant, les auteurs ont observé que la recherche dans l'index des threads donnait des threads plus longs par rapport à la collection dans son ensemble. Cet effet d'enrichissement a été attribué aux fonctions de pondération modernes présentant une préférence pour les documents plus longs. Ils ont également noté que les emails quasi-dupliqués étaient courants dans les collections d'emails, mais le threading a aidé à résoudre ce problème car les quasi-doublons apparaissaient souvent ensemble dans le même fil. Ils ont également souligné l'importance d'un examen attentif lors de la sélection et de l'interprétation des mesures d'évaluation. Ils ont mené des expériences avec une question demandant des informations de contact et ont constaté que la comparaison des mesures

basées sur les messages et les threads nécessitait de la prudence, car les résultats pouvaient être trompeurs.

(Wang et al., 2008) proposent un mécanisme de reconstruction des conversations basé sur des threads qui fournit une analyse et des statistiques efficaces des flux d'emails pour plusieurs personnes. Le mécanisme comprend une règle d'extraction de données pour l'extraction des en-têtes d'emails et le filtrage redondant des emails ; un algorithme de correspondance des messages pour conserver les messages sans identifiant de message dans une relation parent/enfant correcte et une heuristique basée sur les objets d'emails pour fusionner ou décomposer les threads en conversations. Les résultats de leurs expériences montrent que leur mécanisme a de hautes performances en matière de détection, de suivi et de maintien de la relation parent/enfant des conversations.

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 emails en plusieurs groupes, où tous les messages du même groupe ont des objets identiques et la différence de temps maximale entre deux emails 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 fil de conversation. De son côté, *LExTreC* essaie d'apprendre des relations parents-enfants parmi les emails et extrait la structure arborescente des conversations.

De façon plus détaillée, nous présentons les travaux de (Domeniconi et al., 2016) et (Avigdor-Elgrabli et al., 2018) qui s'appuient sur l'utilisation des métadonnées et des contenus d'emails.

(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 construisent ainsi la *feature* sémantique d'un email via une requête sur la plateforme AlchemyAPI [2]. Cette

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

plateforme 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 un indice de confidence compris entre 0 et 1. Ces valeurs ont permis de construire trois vecteurs sémantiques pour chaque message. Ces vecteurs ont ensuite été utilisés 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

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

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 porte sur les temps de réponse et se calcule par une similarité qui est égale au logarithme de l'inverse de la distance temporelle entre deux messages :

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

L'inverse de la norme de distances est utilisé ici pour avoir des valeurs entre 0 et 1. La figure 2.1 montre 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 différentes 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 (deux emails) 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)

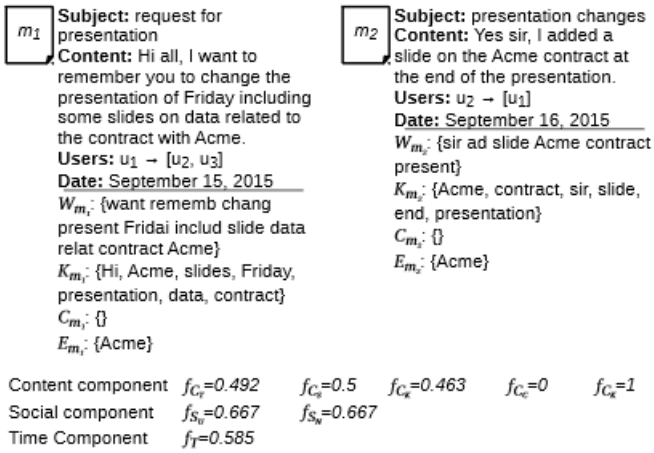


FIGURE 2.1 – Exemple de calcul de similarité des *features* entre deux messages

et à les fusionner progressivement à partir d'une liaison moyenne entre eux.
Dans ce manuscrit 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 clustering. 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 supervisé, 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 non-supervisé sur sept corpus :

— Le corpus BC3 (Ulrich et al., 2008)

— Un sous ensemble de liste de diffusion du corpus Apache d'août 2011 à Mars 2012 [3]

— Des listes de diffusion du corpus Fedrora Redhat Projet collectés pendant 6 mois en 2009 [4]

— le contenu de deux pages Facebook publiques nommées *Healthy Choice* [5] et *World Health Organizations* [6]

— Le contenu de groupes publiques de Facebook au nom de *Healthcare Advice* [7] et *Ireland Support Android* [8]

3. tomcat.apache.org/mail/dev
4. www.redhat.com/archives/fedora-devel-list
5. www.facebook.com/healthychoice
6. www.facebook.com/WHO
7. www.facebook.com/groups/533592236741787
8. www.facebook.com/groups/848992498510493

Enfin ils montrent les meilleures performances de leurs différentes approches comparées à celles de (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 n'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 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 cet 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 de 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 indisponible de boîtes de messagerie de Yahoo. 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 mesurent 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'envoi 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 d'emails) et des poids nommés ***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 positifs ou négatifs 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 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 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 toutes 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 de l'art 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 message, ceci

pouvant 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 email, 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**.

Toutes ces approches de reconstruction de fils de conversation s'appuient fortement sur les méta-

données notamment les en-têtes d'emails et les objets de ceux-ci. Cependant, les emails proviennent généralement de serveurs différents qui ont chacun leur système d'encodage d'en-têtes différents les uns des autres, ce qui rend difficile l'utilisation des en-têtes pour la reconstruction de fils de conversations d'emails. Concernant l'utilisation des objets d'emails, il est tout à fait logique que des emails d'une même conversation aient en commun un même objet parfois précédé des mots clés spécifiant si l'email est une réponse ou un transfert (*forward*). Il peut parfois arriver que ces objets d'emails soient changés par un interlocuteur dans la conversation parce que celui-ci introduit un nouveau sujet dans la dite conversation. S'appuyer sur les objets et les contenus d'emails comme cela a été fait dans certains travaux pour reconstruire les fils de conversations sont des approches intéressantes qui exploitent le contenu entier des emails avec des mécanismes de calcul de similarité et de clustering. Cependant, ces contenus entiers d'emails abordent très souvent plus d'un sujet et ceci est parfois explicitement indiqué par le changement d'objet dans une même conversation. Certains travaux que nous présentons dans la prochaine section s'intéressent à l'identification de différentes thématiques dans ces contenus entiers d'emails, d'autres exploitent ces différentes thématiques existantes pour démêler les conversations.

2.3 Identification de thématiques et démêlage de conversation d'emails

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 dispersése au fil du temps.

Pour extraire ces connaissances, l'identification des thématiques dans les conversations d'emails est une manière d'isoler les segments de texte bien précis . Cependant effectuer une telle isolation reste une problématique difficile à approcher du fait de la petite taille d'une seule et unique conversation d'emails souvent considérée comme un document. Ce document dans lequel la recherche de thématiques ou de sous thématiques va être effectuée et dont les résultats seront imprécis induite par un manque de densité d'informations (à cause de la petite taille du document). L'autre problé-

matique est le démêlage de conversation qui consiste de manière générale à regrouper des segments de texte qui peuvent être de différentes granularités (phrases, paragraphes, etc.) en fonction de leur appartenance à une même thématique. Ces segments de texte sont groupés de façon indépendante des fils de conversations et sur de gros volumes de données.

2.3.1 Identification de thématiques dans les conversations

(Joty et al., 2010) ont annoté et rendu disponible le corpus BC3 annoté manuellement avec des thématiques. 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. Ce modèle commence par calculer des chaînes lexicales pour chaque mot qui ne fait pas partie des stop-words. 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. D'une part une faible similarité indique une faible cohésion lexicale et d'autre part un changement net signale une forte probabilité d'une frontière de sujet réelle.

2.3.2 Démêlage de conversations

En ce qui concerne le démêlage des conversations, Elsner and Charniak (2010, 2011) abordent le démêlage des *chats* en utilisant d'abord un classificateur binaire, puis des modèles de cohérence locale un an plus tard. Jiang et al. (2018) tirent parti de l'apprentissage des représentations linguistiques pour démêler les conversations. Ils calculent les similitudes entre les messages à l'aide d'un modèle qu'ils nomment *Siamese Hierarchical Convolutional Neural Network (SHCNN)*, autrement dit un réseau convolutionnel hiérarchique siamois.
(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 2.2 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

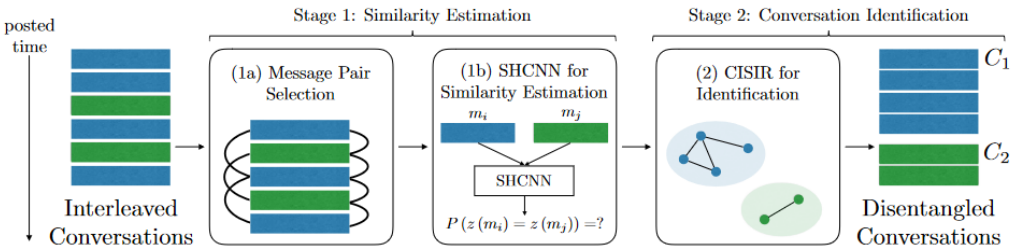


FIGURE 2.2 – Illustration des étapes de la méthode de (Jiang et al., 2018)

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 2.3 montre les différentes convolutions faites par un HCNN pour l'obtention du vecteur de dimension 128 représentatifs des deux niveaux sémantiques. Une fois

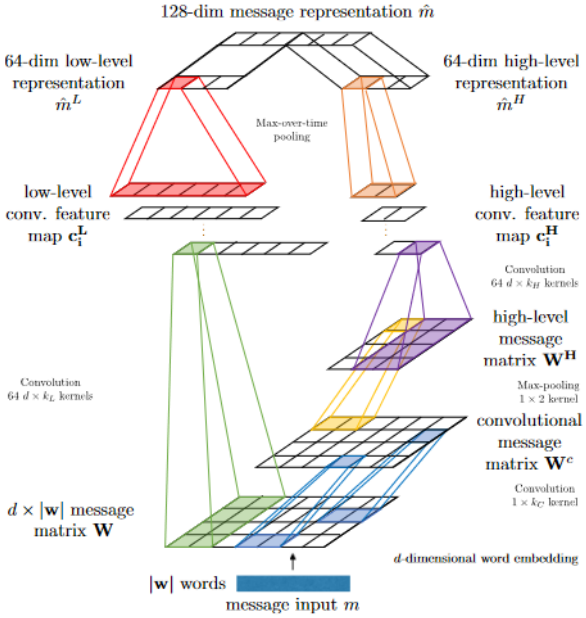


FIGURE 2.3 – Représentation hiérarchique d'un message via HCNN; Les étiquettes avec une taille de police plus grande indiquent les tenseurs correspondants, et les étiquettes avec une taille de police plus petite expliquent les opérations entre les tenseurs.

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 résultat

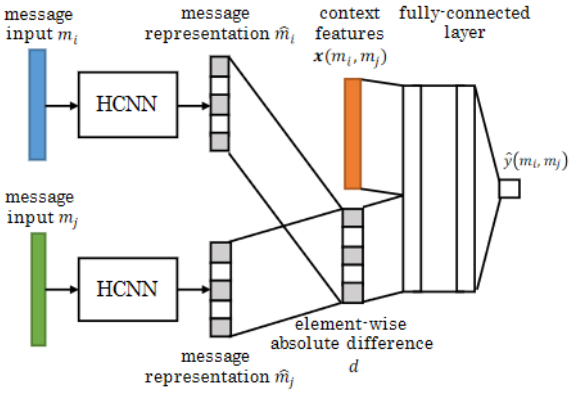


FIGURE 2.4 – Le CNN hiérarchique siamois (SHCNN) pour l'estimation de similitude.

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 2.4.

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'appuie 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 par rapport à six approches de l'état de l'art sur 4 corpus et les résultats montrent que leur approche surpasse celles de l'état de l'art.

Ces travaux de (Jiang et al., 2018) bien présentés dans leur manuscrit, 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 de messages de conversation de plus de 10 mots alors que dans certaines conversations les messages de tailles plus petites peuvent être porteurs d'informations importantes, comme par exemple des réponses d'accord, du genre « ça me va ».

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.

2.4 Conclusion

Nous venons de présenter de façon succincte pour certains et plus détaillée pour les autres quelques avec approches rencontrées dans la littérature pour la résolution des problématiques de construction de fils de discussions, de démêlage de conversations et d'identification de thématique. Il ressort de ces travaux l'importance de la prise en compte du contexte pour avoir de bons résultats, mais aussi de l'exploitation des métadonnées et des contenus de conversations. Certains de ces travaux qui n'exploitent pas assez les contenus de messages dans les conversations résolvent la

problématique de reconstruction de fils de conversation à un niveau « macro » dans lequel les emails vont constituer un fils de discussion seulement à partir des métadonnées des conversations et des similarités avec des contenus. Un des problèmes de telles approches est parfois l'intégration de certains messages (sans ou avec des en-têtes difficiles à parser) dans des fils de conversations alors qu'ils n'en font pas partie à la base. Ce même type de problème s'observe également avec des approches qui effectuent le clustering en exploitant les contenus de messages qui vont très souvent regrouper des messages de conversations qui n'appartiennent pas au même fils de conversations. Ces approches de clustering sont néanmoins intéressantes pour des données dont les métadonnées ont été perdues lors du processus de collecte d'informations.

Ces différents travaux restent au niveau de la reconstruction de fils de conversation, ils tentent en quelque sorte de retrouver les structures originales des conversations. Contrairement à ces travaux, notre intérêt dans cette thèse est la constitution de sous-fils de conversation qui est un niveau plus fin. La granularité très fine ciblée dans nos objectifs pour reconstruire les fils de discussions ne pourrait être atteinte sans des études contextuelles plus fines. Comme mentionné plus haut, il est difficile d'identifier de façon fine des thématiques ou sous-thématiques dans ces conversations, ceci parce que certains segments de texte sont parfois dépourvu d'informations sémantiques et pourtant sont d'une importance capitale pour la compréhension d'une conversation. C'est le cas par exemple de simples accords, d'acquiescements et de bien d'autres phénomènes discursifs que l'on peut appréhender via la notion d'actes de dialogue. Ainsi, pour affiner la compréhension des conversations et au delà des thématiques, l'identification des intentions des interlocuteurs, les actes de dialogues inhérents aux messages et l'appariement des énoncés ou segments de texte portant ces actes de dialogue ou intentions sont des pistes intéressantes à creuser.

3 Identification d'Actes de Dialogue (ADs) et Appariement d'Énoncés (AE)

3.1 Introduction

Les actes de dialogue (ADs), également appelés actes de parole ou actes de communication, sont des actions linguistiques effectuées par les locuteurs au cours d'une conversation ou d'une communication pour transmettre des intentions spécifiques ou atteindre certains objectifs de communication. Ils représentent les différentes fonctions ou objectifs que la langue sert dans la communication. Identifier les actes de dialogue est essentiel pour comprendre les intentions des interlocuteurs et le sens général du discours dans des conversationsqu'elles soient synchrones ou asynchrones.

Une conversation synchrone est une conversation dans laquelle une ou plusieurs personnes échangent de l'information sur un canal vocal ou écrit de façon directe sans temps d'attente de réponses considérables. Les échanges téléphoniques, les réunions, des chats sont des exemples de conversations synchrones. À contrario, les conversations asynchrones sont des conversations dans lesquelles il existe un certain temps de latence entre un message émis par un interlocuteur et une réponse ou retour audit message par un autre interlocuteur. Les emails, les discussions de forum sont des exemples de ce type de conversations. Dans le chapitre 4, nous détaillons les différents types de conversations.

L'Appariement d'Énoncés (AE) quant à lui consiste à mettre ensemble deux segments de texte (qui peuvent être des phrases) qui partagent une certaine similarité soit sémantique ou dialogique ou bien les deux à la fois. Cette notion d'AE introduite ici est centrale dans dans les travaux de thèse. L'AE peut s'appliquer autant sur des monologues que sur des conversations, que celles-ci soient synchrones ou asynchrones. Toutefois, les conversations synchrones sont une succession d'interventions de chacun des interlocuteurs pour l'atteinte d'un objectif commun. Cette succession d'interventions généralement de petite taille (interventions vocales de courte durée ou segment de texte court) est une suite logique et cohérente de partage d'information et donc il est plus facile dans ce type de conversations d'apparier les énoncés qui la constituent. Contrairement aux conversations synchrones, les conversations asynchrones quant à elles ont des contenus un peu plus denses. Dans le corps d'un email par exemple, en plus des formules de politesses (en début et fin), celui-ci contient souvent plus d'un sujet : d'où le qualificatif de contenus entremêlés dans un email. Et dans une conversation d'emails qui possède au moins deux emails, il est difficile d'apparier de courts segments de texte extraits de ces emails, bien qu'ils suivent un ordre chronologique bien précis. C'est dans ce cas de conversations asynchrones que nous allons approcher cette problématique d'AE dans les emails et forums.

Les problématiques d'identification d'ADs et d'AE ont été approchées par différents travaux de recherche. Dans la littérature, nous avons identifié très peu de travaux sur les ADs dans les conversations d'emails et encore moins de travaux qui s'intéressent aux AE dans ce même type de conversations.

3.2 Identification d'Actes de dialogue (ADs) dans les emails

Dans cette section, nous nous intéressons principalement aux travaux d'identification d'ADs dans les conversations d'emails. L'un des premiers travaux mêlant ADs et emails sont ceux de (Cohen et al., 2004) qui proposent des méthodes d'apprentissage automatique pour classer les emails selon une ontologie de verbes et de noms, qui décrivent des « actes de dialogue d'email » voulus par l'expéditeur de l'email. Ensuite ont suivi ceux de (Carvalho and Cohen, 2005) qui décrivent un nouvel algorithme de classification de texte s'appuyant sur une méthode de classification collective, celle-ci basée sur un réseau de dépendances. Leur approche présente des améliorations significatives par rapport à un classificateur de base utilisant des représentations en sac de mots.

Avec l'évolution des représentations des contenus textuels et des algorithmes d'apprentissage automatique, de nouvelles approches d'identification d'ADs ont vu le jour. (Wang et al., 2019) étudient les intentions latentes dans les emails du corpus **Avocado**[1] qui est un grand corpus d'emails de l'ancienne entreprise de technologie de l'information éponyme Avoca. Ce corpus est une collection de 938 035 emails et de pièces jointes pour 279 comptes, majoritairement des comptes employés 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 méthode d'identification de celles-ci. 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 métadonnées. 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.

Pour mieux comprendre les caractéristiques des intentions des utilisateurs, ils sélectionnent 1300 fils discussions de façon aléatoire. Ils ont fait 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 obtenu par une stratégie de vote majoritaire. Les intentions retenues pour chaque email sont celles sélectionnées par au moins deux annotateurs. Pour l'accord inter-annotateur ils obtiennent un score Kappa de **0.694**. La figure 3.1 montre la distribution des intentions sur le corpus sélectionné et annoté. Cette

1. https://catalog.ldc.upenn.edu/LDC2015T03

même figure liste 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 3.2 montre les co-occurrences 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* ont de grandes probabilités d'existence dans un même email, tandis que le *social* et le *rappel/planification de meeting* ont de faibles probabilités.

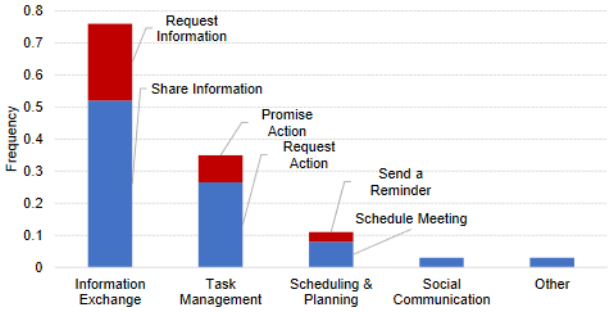


FIGURE 3.1 – Fréquence des intentions sur un sous-ensemble du corpus Avocado

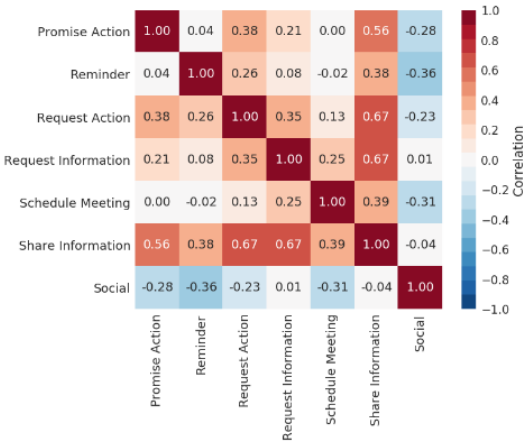


FIGURE 3.2 – Distribution de paires de sous-intentions dans un même email

Dans l'article, ils étudient 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 partir de la vérité de terrain précédemment constituée, 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 annotée par trois annotateurs et la majorité des annotations pour chacune des instances représentent sa prédiction par un humain.

Le tableau 3.1 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 cadre pour cette tâche d'identification d'intentions qui 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 3.3.

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

TABLE 3.1 – Matrice de confusion des prédictions humaines

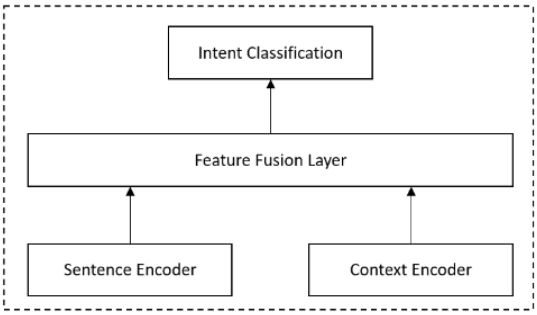


FIGURE 3.3 – Augmentation de contexte d'une phrase pour y identifier les intentions

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 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

$$W \in R^{d \times |V|}$$

où $|V|$ 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 cellules GRU (*Gated Recurrent Unit*). Ce réseau bidirectionnel produit pour chaque mot w_i de la phrase cible un vecteur

$$h_i = [\overleftarrow{h_i}, \overrightarrow{h_i}]$$

avec $\overleftarrow{h_i}$ et
$overset{\rightarrow}{h_i}$ respectivement l'état caché avant et arrière du mot w_i. Ces états permettent d'avoir des informations avant et 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 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 point à point. 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 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 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 3.4 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

	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 3.4 – Aperçu de DCRNN

sont meilleurs que les anciennes approches tant sur les prédictions avec ou sans contexte. Ce travail de (Wang et al., 2019) est l'un des rares qui traitent de la problématique d'identification des ADs avec une granularité fine dans un corpus de conversations d'emails. Cependant la reproductibilité de ces travaux est difficile, ceci est d'une part lié à l'absence du code source de leurs expériences et d'autre part dû au fait que le *Avocado* n'est pas libre d'accès même pour des recherches académiques.

(Taniguchi et al., 2020) ont annoté en actes de dialogue plus de 2k fils de conversations du corpus Enron (cf. section 6.4.5) avec deux granularités différentes : 35k phrases avec une granularité fine et 6k emails annotés avec une granularité moins fine. Ce type de corpus nous aurait été d'une aide précieuse dans le cadre de nos travaux, malheureusement il n'est pas disponible. Les ADs dans les emails sont considérés comme des genres d'emails par (Goldstein and Sabin, 2006) et ils utilisent des algorithmes de machine learning (arbre aléatoire et vecteur à support machine) pour classifier les emails en des genres bien précis.

La problématique d'identification d'ADs sur des corpus d'emails n'a pas été largement abordée dans la littérature comme d'autres problématiques, probablement à cause de la rareté des corpus d'emails annotées. Cependant cette problématique est largement abordée sur d'autres types de corpus de conversations. C'est le cas avec des corpus des réseaux sociaux. (Laurenti et al., 2022) utilisent ainsi deux niveaux de classification d'énoncés en ADs pour la gestion de crise dans les Médias Sociaux. De même (Cerisara et al., 2018) font de la reconnaissance d'ADs et de sentiment avec des modèles multi-tâches sur les données du réseau social *Mastodon* qui ressemble beaucoup à *Twitter*. Le corpus *MRDA (Meeting Recorder Dialog Act Corpus)* qui est un corpus de transcription d'enregistrements de réunions est aussi largement utilisé pour l'identification des ADs (Chen et al., 2018; Li et al., 2019a; Raheja and Tetreault, 2019; He et al., 2021).

Dans le chapitre 5, nous appronfondissons les aspects théoriques autour des ADs, plus précisément, nous détaillons comment les ADs ont vu le jour avec les travaux de (Austin,1962) et ont été largement promus par (Searle, 1975).

3.3 Appariement d'Énoncés (AE)

Comme mentionné à l'introduction (cf. section 3.1), l'Appariement d'Énoncés (AE) consiste à mettre ensemble deux segments de texte (souvent des phrases) qui partagent une certaine similarité soit sémantique ou dialogique ou bien les deux à la fois. Il n'existe pas à notre connaissance de travaux

qui traitent de cette problématique dans le cadre de la constitution de sous-fils de conversations d'emails.

Dans le cadre de la résolution de certaines problématiques comme la segmentation de thématiques dans les dialogues et dans la modélisation de dialogue en général, le terme « **paires adjacentes** » d'après (Schegloff & Sacks 1973), est le terme théorique de base pour décrire l'organisation du discours. (Midgley et al., 2006) pour identifier empiriquement les paires adjacentes significatives dans un dialogue, les considèrent comme des unités atomiques de chaînes de Markov. Ces chaînes sont utilisées en combinaison avec les paires adjacentes par (Boyer et al., 2009) pour modéliser la structure des dialogues. Pour segmenter les thématiques à l'intérieur des dialogues, (Xing and Carenini, 2021) s'appuient sur des scores de cohérence de paires adjacentes. (Krifka, 2022) propose un enrichissement des théories dynamiques de la communication au travers de paires adjacentes et de certaines fonctions communicatives. Il définit les deux éléments constitutifs d'une paire adjacente comme suit :

— la *première partie de la paire (PPP)* émise par un interlocuteur nécessite une réponse spécifique de la part d'un ou d'autres interlocuteurs

— La réponse spécifique en question est la *seconde partie de la paire (SPP)*.

Lorsqu'on s'appuie sur les ADs d'énoncés de dialogue, on peut avoir des exemples de paires adjacentes suivantes : le *salut-réponse au salut*, la *question-réponse*, la *demande-accord (ou refus)*, la *proposition-acceptation (ou déclinaison)*, l'*assertion-confirmation (ou rejet)* et bien d'autres en fonction de la richesse d'un dialogue.

La notion de « **paires adjacentes** » est généralement utilisée dans la littérature pour des problématiques autour des dialogues. Identifier de telles paires adjacentes constitue pour nous la tâche d'**Appariement d'Énoncés (AE)**. Lorsqu'il existe une paire adjacente d'énoncés de deux emails A et B avec B qui suit A dans une conversation d'email, alors nous qualifions l'appariement des éléments de la paire de **transverse**. Comme notre problématique porte sur les emails, nous essayons de construire des sous-conversations à partir des segments de texte d'emails et c'est pour cela que nous utilisons très souvent le terme **appariement transversal d'énoncés**.

3.4 Conclusion

Ce chapitre a mis en avant certains travaux connexes à nos sous-problématiques, plus précisément celle de reconnaissance ou d'identification d'actes de dialogue (ADs) et d'appariement d'énoncés ou de segments de texte. Nous avons relevé la principale contrainte de disponibilité de corpus d'emails annotés pouvant répondre à notre besoin. C'est pour cette raison que dans les chapitres 5 et 6 nous avons défini un référentiel d'ADs (cf. section 5.3) avec 3 niveaux de granularités que nous utilisons pour annoter les corpus pour mener à bien nos expériences.

Concernant l'identification des ADs, les travaux de (Wang et al., 2019) mettent en avant la prise en compte de contexte autour des énoncés à classifier pour de meilleurs résultats. Nous avons aussi remarqué la rareté des travaux pour cette problématique sur des corpus d'emails contrairement à d'autres corpus comme ceux extraits des réseaux sociaux.

L'AE quant à lui a souvent été abordé non pas comme une problématique à résoudre, mais comme un levier pour la modélisation des dialogues et la séparation de thématiques dans ceux-ci. Et ceci au travers de la notion de « paires adjacentes » dans le de communication synchrone.

Dans le contexte de conversations d'emails, nous n'avons pas identifié de travaux qui approchent l'appariement transversal d'énoncés ou segments de texte d'emails comme nous le décrivons dans nos expériences (cf. chapitre 8) pour notre problématique de constitution de sous-fils de conversations d'emails.

Deuxième partie

Modalités de communication et analyse discursive

4 Communication médiée par ordinateur - CMO

4.1 Introduction

Au début des années de 1980, le concept de Communication Médiée par Ordinateur (CMO) a vu le jour dans le milieu universitaire, désignant l'ensemble des modalités de communication qui s'effectuent via des machines. L'évolution de cette notion a poussé à reconsidérer l'ordinateur comme un médium de communication plutôt que comme un outil. De nos jours, faire allusion à ce concept revient à s'intéresser à un domaine très vaste et complexe, parce qu'il inclut non seulement les interactions homme machine dotées de nouvelles technologies, mais aussi des multiples aspects langagiers utilisés lors des échanges. Que ces derniers soient oraux ou écrits, synchrones ou asynchrones, ils présentent tous des avantages et des inconvénients.

4.2 Avantages et inconvénients

Il existe de plus en plus d'applications et d'outils autour des CMO que nous abordons de façon plus explicite dans les deux prochaines sections. Ces outils ou applications présentent des avantages, on a entre autres le fait de rester en contact avec des proches peu importe leur position géographique dans le monde entier, le partage d'informations. Ces outils facilitent aussi l'accès aux informations et la communication dans les entreprises, favorisant ainsi un gain de temps pour leurs collaborateurs. De nos jours, de nombreux professionnels utilisent ces outils dans le cadre de leurs activités pour une meilleure efficacité. Cependant ces outils sont très souvent cause de surcharge informationnelle et de perte de temps, malgré qu'ils aient montré leur grande puissance de maintien de relations entre les personnes, pour leurs activités professionnelles et personnelles. Sur le plan éducationnel et de l'apprentissage, les CMO, au travers des plateformes d'e-learning ou des moocs, permettent de perpétuer les enseignements dans de multiples domaines scientifiques ou littéraires. Un autre avantage est la production des données qui aident énormément dans plusieurs domaines et plus précisément celui de la recherche scientifique.

Cependant ces CMO ont aussi des inconvénients. On peut noter dans le cadre du partage d'informations la propagation des fake news, le piratage d'informations et les demandes de rançon par des techniques allant du simple hameçonnage aux cyberattaques. On a aussi enregistré sur les réseaux sociaux ces dernières années une augmentation du taux des harcèlements en tout genre et même des groupes de radicalisation. Les travailleurs de la connaissance utilisent les CMO dans le cadre de leurs activités pour une meilleure efficacité, mais doivent faire face à la surcharge informationnelle et la perte de temps induits par les CMO. Un autre aspect négatif de certains de ces outils de CMO est l'addiction distractive de personnes impactant ainsi leur productivité au travail par exemple,

sans compter les troubles de concentration chez les plus jeunes. Au vu de la croissance de ces inconvénients, des mesures de sécurité et même des cellules de lutte contre les cyberattaques ont été mises sur pied au niveau des gouvernements et même sur le plan international. Des protocoles sont mis en place dans le cadre de la protection des données privées d'utilisateurs ou d'organismes, c'est par exemple le cas du **RGPD** (Règlement Général sur la Protection des Données). Pour les enfants et adolescents, des contrôles parentaux et les systèmes de limitation de temps d'utilisation sont de plus en plus incorporés dans les téléphones et tablettes afin de limiter l'utilisation de certaines applications ou outils de CMO. Plusieurs applications ou outils qui permettent la production des CMO sous forme orale ou écrite, synchrone ou asynchrone.

4.3 Types de communication

4.3.1 Conversations Orales, Écrites

De nos jours il existe trois principales façons de communiquer : via la parole, les écrits ou les signes. Nous nous intéressons ici aux deux premières parce qu'elles sont les plus mises en avant dans les outils de CMO. L'oral est un discours en interaction exprimé et transmis de vive voix et véhiculé par la parole. Par contre l'écrit est consigné à travers des graphies présentes dans le temps et dans l'espace. Ces deux moyens de communication présentent des caractéristiques langagières et linguistiques qui les distinguent fortement. Toutefois, la présence physique, la distance et le temps sont les principales contraintes expliquant ces différences. L'oral possède des spécificités qui le différencient clairement de l'écrit. On distingue entre autres la spontanéité, la prosodie (pauses, accents d'instance, intonation, débit, etc.), les liaisons, les enchaînements vocaliques, les fréquences de signaux de régularisation, et les disfluences (hésitations, amorces, constructions interrompues, etc.). Ces différents traits permettent de doter les discours oraux de fonctions précises, certains (disfluences) impactant fortement la syntaxe de l'oral lui donnant notamment un aspect disloqué.

Au-delà des aspects que nous venons d'évoquer, notons que des linguistes émérites ont travaillé sur la question de différences et ou de rapprochement entre les communications écrites et parlées. C'est ainsi que (Halliday and University., 1985; McCarthy, 1991; Cook, 2014) et bien d'autres ont tenté de mettre en place des caractéristiques du langage écrit comme une thématique séparée du langage parlé. Cependant, ils sont tous arrivés à la conclusion selon laquelle ces deux types de langages sont interdépendants du fait que le langage parlé nécessite l'écrit et vice-versa. Par le passé, certains linguistes (Bloomfield, 1933) ont suggéré que l'écrit était juste une façon d'enregistrer les paroles, et De Saussure (de Saussure, [1916] 1983) définissait l'objectif de l'écrit comme une représentation du langage oral.

Des opinions récentes stipulent que les communications écrites n'ont jamais été et ne seront jamais une façon de poser ou de transcrire la parole. Ceci peut très bien se voir sur des annotations de conversations parlées qui sont très distantes d'un texte écrit (ouvrages, rapport, romans, etc.), de par l'ensemble des éléments linguistiques utilisés pour ces annotations. De même l'invention du magnétophone a permis de mieux cerner ce qu'était finalement un enregistrement de paroles. Et donc retranscrire de la parole serait à des fins d'analyses de discours ou de conversations parlés.

En complément des paragraphes précédents, il existe d'autres aspects de différenciation de ces modalités de communication. Ce sont ceux de leur typage fonctionnel et de leurs caractéristiques linguistiques.

Concernant les caractéristiques linguistiques, la différence entre la parole et l'écrit peut se faire sur plusieurs plans ci-dessous :

— Prosodie (ou traits suprasegmentaux) et la ponctuation : la segmentation de la parole s'identifie premièrement par les critères paralinguistiques tels que les pauses, l'intonation et le rythme de la parole. L'écrit quant à lui utilise les ponctuations et les paragraphes.

— Compréhension et incompréhension : en général des interlocuteurs impliqués dans un échange partagent un grand champ contextuel parce qu'ils ont l'opportunité d'avoir des retours immédiats, ce qui permet d'éviter des incompréhensions. Cependant le langage conversationnel est souvent peu explicite et vague, caractérisé par une haute fréquence d'expressions déictiques et de pronoms, mais aussi d'une basse fréquence d'utilisation des noms. Par contre l'écrit ne peut s'appuyer que sur une contextualisation parce qu'il n'y a pas de possibilités d'avoir des feedbacks intermédiaires. Ainsi l'écrit tend à être précis et explicite et est caractérisé par une moindre utilisation d'expressions déictiques.

— Densité lexicale : sur cet aspect, l'écrit est largement plus dense que la parole.

— Interactivité : de par la fonction prédominante phatique de la parole et peu représentée dans l'écrit, il ressort très logiquement que l'interactivité est beaucoup plus marquée lors de l'utilisation de la parole.

— Informel : en tenant compte des traits externes de la parole et de l'écrit, il n'y a point de doute que l'écrit reste très formel au niveau de l'élaboration des structures syntaxiques afin de produire des phrases complexes. Ce qui n'est pas le cas de la parole qui est utilisée par des interlocuteurs au travers des phrases courtes et simples et parfois peu structurées syntaxiquement. L'élaboration d'une structure syntaxique lorsqu'on parle aboutit quelques fois à des incompréhensions. La parole est fortement caractérisée par des répétitions prépondérantes, des mots d'argots et des contractions, ce qui montre bien le caractère informel de la parole.

Ces précédents points présentent les différences entre les médiums de communication écrits et parlés sous plusieurs aspects linguistiques et paralinguistiques. Ces caractéristiques ici présentées ont pour sources les travaux de (Pardubová, 2006). Au-delà de la dimension écrite ou orale des communications, celles-ci peuvent aussi être synchrones et asynchrones.

4.3.2 Conversations synchrones, asynchrones

Les téléphones ou ordinateurs possèdent une pléthore d'applications permettant de communiquer entre particuliers ou entre collaborateurs en entreprise et ceci de manière synchrone ou asynchrone.

La communication synchrone est une communication temps réel entre deux ou plusieurs interlocuteurs, les échanges y sont directs et spontanés. La conversation en face-à-face, les réunions, appels téléphoniques, visioconférences ou messageries instantanées sont des exemples de communications synchrones. Ce type de communication a l'avantage d'être très riche du fait des informations non verbales qui leur sont inhérentes. Ces conversations synchrones ont un avantage sur le plan social parce qu'elles permettent la transmission directe des émotions qui peuvent être perceptibles même

sur une conversation téléphonique. L'abondance d'interruptions, le stress et le manque d'assurance souvent induits par l'influence d'autres interlocuteurs sont les inconvénients qu'on peut retrouver lors des conversations synchrones.

La communication asynchrone quant à elle est une communication qui se déroule en différé, les contraintes spatiales ou temporelles sont alors inexistantes. Les locuteurs décident eux-mêmes quand et où échanger. La majorité des outils de communications professionnelles, emails, Skype, Slack, Workplace... permettent la communication asynchrone caractérisée par la présence d'un délai plus ou moins long entre le moment où l'information a été émise et le moment où elle est reçue et le temps de réponse à celle-ci. Toutefois il existe certains canaux de communications qui sont considérés comme des communications à la fois synchrones et asynchrones. Prenons l'exemple des SMS ou de la messagerie instantanée; lors de la réception d'un message via ces canaux, on peut répondre sur le champ si on est disponible pour cela ou le faire plus tard sinon. Les avantages liés à la communication asynchrone sont nombreux. On distingue, entre autres, une plus grande liberté pour mieux exprimer ses idées, le contrôle de l'information reçue, le respect des autres interlocuteurs et aussi de leur emploi de temps. Cependant, dans ces communications asynchrones il est difficile de pouvoir détecter l'état émotionnel réel d'un interlocuteur lorsque celui-ci a rédigé un message dans la conversation. Un autre inconvénient est le retard entre, par exemple, un email et sa réponse qui peut souvent durer plusieurs heures, jours ou semaines parce que noyé dans une boîte de messagerie avec plusieurs autres emails reçus par son destinataire.

Certains outils de communication comme les SMS, les applications de conversations instantanées (e.g. Slack, Skype), les forums en fonction de la spontanéité ou pas de réaction des interlocuteurs vont produire des communications synchrones ou asynchrones. Dans les forums, du fait que plusieurs interlocuteurs peuvent prendre part à une ou plusieurs conversations, de multiples sujets sont ainsi abordés et plus tard, il sera difficile, même pour un interlocuteur ayant participé à ces conversations, de retrouver une information explicite dont il est conscient de son existence dans lesdites conversations. Bien que les labels, les titres de conversations et d'autres métadonnés de forums peuvent faciliter la recherche d'informations. On retrouve aussi cette difficulté de recherche d'information dans les conversations d'emails en entreprise parce que, malgré les objets plus ou moins ciblés des emails, les interlocuteurs vont souvent aborder de nouvelles thématiques.

Ce problème de recherche d'information dans les forums ou les emails poussent à des problématiques spécifiques autour des conversations. Le démêlage de conversation, la constitution de sous-fils de conversations et l'identification des intentions des interlocuteurs sont des exemples de ces problématiques. Ces dernières, une fois résolues faciliteraient l'accès à l'essence des informations contenues dans ces conversations extraites d'emails ou de forums, riches en connaissances et souvent très peu structurées. Dans les prochaines sections, nous allons présenter les caractéristiques des emails parce qu'ils constituent le principal médium de communication étudié dans nos travaux. Dans ces travaux, nous utilisons des corpus de forums disponibles et annotés pour approcher des sous-problématiques (comme la classification d'énoncés en acte de dialogues) connexes à notre problème de constitution de sous-fils de conversations. C'est pour cette raison que les caractéristiques des forums sont aussi décrites dans l'une des sections suivantes.

4.4 Caractéristiques d'emails

Apparu au Québec dans les années 90, le mot « email » est dérivé du mot anglais « mail » (« courrier » en français), avec le préfixe e- (contraction de « électronique »). En français le synonyme « courriel » (contraction de « courrier électronique ») a été accepté par l'Académie française et rendu obligatoire en France pour les textes officiels depuis le 20 juin 2003. Le courriel ou email est donc un courrier électronique destiné à un ou plusieurs tiers, pouvant contenir des messages de différentes natures (travail, publicité, loisirs...) qui transite par le biais d'une connexion dans un réseau informatique pour enfin terminer son parcours dans une boîte de messagerie électronique où il sera consulté.

D'après le rapport Radicati Group de novembre 2022, Le nombre total d'emails envoyés et reçus par jour par les entreprises et le grand public a dépassé 333 milliards en 2022 et devrait atteindre plus de 392 milliards d'ici la fin de l'année 2026. Ce même rapport indique que plus de la moitié de la population mondiale (soit environ 4,2 milliards) utilise les emails et ce nombre devrait atteindre plus de 4,7 milliards d'ici la fin de 2026. D'après Médiamétrie (janvier 2019), on a dénombré en 2019 42,2 millions de Français se connectant chaque mois à un webmail et 22,7 millions se connectant à au moins un compte mail chaque jour. Pour Map global Provider Report (octobre 2019), les principales messageries utilisées en France sont, par ordre décroissant : Gmail (27 %), Outlook.com / Hotmail (26%), Orange (18%), Yahoo mail (12%) et SFR (6%). Dans le monde de l'entreprise, d'après une étude d'Adobe en août 2015, les cadres estiment passer plus de 5 heures par jour en moyenne à consulter leur messagerie ; en France leur estimation était autour de 5,6 heures et 5,4 heures en Europe. Aux Etats-Unis ce chiffre monte à 6,3 heures. Au vu de ces chiffres, on peut en effet s'imaginer la grande quantité d'informations produite par les emails et aussi l'urgence d'optimiser ces temps de consultation de messagerie par des cadres en facilitant la recherche d'informations comme mentionné plus haut. Cependant pour atteindre un tel objectif et bien d'autres, il faut s'intéresser aux différentes caractéristiques des courriels, tant sur les aspects linguistiques, paralinguistiques, lexicaux et syntaxiques que structurels.

4.4.1 Structure d'un email

La structure d'un email suit généralement un format standard qui comprend 3 principales zones :

1. **L'en-tête d'un email** constituée de :
    - L'en-tête ou le champ « De » : c'est le champ où vous écrivez votre adresse email. Cette adresse apparaîtra comme l'expéditeur de l'email.
    - Le champ « À » : c'est le champ où vous écrivez l'adresse email du ou des destinataire(s) principal(aux) de l'email.
    - Le champ « Cc » (Copie carbone) : ce champ permet d'envoyer une copie de l'email à une ou plusieurs personnes en plus du destinataire principal.
    - Le champ « Cci » (Copie carbone invisible) : ce champ permet d'envoyer une copie de l'email à une ou plusieurs personnes sans que le destinataire principal ne puisse les voir.

— L'objet : c'est l'en-tête de l'email qui décrit brièvement le contenu du message. L'objet d'un email est souvent précédé des mots expressions *Re :* ou *Tr :* suivant que l'email est une réponse ou un tranfert.

2. Le **corps de l'email** : c'est le contenu principal de l'email. Il peut inclure du texte, des images, des liens, des pièces jointes, etc. Dans cette partie de l'email on retrouve généralement l'historique d'une conversation sous forme de messages cités séparés par des expressions spécifiant la date, l'heure, l'émetteur de chacun de ces messages et les destinataires. Ces messages cités pour chaque précédent email dans la conversation se distinguent souvent par des empilements du signe supérieur (>) et donc l'email le plus ancien dans la conversation en aura beaucoup que le plus récent qui en aura qu'un seul. Ce signe diffère aussi en fonction des serveurs de messageries.
3. le **pied de page** de l'email : principalement constitué de la signature qui comprend généralement les coordonnées de l'expéditeur, telles que son nom, sa société, le nom son équipe, son numéro de téléphone et son adresse email, mais aussi parfois des liens vers des réseaux sociaux. La signature est souvent suivie d'une clause de non-responsabilité légale dans des emails professionnels.

La formulation d'un email peut varier en fonction de la situation et du contexte. Par exemple, dans les emails professionnels , il est respectivement courant d'être plus formel, à contrario dans des emails personnels, on utilise des formulations moins formelles et plus familières.

Il est également important de noter que l'utilisation de certains éléments, tels que les champs « Cc » et « Cci », peut varier en fonction de l'objectif et de la relation entretenue avec le destinataire de l'email. Par exemple, dans un email professionnel envoyé à un supérieur hiérarchique, il est courant de ne pas inclure d'autres personnes en copie, sauf si cela est nécessaire.

En résumé, la structure d'un email comprend généralement les éléments suivants : l'en-tête, le champ « À », les champs « Cc » et « Cci », l'objet, le corps de l'email et la signature. La structure peut varier en fonction de la situation et du contexte. En plus de leur structure, les emails possèdent aussi de nombreuses caractéristiques.

4.4.2 Caractéristiques d'emails

Les caractéristiques d'emails peuvent être détaillées sous les aspects linguistiques, paralinguistiques et lexicaux.

◇ **Caractéristiques linguistiques** :

— **Niveau de langue** : le niveau de langue utilisé dans un email peut varier en fonction de la relation entre l'expéditeur et le destinataire, ainsi que du contexte de l'email. Dans un contexte professionnel, un niveau de langue plus formel est utilisé, tandis que dans un contexte personnel, un niveau de langue plus familier peut être utilisé.

— **Ton** : le ton d'un email également varie en fonction du contexte et de la relation entre l'expéditeur et le destinataire. Le ton peut être informel, amical, professionnel, directif, etc.

— **Grammaire** : la grammaire d'un email doit être correcte afin de garantir une compréhension claire et précise du message. Les erreurs grammaticales peuvent affecter négativement la crédibilité de l'expéditeur.

◇ **Caractéristiques paralinguistiques** :

— **Politesse** : l'utilisation de formules de politesse est importante dans un email. Certaines comme « S'il vous plaît » et « Merci » établissent une relation positive entre l'expéditeur et le destinataire.

— **Emoticons** : les emojis, qui sont des symboles ou des images utilisés pour exprimer des émotions, sont utilisés dans un email pour renforcer le ton ou exprimer des émotions.

— **Ponctuation** : la ponctuation est importante car elle peut affecter la compréhension du message. L'utilisation excessive ou inappropriée de la ponctuation peut également affecter la crédibilité de l'expéditeur.

◇ **Caractéristiques lexicales** :

— **Vocabulaire** : le choix du vocabulaire doit être adapté au contexte et au destinataire de l'email. L'utilisation d'un vocabulaire trop complexe peut rendre le message difficile à comprendre, tandis qu'un vocabulaire trop simple donne l'impression d'un manque de professionnalisme.

— **Jargon** : dans un contexte professionnel, l'utilisation de jargon peut être appropriée pour communiquer des concepts spécifiques. Cependant, il est important de s'assurer que le destinataire comprend le jargon utilisé.

Nous constatons que les emails possèdent plusieurs caractéristiques en plus de leurs structures spécifiques, ces richesses d'informations tant structurelles que linguistiques, paralinguistiques et lexicales font des emails une modalité de CMO intéressante à étudier dans le cadre d'analyse discursive.

4.5 Analyse discursive d'emails

L'analyse discursive est une méthode d'analyse qui se concentre sur la façon dont les personnes utilisent la langue pour construire des significations dans des situations sociales spécifiques. En ce qui concerne les emails, l'analyse discursive peut aider à comprendre comment les émetteurs utilisent la langue pour construire des relations avec les destinataires, pour exprimer des émotions, pour négocier des significations, et le plus important surtout en milieu professionnel : pour atteindre des objectifs de communication spécifiques.

L'analyse discursive peut être utilisée pour examiner les choix linguistiques et les motifs discursifs dans les emails. Par exemple, l'analyse discursive peut se concentrer sur les choix de mots, les constructions de phrases, les tons, les formules de politesse, les signes de ponctuation, les emojis, et les stratégies discursives utilisées pour persuader, convaincre ou influencer le destinataire. L'analyse discursive peut également aider à comprendre les normes et les attentes professionnelles associées à la communication par email dans une entreprise. Elle peut aussi s'étendre à plusieurs de ses attributs comme, par exemple, les actes de dialogue afin de permettre une meilleure compréhension des conversations d'emails. Cette dernière façon d'exploiter les emails sous le prisme de l'analyse discursive est celle que nous utilisons dans nos travaux.

4.6 Caractéristiques des forums

Les forums sont un autre type de CMO ; ce sont des espaces en ligne où les utilisateurs peuvent échanger des informations, des idées et des opinions sur des sujets spécifiques. Les utilisateurs peuvent créer des sujets de discussion et répondre aux messages des autres utilisateurs.

Sur le plan linguistique, paralinguistique, lexical et syntaxique, les forums présentent certaines caractéristiques spécifiques :

— **Linguistique** : les forums sont caractérisés par un style d'écriture informel et interactif, qui encourage les échanges entre les utilisateurs. Les messages peuvent inclure des expressions familières, des abréviations et des erreurs d'orthographe, qui reflètent la spontanéité de la communication en ligne. Les échanges peuvent inclure des éléments de persuasion et de débat, car les utilisateurs cherchent souvent à convaincre les autres de leur point de vue.

— **Paralinguistique** : les forums ne permettent pas l'expression de la parole, ce qui signifie que les utilisateurs ne peuvent pas utiliser des indices paralinguistiques tels que l'intonation, les gestes ou les expressions faciales pour communiquer. Cependant, les forums peuvent inclure des éléments de paralangage tels que les émoticônes, les emojis ou les réactions visuelles (par exemple, des « j'aime » ou des "j'adore") pour exprimer les émotions ou les attitudes.

— **Lexical** : les forums peuvent inclure un jargon spécifique ou des termes techniques, liés aux sujets discutés. Ceci est également le cas pour les emails en général, et en particulier les emails professionnels. Les utilisateurs peuvent également créer de nouveaux termes ou des néologismes pour décrire des concepts spécifiques, qui peuvent devenir courants dans la communauté de l'utilisateur. En outre, les forums peuvent être caractérisés par un style d'écriture concis et direct, qui utilise souvent des phrases courtes et simples pour communiquer l'information.

— **Syntaxique** : les forums ne suivent pas toujours de normes grammaticales strictes, ce qui permet aux utilisateurs d'écrire de manière plus informelle et spontanée. Les phrases peuvent être courtes et simples, ou au contraire longues et complexes. Les utilisateurs peuvent également utiliser des structures syntaxiques non conventionnelles, telles que des phrases nominales ou des phrases incomplètes, pour communiquer leur message de manière plus efficace.

En somme, les forums ont leur propre style d'écriture, qui se caractérise par un langage informel, interactif et direct. Les forums peuvent inclure des éléments de persuasion et de débat, ainsi que des termes spécifiques et des néologismes liés aux sujets discutés. Les forums peuvent également présenter des erreurs grammaticales et des structures syntaxiques non conventionnelles, qui reflètent la spontanéité de la communication en ligne.

Les caractéristiques discursives des forums sont similaires à celles des emails, mais avec quelques différences notables.

Les caractéristiques discursives des forums incluent des éléments tels que le sujet du forum, le titre de la discussion, les messages individuels, les réponses et les commentaires, ainsi que les règles et les normes de participation au forum. Les sujets de discussion sont souvent organisés en catégories thématiques pour faciliter la navigation et la recherche de sujets spécifiques. Les titres de discussion sont souvent des résumés concis du sujet de discussion, qui doivent attirer l'attention des utilisateurs et les inciter à lire et à participer à la discussion.

Les messages individuels dans les forums peuvent être plus longs que les emails, car ils sont souvent destinés à fournir des informations détaillées sur un sujet spécifique ou à exprimer une opinion personnelle. Les messages peuvent également inclure des liens vers d'autres ressources en ligne, des images ou des vidéos pour appuyer ou illustrer les arguments de l'auteur. Les réponses et les commentaires sont souvent utilisés pour discuter ou débattre des idées soulevées dans les messages individuels.

Les règles et les normes de participation au forum sont également importantes pour le fonctionnement du forum. Ces règles peuvent inclure des normes de conduite, des règles de modération et des directives sur la façon de participer aux discussions de manière constructive et respectueuse.

En analysant les caractéristiques discursives des forums, on peut examiner comment les utilisateurs utilisent la langue pour discuter de sujets spécifiques, pour persuader les autres utilisateurs, pour exprimer des émotions et des opinions, et pour construire des relations sociales en ligne. On peut également examiner les normes et les attentes culturelles associées à la participation aux forums, ainsi que les impacts sociaux et politiques de ces formes de communication en ligne.

De même, l'identification des actes de dialogues dans les forums permet une meilleure compréhension des conversations des forums et permet ainsi de mieux identifier les différents sujets abordés de telles conversations.

4.7 Conclusion

Dans ce chapitre, les communications médiées par ordinateur (CMO) ont été présentées sur plusieurs niveaux : avantages et inconvénients, synchrones et asynchrones. La parole et l'écrit ont également été analysés sur différents aspects linguistiques. Les caractéristiques linguistiques, syntaxiques et paralinguistiques des emails et forums ont aussi été présentées. Les emails et forums sont un type de CMO écrites, respectivement asynchrones et synchrones, que nous exploitons largement et analysons dans le cadre de nos travaux, notamment sous le prisme d'une analyse discursive pour aborder notre problématique de constitution de sous-conversations. Dans le prochain chapitre, nous nous attardons sur les différentes propriétés de l'analyse discursive, plus précisément sur les actes de dialogue.

5 Analyse discursive

5.1 Introduction

Dans ce chapitre, nous détaillons ce qu'est l'analyse discursive, plus précisément, nous nous intéressons ici aux actes de dialogue qui sont nécessaires à l'étude des phénomènes conversationnels, à la conception d'agents conversationnels, et même à l'annotation de dialogues, de conversations ou de segments de texte de CMO.

Tout d'abord nous allons expliciter la notion de théorie d'actes de dialogue, son évolution au travers des travaux d'Austin et Searle, et montrerons comment ces actes de dialogues sont utilisés ou appliqués dans l'analyse des conversations fonctionnelles, leur rôle dans les contextes conversationnels et leur influence sur les marqueurs pragmatiques. Par la suite, nous présenterons les principaux schémas d'annotations : DAMSL (Dialog Act Markup in Several Layers), DIT++(Dynamic Interpretation Theory) et la norme ISO 24617-2 établie en 2012 et dont la dernière version a été publiée en décembre 2020 (Bunt et al., 2020) . Enfin nous exposerons un référentiel d'actes de dialogue que nous avons défini, qui s'appuie fortement sur la norme ISO 24617-2 et que nous utilisons plus tard dans nos travaux pour annoter nos données.

5.2 Actes de langage ou de discours

5.2.1 Théorie des actes de Langage – Premières taxonomies (Austin et Searle)

Les pratiques linguistiques aujourd'hui déployées dans plusieurs domaines trouvent leur origine dans la théorie des actes de langages ou actes de discours. Cette dernière stipule que le langage a pour fonction essentielle non seulement celle de décrire le monde, mais aussi celle d'accomplir des actions. Cette théorie a été initiée par un philosophe britannique du nom de J.L Austin en 1969 dans son livre bien connu, intitulé « *How to do things with words* ». Austin développe une nouvelle théorie des actes de langage selon laquelle tout énoncé peut être analysé sur trois niveaux :

— Niveau **locutoire** : production d'une suite de sons, évoquant et reliant syntaxiquement les notions représentées par les mots.

— Niveau **illocutoire** : production d'un énoncé porteur d'intention rhétorique du locuteur.

— Niveau **perlocutoire** : l'énonciation vise des effets plus lointains ou bien s'intéresse aux conséquences qui seront produites ou même de son interprétation par les allocutaires.

Ces trois niveaux peuvent respectivement se traduire par les questions, « que dit-il ? », « que fait-il ?» et « pour quoi faire ? ». Le niveau *illocutoire* est l'un des plus importants parce que ses énoncés décrivent des fonctions communicatives (questions, réponses, remerciements...) qui sont

Austin (1969)	Searle (1975)	Quelques verbes
Expositifs (qui exposent de l'information)	Assertifs (qui affirment un état de fait)	affirmer, nier, postuler, remarquer...
Exercitifs (qui exercent un pouvoir)	Directifs (qui poussent l'interlocuteur à agir)	commander, conseiller, ordonner, pardonner, léguer...
Promissifs (qui engagent le locuteur)	Promissifs (qui engagent le locuteur)	promettre, faire vœu de, garantir, jurer de..
Comportatifs (qui expriment l'attitude)	Expressifs (qui expriment un état psychologique)	s'excuser, remercier, féliciter, déplorer, critiquer.
Verdictifs (qui donnent un verdict)	Déclaratifs (qui ont un impact réel)	acquitter, condamner, décréter, baptiser

Table 5.1 – Taxonomies primaires de la théorie d'actes de dialogue

étroitement liées aux actes de discours. Dans ses travaux, Austin propose cinq classes d'actes de discours regroupées comme dans le tableau 5.1.

Ces actes de langage sont aussi étudiés en profondeur par John R. Searle (Searle, 1975). Il défend sa théorie selon laquelle tout acte de discours est illocutoire et cette théorie rejoint celle de (Austin, 1975) sur ce que ce dernier appelle les actes de langage. Cette notion a été explicitée par (Bunt, 1989, 1995) qui attribue trois aspects aux énoncés qui portent ces actes de langage. Ces attributs sont la forme de l'énoncé, sa fonction communicative et son contenu sémantique. Searle aussi mettait en avant la notion de fonction communicative des énoncés qui pour lui est essentielle. Il a ainsi, tout comme Austin, proposé cinq classes d'actes de langage (tableau 5.1).

Cette théorie d'actes de langage, au fil des années, a été utilisée dans plusieurs domaines tels que la philosophie, la littérature et dans son domaine de base la linguistique où elle s'est d'autant développée pour donner place à la pragmatique cognitive issue de la théorie de la pertinence. En informatique, la modélisation des conversations et des dialogues, les applications de dialogue homme-machine trouvent leur essence dans l'utilisation des actes de discours. Dans le cadre des analyses de conversations entre plusieurs participants, les actes de langage jouent un rôle prépondérant.

5.2.2 Actes de dialogues

Dans cette partie nous abordons le passage des actes de langage aux actes de dialogue et comment ces derniers contribuent à l'analyse des conversations. Le contexte conversationnel sera aussi présenté avec l'influence sémantique de surface et sous-jacente qu'il peut permettre de percevoir sur des énoncés.

A) **Analyse des conversations fonctionnelles**

Les travaux d'analyse de (Vanderveken, 1992) sur la limite des actes de discours largement répandue dans les précédentes années ont montré que les actes de discours permettaient seulement de faire des études sur des énoncés de façon isolée ou indépendante de l'existence des autres énoncés lors d'échanges entre différents locuteurs, par exemple. Cette limitation est d'autant plus marquée que, pour Vanderveken, elle a un impact sur l'atteinte d'un objectif commun par des participants à une conversation qui produisent néanmoins des énoncés illocutoires ayant des fonctions communicatives spécifiques. Cet objectif commun pouvant être l'accomplissement d'une mission ou la résolution d'un problème. Ceci montre davantage l'impact des actes de discours l'analyse de conversations.

Les conversations fonctionnelles sont un type de conversations précis dont les échanges contribuent à la réalisation d'un objectif spécifique. On peut ainsi affirmer que les conversations d'emails appartiennent à ce type de conversation parce qu'en général les messages échangés dans les emails

ont un but bien précis qui peut être l'obtention d'une information, la résolution d'un problème, ou en entreprise la mise en place d'un processus. L'extension des actes de discours avec la prise en compte de dépendances entre les énoncés a été faite par (Traum and Hinkelman, 1992) donnant lieu aux actes de dialogue ou actes de conversation. Ces derniers sont très utilisés dans la littérature pour modéliser les interactions dans les CMO entre les humains, mais aussi entre les humains et les machines. Dans ces modélisations, l'un des aspects importants est le contexte qui contribue à différencier les actes de discours des actes de dialogues. Dans la prochaine sous-section, le contexte sera décrit sous l'angle de l'interprétation des conversations.

B) **Prise en compte de contexte**

La théorie du contexte autour des conversations s'est davantage développée à partir des observations faites par (Poesio and Traum, 1997) pour qui les conversations fonctionnelles possèdent des propriétés qui vont au-delà de l'atteinte d'un objectif. Pour eux déterminer une action coordonnée enfouie dans les énoncés est non négligeable. D'où la prise en compte du contexte principalement utilisé pour représenter les effets des actes de discours sur les participants d'une conversation. Entre autres de ces effets, les besoins, les obligations et les croyances de ces interlocuteurs qui sont induits par des compréhensions pouvant varier d'un participant à un autre bien qu'étant dans la même conversation. Pour Poesio et Traum, le contexte représente l'information sur laquelle les participants s'appuient pour interpréter les énoncées d'une conversation. Vu la polysémie que peuvent porter des énoncés d'une conversation, seul le contexte est le vecteur principal qui permet ainsi aux participants de converger vers une compréhension monosémique d'énoncés. Il est ainsi considéré comme un ensemble de connaissances communes partagé par les participants à une conversation. Ainsi donc, pour bien modéliser une conversation, il faut à tout niveau d'évolution de la conversation mettre le contexte à jour. Et donc, pendant que les actes de discours s'attardent à capturer l'intention communicative du locuteur, les actes de dialogues vont plus loin à travers la prise en compte du contexte. En plus de la prise en compte du contexte dans une conversation, les marqueurs pragmatiques sont d'autres caractéristiques importantes dans l'analyse des conversations.

C) **Marqueurs pragmatiques multimodaux**

La pragmatique est un domaine vaste des sciences du langage qui s'intéresse aux éléments langagiers dont la signification ne peut être comprise qu'en connaissance du contexte de leur emploi. Les marqueurs pragmatiques sont une généralisation des marqueurs discursifs qu'on retrouve dans les conversations orales ou écrites. Les marqueurs pragmatiques se divisent en deux grands groupes : les marqueurs pragmatiques verbaux et les marqueurs pragmatiques non-verbaux. Les marqueurs pragmatiques, tout comme les actes de dialogues, incorporent le contexte, ce qui n'est pas le cas avec les actes de discours. Les propriétés pragmatiques d'un énoncé peuvent s'identifier sous deux aspects : les actes expressifs (satisfaction, félicitation, remerciement, mécontentement, consternation, indignation, surprise) et les actes illocutoires (admettre une adéquation ou une inadéquation à un propos, une action, un comportement, approuver ou désapprouver ces derniers). Dans le cadre de l'analyse des CMO, la problématique d'identification de pragmatique sous l'angle d'actes expressifs pourrait permettre d'avoir les informations émotionnelles des participants à une conversation lors de leurs interventions. Partant ainsi d'un glossaire de marqueurs pragmatiques verbaux et non

verbaux, cette identification pourrait se mener facilement, mais reste la contrainte selon laquelle les pragmatiques sont fonction de leur contexte d'utilisation. Il en découle une fois de plus que la modélisation de contexte reste un verrou pour une bonne modélisation de conversation. Vu que les marqueurs pragmatiques dépendent du contexte et que celui-ci a une grande incidence sur la détermination des actes de dialogue selon (Poesio and Traum, 1997), on peut facilement inférer par transitivité que les marqueurs pragmatiques font partie intégrante des actes de dialogues (ADs) et peuvent ainsi être déduits de ceux-ci.

Les actes de discours ou de langage ont ainsi évolué depuis les travaux d'Austin et de Searle pour devenir les actes de dialogue (ADs) au travers de la prise en compte de contexte qui doit être partagée par tous les participants à une conversation. Les ADs permettraient ainsi une bonne analyse et compréhension d'une conversation, c'est pour cela qu'ils sont largement utilisés dans les travaux autour de la modélisation des dialogues/conversations écrits ou parlés (Stolcke et al., 2000; Colombetti et al., 2000; Cohen, 1996a; Traum and Hinkelman, 1992). Nous avons pu établir le fait que les conversations d'emails ou de forums, puisqu'elles sont initiées pour des objectifs bien spécifiques, sont des conversations fonctionnelles qui possèdent des contextes respectifs et, ainsi, leur analyse ou modélisation dans le cadre de la résolution d'une problématique générale, par exemple la constitution de sous-fils de conversations, passerait forcément par l'identification des ADs inhérents aux différents énoncés desdites conversations. L'utilisation des ADs est devenue depuis quelques décennies incontournable dans les études langagières et l'interprétation des discours écrits ou oraux, des conversations synchrones ou asynchrones. Différents schémas d'annotation pour les ADs ont ainsi vu le jour et se sont progressivement améliorés pour une meilleure compréhension des différents énoncés qu'on peut rencontrer des CMO afin de mieux modéliser et automatiser certaines tâches autour de ceux-ci. Dans les prochaines sections, nous présentons quelques-uns de ces schémas d'annotations.

5.2.3 Schémas d'annotation

Les schémas d'annotation sont utilisés pour représenter ou modéliser les conversations en utilisant les ADs. **DAMSL**, **DIT++** et la **norme ISO 24647-2** font partie des normes de schémas d'annotation largement utilisés. Ils ont tous été développés sur le même corpus TRAINS (Heeman and Allen, 1995) qui est un corpus de dialogues de résolution de problèmes. Dans ces schémas d'annotation, deux idées semblent prédominer dans la littérature : (1) la notion de « dimensions » (les dimensions correspondent à différents types d'informations) et (2), une dimension est formée par un ensemble d'étiquettes mutuellement exclusives. Dans DAMSL, par exemple, les termes « dimension » et « couche » sont parfois utilisés dans le sens de (1) et parfois dans celui de (2). Avant de détailler chacun de ces schémas d'annotation, nous allons présenter les principes qu'ils ont en commun :

— La **mise à jour du contexte** : les trois schémas d'annotation prennent en compte l'évolution du contexte partagé par les participants d'une conversation. Et pour un meilleur encodage dudit contexte, les schémas d'annotation prennent en compte les fonctions communicatives, qui selon (Core and Allen, 1997b) représentent de façon directe le contexte évolutif d'une conversation.

— La **multi-dimensionnalité** : la principale limite identifiée dans les taxonomies d'Austin et Searle est la prise en compte du fait qu'un même énoncé peut exprimer plus d'une intention d'un locuteur. Par exemple, dans une conversation un locuteur peut détailler son propos dans un énoncé en posant une question à un autre allocutaire. C'est la muti-dimensionnalité d'un énoncé. Dans ses travaux (Bunt, 2007) prend en compte cette problématique et permet par les schémas d'annotation qu'il propose d'annoter des énoncés sur plusieurs couches (layers).

— La **généricité** : ces schémas permettent en fait d'annoter les énoncés de conversations en ADs suivant deux approches dont une basée sur une taxonomie propre à un domaine ou à une tâche bien précise, la seconde pour une couverture plus large de dialogue (Leech and Weisser, 2003); d'où la notion de généricité.

Leur trait distinctif générique est l'un de leurs atouts les plus importants et probablement celui qui a le plus contribué à leur renommée. Les annotations suggérées sont toutes de haut niveau, ce qui les rend applicables à divers types de dialogues.

A) **DAMSL**

Dialogue Act Markup in Several Layers (DAMSL), en français balisage d'ADs en plusieurs couches, a été le premier schéma d'annotation permettant d'assigner de multiples labels à des énoncés de conversations. Développé par (Core and Allen, 1997b), DAMSL est constitué de quatre principales couches et de plusieurs dimensions comme le montre la figure 5.A ci-dessous. Les super-couches de DAMSL sont les fonctions prospectives, les fonctions rétrospectives, le niveau d'information et le statut communicatif. Les deux premières permettent d'annoter les énoncés en fonction de leur intention communicative. Les deux autres indiquent sur quoi portent les énoncés; ce sont des catégories du niveau informationnel. Les dimensions sur ladite figure correspondent aux premiers niveaux en dessous des super-couches, elles sont indépendantes les unes des autres et possèdent pour certaines des sous-dimensions qui expriment des opinions d'acceptation, de rejet ou de neutralité suivant que l'on soit dans un processus de compréhension ou de validation d'une idée, d'un projet ou bien d'autre chose.

Certains travaux (Jurafsky et al., 1997; Core and Allen, 1997a; Yu and Yu, 2019) sur l'analyse des conversations avec annotations en ADs, et même de constitution de corpus de conversations annotées, s'appuient sur la taxonomie DAMSL. Dans ses travaux (Bunt, 2006) porte des critiques sur les éléments de DAMSL qui, selon lui, manquent de signification conceptuelle et ne s'appuient pas sur des fondements théoriques. Par exemple, il argue que les dimensions DAMSL telles que la « demande d'information, la déclaration et la réponse » ne sont pas considérées comme des dimensions appropriées, et que les fonctions communicatives dans ces catégories ne relèvent d'aucune dimension spécifique, mais doivent être considérées pour des « usages généraux » dans le sens où elles peuvent être utilisées dans n'importe quelle dimension. C'est ainsi qu'il propose le schéma *DIT++* (Bunt et al., 2009).

B) **DIT++**

Cette taxonomie est une extension de la taxonomie **DIT (Dynamic Interpretation Theory)** (Bunt, 1989, 1995) prenant en compte des concepts de DAMSL qui ont été améliorés et d'autres

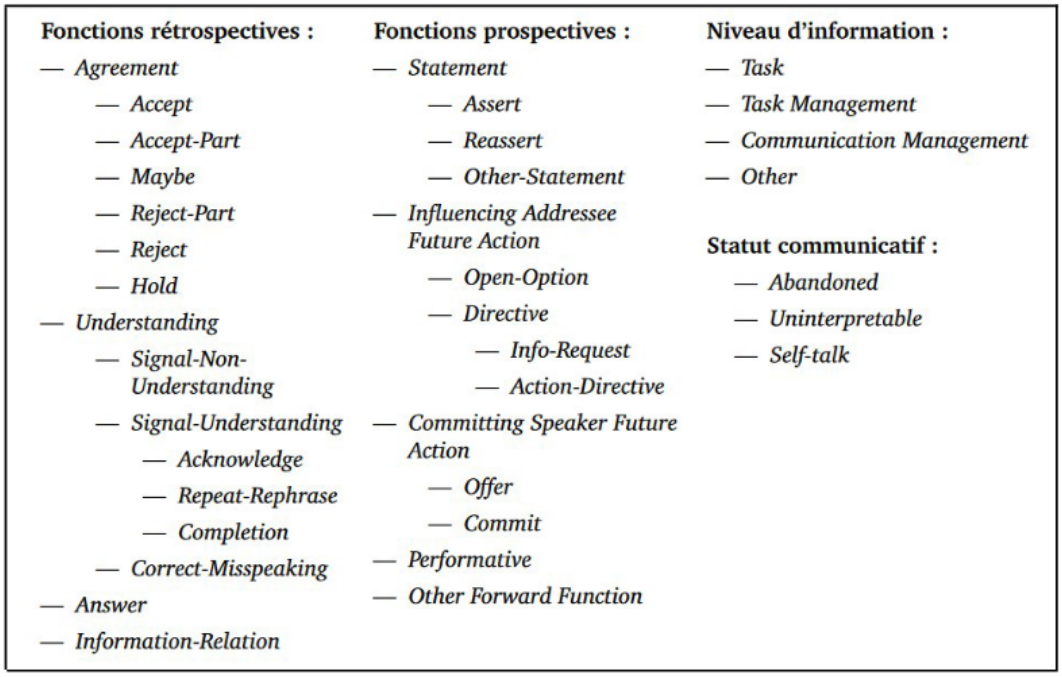


FIGURE 5.A – Taxonomie DAMSL

études sur les ADs. *DIT++* est un framework sémantique développé pour l'analyse des conversations entre humains, mais aussi entre hommes et machines. Il permet aussi d'annoter en ADs des segments de texte contenant des fonctions communicatives. Cette taxonomie est constituée :

1. d'une taxonomie compréhensible multidimensionnelle de fonctions communicatives sémantiquement définies sur plusieurs états de changements d'informations.
2. d'une définition de 10 dimensions orthogonales auxquelles appartiennent les ADs; ces dimensions offrent une base de compréhension de la multifonctionnalité des énoncés de dialogue.
3. d'une définition de plusieurs types de relations sémantiques et pragmatiques entre les ADs.
4. d'un petit ensemble de qualificatifs permettant d'indiquer l'incertitude, la réserve ou les sentiments d'un interlocuteur.

La taxonomie *DIT++* a beaucoup évolué jusqu'en 2009 (Bunt, 2009). Cette évolution a donné lieu à des caractéristiques plus solides tant sur ses dimensions que sur le plan de ses différentes fonctions communicatives. Ces caractéristiques sont présentées dans les deux figures 5.B.

La structure de la taxonomie *DIT++* est basée sur une hiérarchie de catégories ou de couches qui sont définies en fonction de leur contenu sémantique. Les catégories sont organisées en plusieurs niveaux, de la catégorie la plus générale à la plus spécifique. Ces catégories sont des fonctions communicatives. Entre autres, on a les fonctions de transfert d'information, les fonctions de discussion autour des actions inhérentes aux conversations.

Dans la première catégorie, on distingue deux sous-couches ou dimensions qui sont, d'une part, les fonctions de demande d'information et, d'autre part, celles de partage d'informations. Les

demandes d'information sont en général formulées sous forme de questions directes ou indirectes, mais elles ont un objectif bien défini qui est d'obtenir une réponse qui peut être un simple partage d'information, une acceptation ou approbation, un refus ou désaccord, une explication ou élaboration, des réponses sous forme de propositions, etc.

Les fonctions commissives et directives constituent la première couche autour des actions. Les offres, promesses et des prises en compte de requêtes d'actions sont les principaux ADs de type commissif, ceux-ci peuvent aussi s'exprimer avec l'utilisation des verbes performatifs. Les directives, quant à elles, sont, comme leur nom l'indique, des instructions, des requêtes indirectes, des suggestions, des recommandations et même des ordres déguisés avec des formules de politesse.

Au-delà de ces éléments qui sont d'ordre général et qui constituent la structure de *DIT++*, il existe 10 fonctions (détaillées en sous-figure [b] dans l'image 5.B) de communication spécifique liées à des domaines spécifiques dans une conversation. Ces domaines portent, par exemple, autour de la gestion d'activités, de tâches, de feed-back par les allocutaires, de gestion de tours de paroles, de temps, d'obligations sociales, etc.

Au travers de ses différentes fonctions communicatives que nous venons de brièvement présenter, la taxonomie *DIT++* permettrait d'encoder en ADs les énoncées de différents types d'interactions qui peuvent émerger de conversations orales ou écrites comme les forums et les conversations d'emails. Malgré son large champ de couverture de types d'énoncés, il peut arriver que certains énoncés dans les conversations ne trouvent pas d'AD qui leur corresponde exactement en fonction des besoins précis de modélisation de dialogue. C'est par exemple le cas d'un énoncé qui est une hypothèse, qui est annoté comme une suggestion ou partage d'information. Ce type de limite peut ainsi conduire à des erreurs de modélisation de conversations ou de dialogues ou bien de conception d'outils automatiques comme les chatbots.

En 2010, (Bunt et al., 2012) effectuent des travaux de mise en place d'une norme (ISO 24617-2) qui s'appuie sur la taxonomie *DIT++* pour l'annotation d'énoncés en ADs.

- *Information Transfer Functions*
  - *Information-Seeking Functions*
    - *Direct Questions*
      - propositional question, set question, alternatives question, check question, etc.
    - *Indirect Questions*
      - indirect propositional question, set question, alternatives question, check question, etc.
  - *Information-Providing Functions:*
    - *Informing Functions:*
      - inform, agreement, disagreement, correction;
      - *Informs with Rhetorical or Attitudinal Functions, such as* elaboration, justification, exemplification.. and warning, threat,..
    - *Answer Functions:*
      - propositional answer, set answer, confirmation, disconfirmation
- *Action Discussion Functions*
    - *Commissives*
      - offer, promise, address request
      - *other commissives, expressable by means of performative verbs*
    - *Directives:*
      - instruction, address request, indirect request, (direct) request, suggestion
      - *other directives, such as advice, proposal, permission, encouragement, urge,..., expressable by means of performative verbs*

(a) Structure de la taxonomie DIT++ des fonctions communicatives à usage général (Bunt, 2009)

1. *Task/Activity*, pour tout ce qui se rapporte à la tâche qui est l'objet de la conversation ;
2. *Auto-Feedback*, pour les actes signifiant le niveau de compréhension et d'interprétation du locuteur ;
3. *Allo-Feedback*, *idem* pour l'allocutaire ;
4. *Turn Management*, pour les actes portant sur la gestion du tour de parole ;
5. *Time Management*, pour les situations où il est nécessaire de signifier que le locuteur a besoin de plus de temps pour contribuer ou qu'il faut faire une pause ;
6. *Contact Management*, pour les actes qui servent à établir et maintenir la communication ;
7. *Own Communication Management*, pour les actes servant à indiquer que le locuteur prépare ou modifie sa contribution au dialogue ;
8. *Partner Communication Management*, pour les actes effectués par un participant endossant le rôle d'allocutaire, servant à assister son partenaire dans la formulation de sa contribution ;
9. *Discourse Structure Management*, pour les actes servant à structurer thématiquement la conversation ;
10. *Social Obligations Management*, pour les actes de gestion sociale du dialogue.

(b) Exemples de fonctions communicatives spécifiques de DIT++

FIGURE 5.B – Taxonomie DIT++

C) **Norme ISO 24617-2**

En 2012, (Bunt et al., 2012) développent la norme ISO 24617-2. Cette norme est un framework d'annotations sémantiques, qui n'est rien d'autre qu'une mise à jour d'une version antérieure de la norme ISO DIS 24617-2 :2010 développée par (Bunt et al., 2010). Pour passer de cette dernière à la nouvelle norme (Bunt et al., 2020), des concepts d'annotation de relations rhétoriques, de dépendances fonctionnelles et de feedback entre des unités de dialogue y ont été ajoutés. Le langage de balisage d'ADs, **DiAML (Dialog Act Markup Language)** a été utilisé et certains de ses éléments et attributs ont été restructurés et adaptés afin de produire la norme ISO 24617-2.

Avant de détailler les éléments qui ont été ajoutés à la nouvelle norme, nous allons ci-dessous présenter les principales caractéristiques de la version ISO DIS 24617-2 :2010 :

1. Les aspects de dimensions sémantiques dans l'analyse d'ADs y sont incorporés et ces dimensions 5.B sont au nombre de 9. Ce sont en fait celles que Bunt a définies dans la taxonomie *DIT++*. La seule différence est que la dimension «Contact Management» qui permettait d'annoter les actes d'établissement et de maintien de communication, a été supprimée et sa fonction a été incorporée dans la dimension «Discourse Structuring» qui, elle, permet d'annoter les actes de gestion de sujet, d'ouverture et de fermeture des dialogues (sous-dialogues aussi) et de les structurer. Cette fonction de structuration était la seule de la dimension «Discourse Structuring» dans *DIT++*. Ces dimensions se distinguent les unes des autres sur les fondements empiriques et théoriques qui permettent ainsi d'annoter les énoncés de façon multidimensionnelle, c'est-à-dire qu'un segment texte de dialogue peut être annoté par plus d'une de ces dimensions sans ambiguïté.
2. Deux classes de fonctions communicatives dont une spécifique à une dimension et l'autre à un usage plus général constituent aussi l'une des caractéristiques de cette norme. Ces fonctions peuvent ainsi être combinées avec n'importe quel contenu sémantique pour former un acte de dialogue (AD) dans la dimension correspondante.
3. Des **fonctions qualificatives** utilisées pour annoter un énoncé en AD de façon conditionnelle ou non, avec certitude ou pas ou bien avec un sentiment particulier.
4. Les relations de dépendance fonctionnelle ou de feedback sont définies pour mettre en relation des énoncés annotés en ADs avec des segments de texte préalablement identifiés. Rendant ainsi plus explicite l'association d'une réponse à une question en amont ou bien d'un feedback à l'énoncé d'un interlocuteur ou, encore, une approbation à une suggestion.
5. La notion de **segment fonctionnel** est utilisée comme unité d'annotation d'AD. C'est l'entité minimale comportementale possédant une ou plusieurs fonctions communicatives.
6. La segmentation multifonctionnelle est appliquée, montrant ainsi la distinction que peut avoir chaque dimension sur un segment fonctionnel. Un segment portant une fonction de feedback peut se chevaucher avec un segment qui est une fonction liée à une tâche.
7. **DiAML** est tout aussi représenté sous trois angles : (i) une syntaxe abstraite qui spécifie les annotations possibles avec un ensemble de termes théoriques, (ii) une sémantique qui spécifie les interprétations des structures définies par la syntaxe abstraite, (iii) une syntaxe concrète fixant une représentation XML des structures d'annotations.

À la différence du standard ISO DIS 24617-2 :2010, le méta modèle de la figure 5.C fait ressortir de nouveaux concepts tels que les relations rhétoriques (qui n'existaient pas dans ISO DIS 24617-

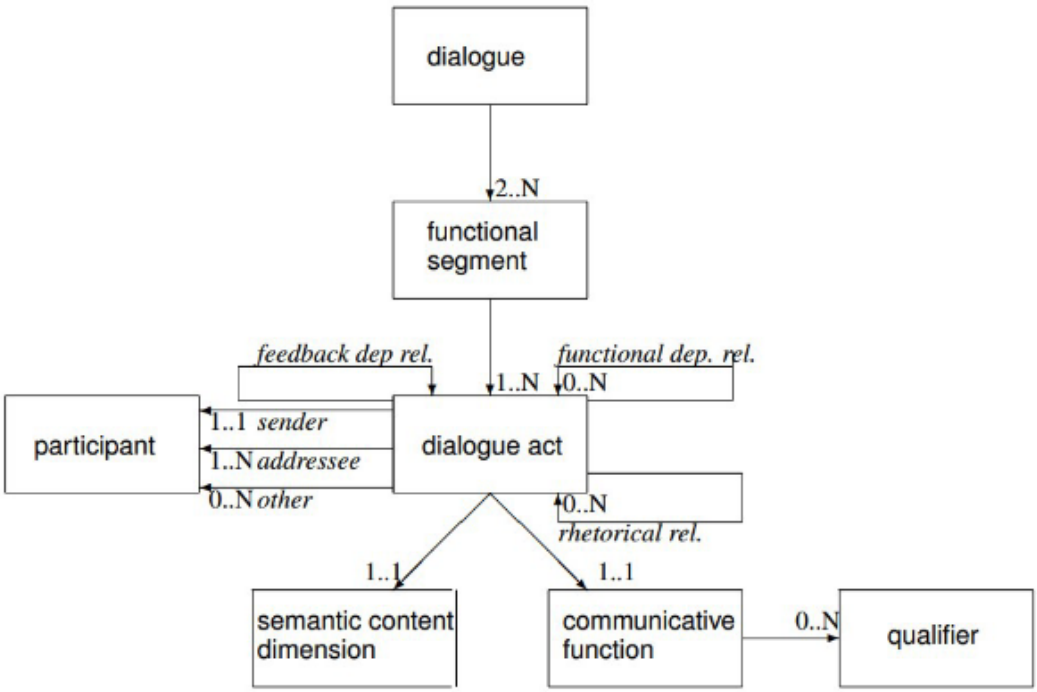


FIGURE 5.C – Méta-modèle d'annotation d'ADs de (Bunt et al., 2012)

2 :2010) entre ADs. Ces relations sont plus marquées dans les textes écrits, puisqu'elles ont été intensivement étudiées dans ce contexte avant de se rapporter aux dialogues. Dans (HOBBS, 1985; MANN and THOMPSON, 1988; Prasad et al., 2008), ces relations rhétoriques ont des appellations différentes : **relations de cohérence** ou **relations de discours**. Un dialogue au travers de ses différents tours de paroles peut être considéré comme une conversation écrite (chats ou échange d'emails), et on y distingue deux types de relations rhétoriques :

1. Relations rhétoriques de dépendance fonctionnelle : par exemple, une réponse d'un interlocuteur à une question rhétorique dans un énoncé antérieur à cette réponse. Ceci pouvant s'effectuer de façon bidirectionnelle entre deux intervenants ou dans le cadre d'une conversation multipartite (forum).
2. Relations rhétoriques de dépendance de feedback : celles–ci se matérialisent par une relation entre un énoncé du type **«sûr»** et un **hochement de tête** d'un autre interlocuteur. Ce type de relation est plus visible en communication parlée, mais avec l'avènement des émoticones/émojis, on peut aussi les retrouver dans les conversations écrites. La relation de dépendance de feedback peut aussi se manifester par une explicitation d'un concept ou l'enrichissement d'une information suite à un énoncé du genre **«suis pas sûr d'avoir bien compris»**.

Dans cette nouvelle norme ISO 24617-2, les relations de dépendance fonctionnelle (DF) surviennent avec des ADs de nature réactive tels que des réponses, confirmations, accords, acceptations, excuses, refus ; les contenus sémantiques de ces types d'AD dépendent fortement des précédents

actes auxquelles ils répondent. C'est pour cette raison qu'elles peuvent s'exprimer avec des énoncés courts tels que « oui, non, merci, ok, etc » qui, eux, ne possèdent aucun contenu sémantique.

La norme ISO 24617-2 est un framework complet d'annotation d'énoncés de conversations écrites ou orales ; de plus, elle suit les directives de l'encodage et de l'échange de texte électronique [1] du standard *TEI* (Text Encoding Initiative) d'encodage de données au format XML. Cette norme est très utilisée pour annoter des énoncés (Papalampidi et al., 2017; Mezza et al., 2022) dans le cadre des problématiques autour des conversations et des dialogues, par exemple la nôtre qui s'intéresse à la constitution de sous-fils de conversations d'emails avec chacune de ces sous-conversations traitant d'un sujet bien spécifique. Elle est aussi utilisée et parfois même adaptée en amont de la résolution de ces problématiques, pour la construction de corpus (Bunt et al., 2016; Asri et al., 2017; Feng et al., 2020).

Comme nous venons de le voir, il existe nombre de travaux qui utilisent la norme ISO 24617-2. Cependant pour notre problématique de constitution de sous-fils de conversations d'emails, à notre connaissance, il n'existe pas de corpus de conversations d'emails annotées selon cette norme et qui nous permettrait d'approcher notre problématique. Nous nous sommes ainsi intéressées à différents corpus en plus d'un corpus d'emails constitué chez Orange, toutes les caractéristiques de ces corpus sont présentées dans le chapitre 6. L'analyse de ces corpus nous a permis d'identifier des aspects qui ne permettaient pas une annotation avec une granularité fine et facilement compréhensible de notre point de vue des énoncés d'emails en ADs des énoncés. Ces aspects peuvent être considérés comme des insuffisances partielles pour notre besoin et sont d'une part dus au fait que le corpus que nous avons exploité pour la classification en ADs des énoncées n'est pas un corpus de conversations d'emails, mais plutôt un corpus de transcriptions d'enregistrements de réunions. Ces insuffisances partielles nous ont incités à définir un référentiel d'ADs adapté à notre besoin, mais qui peut aussi être utilisé pour résoudre d'autres problématiques liées aux ADs.

5.3 Notre référentiel d'ADs

Nous avons fait mention dans la précédente section de l'analyse de certains corpus qui nous a permis d'identifier dans ceux-ci des aspects qui sont considérés comme des insuffisances partielles autour de l'utilisation des ADs. L'un des principaux corpus que nous avons exploités est le corpus *MRDA* [2] qui est un corpus de conversation orale et synchrone, contrairement à un corpus de conversation d'emails qui est écrit et asynchrone. On constate ici les premières différences qui peuvent justifier ces insuffisances et se matérialisent par l'existence des marqueurs de parole dans le corpus MRDA qu'on ne retrouvent pas dans des conversations écrites formelles. Ces marqueurs sont par exemple des interjections utilisées pour exprimer différents types de sentiment :

— « Ah » : utilisée pour exprimer une surprise, une douleur, une satisfaction ou une reconnaissance,

— « Oh » : utilisée pour exprimer une surprise, une admiration, une déception ou une tristesse,

— « Oups » : utilisée pour exprimer une gêne, une maladresse ou une erreur,

— « Aïe » : utilisée pour exprimer une douleur physique,

1. P5: Guidelines for Electronic Text Encoding and Interchange
2. Meeting Recorder Dialogue Act Corpus

— « Eh bien » : souvent utilisée pour exprimer une approbation, une surprise, une impatience ou une résignation.

Ces interjections ont leurs correspondants en anglais, langue du corpus MRDA. En plus de ces interjections, on retrouve l'expression « you know », fortement utilisée dans les dialogues. Ces marqueurs dans le corpus MRDA sont parfois les seuls constituants d'énoncés annotés ; on les retrouve aussi parfois en début et milieu d'énoncés. Les ADs utilisés pour l'étiquetage des énoncés de MRDA sont une version modifiée de *SWBD (Switchboard)-DAMSL* [3]. Dans cet ensemble d'ADs, nous avons constaté que les acronymes utilisés pour certains de ces ADs ne permettaient pas une identification aisée de ceux-ci. C'est le cas par exemple pour « Acknowledge-answer, Reject, Offer, Yes answers » qui ont respectivement pour acronymes « bk, ar, co, ny ». Un autre aspect qui peut pousser à la confusion lors de la phase d'annotation est la diversité des ADs utilisés dans MRDA. Aussi, certains ADs que l'on retrouve dans MRDA ne seront que très peu ou presque pas utilisés dans le cadre des conversations écrites asynchrones comme les emails ; par exemple, « Self-talk, 3rd-party-talk, Hold before answer/agreement, Signal-understanding » sont très souvent fortement liés aux dialogues.

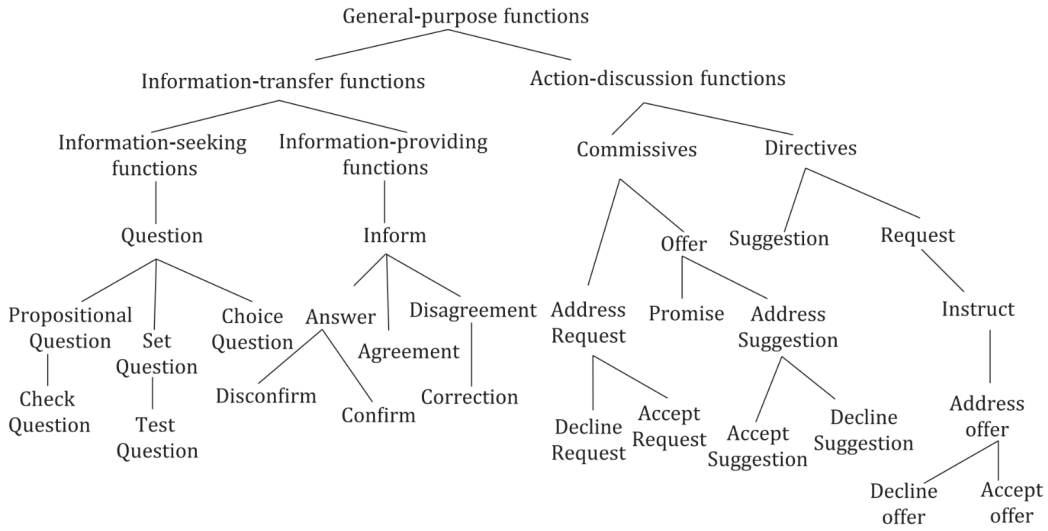


FIGURE 5.4 – Structure arborescente des fonctions communicatives à usage général (extrait de la norme ISO 24647-2 (Bunt et al., 2020))

C'est pour ces différentes raisons que nous avons opté pour une adaptation des ADs utilisés dans MRDA. L'adaptation que nous avons effectuée s'est déroulée en trois étapes :

— premièrement nous avons identifié sur les différentes fonctions communicatives (figure 5.4), les actes de gestion de tâches et les feedbacks définis dans la norme ISO 24617-2 (Bunt et al., 2020) qui répondaient le mieux à l'identification des ADs dans les emails et aussi aideraient à une structuration dialogique de conversation d'emails,

3. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation

— ensuite nous avons défini des acronymes qui permettent de facilement identifier chaque AD sans ambiguïté,
— et enfin nous avons fait correspondre les acronymes d'ADs 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.

Lors de ce processus de correspondance d'ADs et sur la centaine d'énoncés de MRDA consulté nous avons progressivement défini un ensemble de règles qui nous ont permis d'automatiser ces correspondances sur l'ensemble du corpus MRDA. Ces règles s'appuient sur les différentes catégories des ADs de chaque énoncé du corpus, à savoir : basique, générale et fine. Nous avons explicité les ADs de la catégorie fine (Annexe C), ce qui nous a permis une meilleure compréhension de ceux-ci. Pendant que certaines de ces règles pour la majorité reposent uniquement sur l'AD fine de l'énoncé, les autres en revanche s'adossent sur une combinaison des ADs généraux et fins. Ces 3 catégories sont représentées par un triplet sous la forme **x|y|z** où x, y et z sont à leur tour soit un chiffre, une lettre, une combinaison de 2 ou 3 lettres ou de lettres et chiffres (Annexe B pour plus de détails).

L'ensemble des règles que nous avons définies sont les suivantes et sont catégorisées en fonction de la première couche des fonctions communicatives de la norme ISO 24617-2.

1. **Fonction de demande ou de recherche d'informations (*Information Seeking*)**
   Les demandes ou recherches d'information sont en général des questions qui se subdivisent en sous-catégories de questions comme dans la figure 5.4. Ci-dessous sont les règles mises en oeuvre pour ces fonctions de recherche d'information.
   — Dans MRDA tous les énoncés avec « br » (*Signal-non-understanding*) comme AD fin sont des questions « Q » dans notre référentiel.
   — Les énoncés taggés par « Q|fh|fh » (*Q=Question, fh=Floor Holder*), « Q|q ?|fe » (*q ?=any type of question, fe=Exclamation*), « qy » (Yes-No-question), « qo » (Open-ended Question), « qy|cs » (cs=Offer) sont pour nous des **questions avec proposition** avec l'acronyme **PrQ** (*Propositional Question*).
   — Les énoncés dans MRDA avec « bu » (*Understanding Check*), « d » *Declarative-Question*), « g » *Tag-Question*) comme AD fin rentre dans la catégorie **questions de vérification** (*Check Question* = **CkQ**).
   — Dans notre référentiel, les **questions fermées ou à choix prédéfinis** (en anglais *Set Questions* = **SQ**) regroupent des énoncés qui ont « qw » (*Wh-Question*) comme AD fin ou « qw|cs » comme AD général et fin.
   — les ADs « qr » (*Or Question*) et « qrr|qrr » (qrr= *Or-Clause*) sont pour nous des **questions à choix** (*Choice Question* = **ChQ**)
2. **Fonctions de fourniture d'informations (*Information Providing*)**
   La mise à disposition d'information survient en général sous deux situations : c'est soit une réponse à une demande d'information ou bien un simple partage d'information. Pour ce type de fonctions communicatives, nous avons défini les règles suivantes :
   — La catégorie *Inform* (**I**) du référentiel correspond aux énoncés de MRDA qui ont « S » (*Statement*) pour AD basique et la combinaison « S|fh|fh » pour les différentes catégories d'AD.

— Les énoncés taggés « no » (*Other Answers*) de MRDA sont des **réponses** (*Answers* = **An**)

— Les confirmations (**Cf**) de notre référentiel correspondent aux ADs « na » (*Affirmative Non-yes Answers*) de MRDA.

— Les refutations (*Disconfirm* = **Dcf**) de notre référentiel se prêtent aux ADs « nd, ng » (*Dispreferred Answers*, *Negative Non-no Answers*).

— Dans MRDA, les énoncés avec « aa » (*Accept*) pour AD fin sont des **accords** (*Agreement* = **Ag**). De même les rejets « ar » (*Reject*) de MRDA sont pour nous dans la classe **désaccord**(*disagreement* = **Dag**)

— Les partages d'information sous forme de **correction** (**Cr**) du référentiel se rapportent aux « bc » (*Correct-misspeaking*) de MRDA

Le tableau 5.2 récapitule le référentiel que nous avons défini. Dans la colonne la plus à gauche, les ADs de MRDA qui ont été mappés. Les ADs qui n'apparaissent pas ici sont ceux que nous avons estimés inadaptés pour les conversations d'emails. En orange, les fonctions communicatives de la norme ISO 24617-2 et enfin, en bleu, les acronymes et ADs que nous avons définis. Les acronymes s'interprètent facilement avec leurs ADs respectifs, ce qui n'est pas le cas pour ceux de MRDA et de Switchboard Dialogue [4].

4. Switchboard Dialogue Act Corpus

		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		Address Request		
	AcR				Accept Request	
	DeR				Decline Request	
cs	O		Offer			
	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 5.2 – Référentiel d'ADs définis en s'appuyant sur la norme ISO 24617-2 et leur correspondance avec les ADs de MRDA

5.4 Conclusion

Dans ce chapitre, nous avons parcouru l'analyse discursive, partant des premières taxonomies d'actes de langage jusqu'à la dernière version de la norme ISO 24617-2 (Bunt et al., 2020), établie en 2012 et qui permet d'annoter des conversations écrites, parlées ou même transcrites. Ces taxonomies et normes ont aussi permis de mieux cerner et capturer les différentes caractéristiques linguistiques dans des conversations, ceci au travers des différentes couches, dimensions et relations de dépendances entre elles. Tous ces travaux donnent ainsi la possibilité d'implémenter des systèmes automatiques de dialogues ou de conversations (chabots). Force est de constater que la majorité des briques autour de l'analyse des discours a été abordée depuis 1969 par Austin. Et depuis leurs prémices, ces éléments de l'analyse discursive sont utilisés dans différents projets de recherche, d'annotation des corpus et même pour la conception d'applications de communication intelligente et autonome (Djingo, Siri, Alexa...). Malgré les nombreux travaux effectués sur l'analyse de conversations, il n'est toujours pas évident, même avec les évolutions des récents larges modèles de langage, de répondre de façon souhaitée par les humains à certaines problématiques dans ce domaine d'analyse conversationnelle. Cependant, pour mieux approcher notre problématique, les différents aspects de l'analyse discursive abordés dans ce chapitre, et plus précisément notre référentiel d'actes de dialogue (ADs), nous seront très utiles en première ligne notamment pour annoter des corpus ou bien pour catégoriser des segments de textes d'emails en ADs.
Dans la suite de ce document, nous détaillons les différentes expériences que nous avons menées en vue de la résolution de notre problématique. Il s'agira tout d'abord de présenter les différents corpus que nous avons utilisés, ensuite de détailler les modèles d'intelligence artificielle entraînés pour la classification de segments de texte d'emails en ADs et pour l'appariement desdits segments de texte de façon transverse dans une conversation et, enfin, nous exposerons les évaluations de ces modèles.

Troisième partie

Corpus utilisés et méthodes proposées

6 Corpus et Exploitation

6.1 Introduction

Constituer un corpus est souvent nécessaire pour atteindre un objectif bien précis qui est en général la résolution d'une problématique. Dans le cadre des conversations asynchrones, le démêlage de conversation, l'analyse de discours, la segmentation d'emails, la reconnaissance d'entités nommées etc. sont quelques exemples de problématiques qui requièrent la constitution de corpus. C'est dans ce sillage que (Chanier et al., 2014), pour analyser des discours et effectuer des études linguistiques des idiolectes qui apparaissent dans différents types CMO, ont construit le corpus CoMeRe constitué de contenus hétérogènes en langue française dont 3 millions de tchats, 44K SMS, 2300 emails, 2700 messages de forum et 34k Tweets. Classifier de façon automatique des emails dans des répertoires spécifiques de messagerie et extraire des informations de ces emails sont des problématiques qui ont poussé à la construction du Corpus ENRON par (Klimt and Yang, 2004a), un corpus d'emails en langue anglaise extrait des échanges d'une entreprise privée. Ce corpus ENRON a largement été utilisé après sa constitution ; les formalités de courrier électronique dans des environnements de travail ont été étudiées par (Peterson et al., 2011), de même (Chhaya et al., 2018) ont analysé les sentiments et les tons utilisés dans les emails. Afin de démêler des conversations dans un même flux de messages, (Kummerfeld et al., 2019) ont développé un grand corpus de 77563 messages dont 74963 proviennent d'un forum autour d'Ubuntu et 2600 messages de celui de Linux contribuant fortement aux recherches autour de l'analyse des dialogues. Tout récemment (Bevendorff et al., 2020) ont développé le plus grand corpus disponible nommé Webis Gmane Email Corpus 2019 constitué de 153 millions d'emails extraits de 14669 listes de diffusion. Ce corpus est disponible en trois langues majeures : anglais, français et allemand. Tous ces travaux montrent en effet que la constitution d'un corpus répond à des problématiques bien identifiées et les corpus une fois constitués contribuent énormément à l'amélioration d'approches existantes et à la résolution de nouvelles problématiques. Dans les prochaines sections, nous présentons quelques corpus en langue anglaise que nous avons utilisés dans le cadre de nos travaux et nous nous attardons sur un corpus d'emails d'entreprise que nous avons constitué au sein d'Orange.

6.2 Corpus Orange

Comme mentionné plus haut et avec les travaux sus-présentés, on peut constater la rareté ou l'inexistence des corpus d'emails d'entreprise en langue française. Et c'est ce qui nous a emmenés à construire un corpus à Orange qui est soumis à différentes contraintes que nous présentons dans la section 6.2.1.

6.2.1 Contraintes juridiques liées à l'utilisation des données

Certaines conditions encadrent l'exploitation de données publiques notamment les données de forum ou de listes de diffusion. Ces types de condition sont en général soumis lors des inscriptions à ces forums et/ou listes de diffusions. Les mentions contenues dans ces conditions font souvent référence à l'exploitation des données des utilisateurs à des fins de formation ou d'améliorations de services. Contrairement à ce processus d'inscription à des forums ou listes de diffusions, les collaborateurs d'une entreprise ne valident pas de telles conditions, mais sont tenus de garder secret certaines informations d'entreprise d'une part et d'autre part les informations à caractère privé échangées en entreprise via des CMO sont régies à la fois par l'article L33-1 du code des postes et des communications électroniques sur le respect du secret de la correspondance et par le Règlement Général sur la Protection des Données (RGPD) adopté en 2016 et entré en vigueur le 25 mai 2018 dans l'Union Européenne. Le principe de protection de données personnelles existe depuis 1970 et la paternité de la confidentialité des données dans toute conception technologique est attribuée à Ann Cavoukain, la Commissaire de la protection de la vie privée de l'Ontario, Canada, coauteure du rapport international de 1995 sur les technologies d'amélioration de la protection de la vie privée (Privacy Enhancing Technologies - PET en anglais). (Cavoukian, 2010) a d'ailleurs précisé que *« l'avenir de la vie privée ne peut être assuré uniquement par le respect des cadres réglementaires; plutôt, la protection de la vie privée doit idéalement devenir le mode de fonctionnement par défaut d'une organisation »*. Dans ses travaux elle liste sept principes fondamentaux de la confidentialité dès la conception.

Concernant le secret de la correspondance, il y a violation de celui-ci si une personne tierce parvient à accéder à n'importe quelle partie d'une conversation privée entre collaborateurs sans le consentement de ces derniers. Notre objectif de construction de sous-fils de conversations d'emails d'entreprise ne peut se faire qu'en respectant le secret de la correspondance, nous contraignant ainsi à émettre des requêtes de consentement aux différents collaborateurs impliqués dans ces conversations. Ce qui se révèle une tâche fastidieuse vu le nombre de collaborateurs dans les différentes équipes ou entités au sein d'Orange.

Dans le même ordre idée de protection des données personnelles, le RGPD a été institué sur le territoire de l'Union Européenne. Il s'applique à toute organisation, publique et privée, qui traite des données personnelles pour son compte ou non, dès lors qu'elle est établie sur le territoire de l'Union Européenne, ou que son activité cible directement des résidents européens. Orange est un exemple de ce type d'organisation. Aussi, dans le cadre de travaux comme les nôtres où nous sommes amenés à traiter des données personnelles, le RGPD renforce l'obligation d'information et de transparence à l'égard des personnes (principalement des collaborateurs internes dans notre cas) dont nous traitons et analysons les données. Cette obligation de transparence à l'endroit des collaborateurs, la possibilité pour ceux-ci d'exercer leur droit de retrait et la sécurisation de ces données sont des conditions à respecter afin de satisfaire le RGPD. Ces conditions sont contraignantes pour l'acquisition des données parce qu'elles nécessitent de mettre en place des processus stables et épurés qui incluent des aspects administratifs et matériels.

Les deux précédentes contraintes positionnent les collaborateurs au centre de tout processus d'acquisition de données en entreprise. Ces collaborateurs doivent savoir pourquoi, comment et pendant combien de temps leurs données seront exploitées. Et en fonction des réponses qui leur sont fournies ils sont libres de donner leur accord ou pas. La réticence au partage de leurs contenus

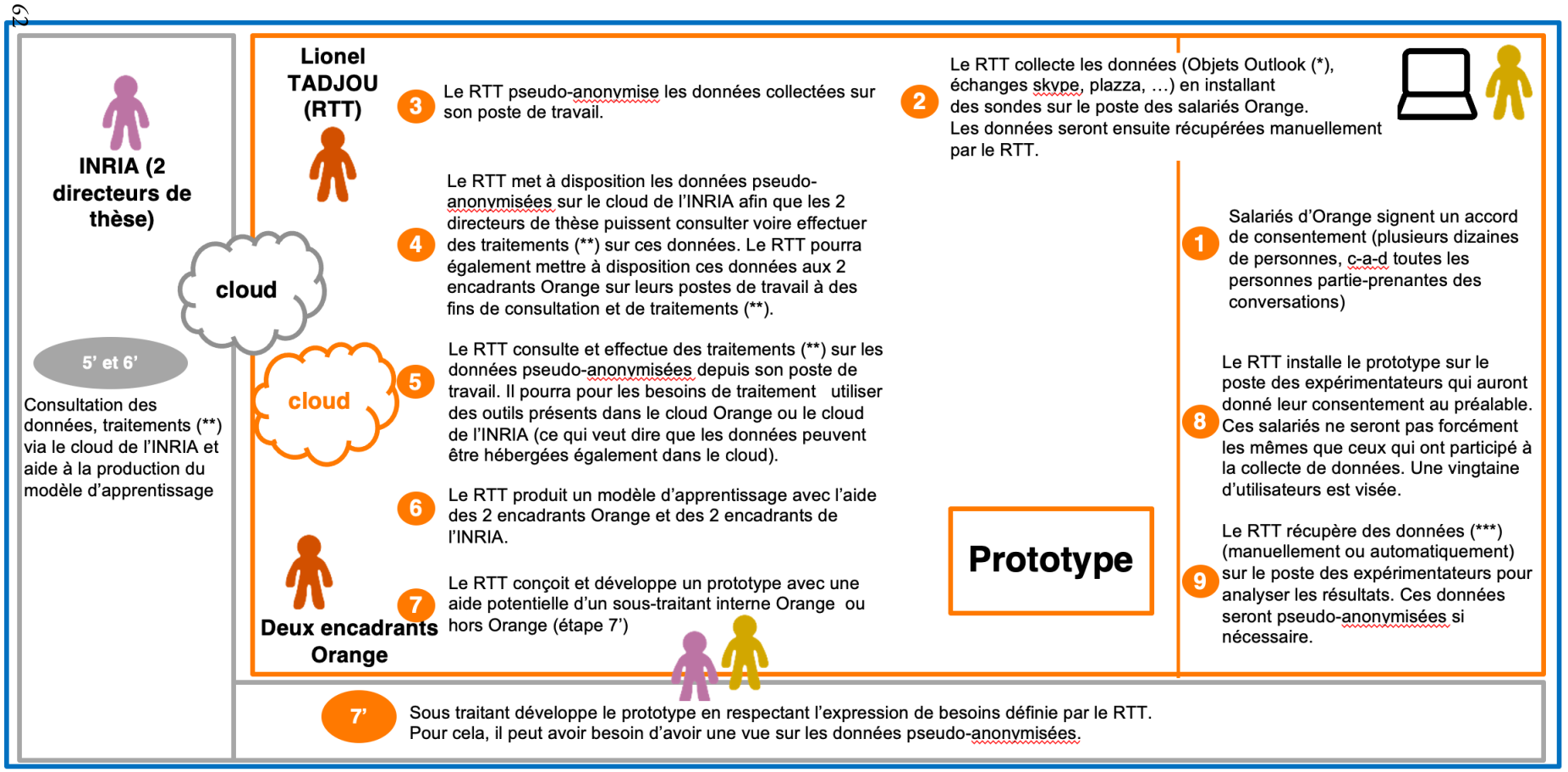
qui comportent potentiellement des données personnelles ou très sensibles fait ressortir un autre groupe de contraintes qui sont d'ordre personnel, d'acceptabilité et même psychologique.

Toutes les contraintes précédemment présentées constituent des verrous majeurs dont les déblocages sont nécessaires pour toute acquisition et exploitation de données en entreprise.

Ainsi pour être en conformité avec le RGPD et le secret de la correspondance, notre projet a été soumis à l'entité juridique d'Orange qui nous a demandé de remplir un formulaire et de produire un cycle de vie des données (figure 6.1)respectant certaines contraintes. Le formulaire en question était une analyse des risques et des incidents probables qui pourraient résulter de l'extraction et de l'exploitation des données de messagerie. Le cycle de vie des données contient des informations sur quelles données sont collectées et analysées, comment et par qui. Dans le cadre d'un doctorat, ce qui est notre cas, le cycle de vie des données précise également qu'avant l'anonymisation ou la pseudo-anonymisation, seul l'étudiant en doctorat appelé *contrôleur technique des données (CTD)* peut accéder à l'ensemble des données collectées sur un dispositif de stockage sécurisé. Les collaborateurs ne peuvent accéder qu'aux données collectées sur leurs ordinateurs respectifs, ce qui signifie qu'ils ne peuvent consulter que les emails des conversations auxquelles ils participent. Une commission juridique à Orange a évalué notre besoin de collecte d'emails sur la base du formulaire et du cycle de vie des données et a défini le processus à suivre. Ce processus a permis d'une part d'obtenir le consentement d'un nombre significatif de collaborateurs et d'autre part de substituer ou de pseudo-anonymiser des données sensibles et des chaînes de caractères qui pourraient conduire à l'identification d'une personne de façon directe ou indirecte. Un aspect à mentionner dans ce processus est que le consentement fourni par les collaborateurs n'est disponible que pendant une année au maximum et doit être renouvelé si l'analyse des données doit se poursuivre au-delà de ce délai.

Pour obtenir le consentement des collaborateurs, nous avons conçu un support de communication au format vidéo et diapositives expliquant pourquoi, comment et pendant combien de temps le processus de collecte et d'analyse des données aura lieu. Ceci afin d'être transparent auprès de ceux dont nous voulons collecter les données, comme l'exigent les contraintes juridiques dont le RGPD.

Dans les prochaines sections, nous présentons les différentes approches méthodologiques que nous avons adoptées afin de débloquer ces verrous. Ces approches font partie intégrante d'un processus en trois étapes principales : la collecte des conversations, leur prétraitement et la pseudo-anonymisation des informations sensibles contenues dans ces conversations.



(*) Les objets Outlook sont les mails, RDV Calendrier, tout élément se trouvant dans le client Outlook.
(**) Exemples de traitement : preprocessing (nettoyage, lemmatisation...), analyse (corrélation...), mise en œuvre d'algorithmes de TAL (traitement automatique du langage) et d'autres algorithmes (non identifiés précisément à ce stade)
(***) Des mesures d'efficacité des modèles, voire des modèles d'apprentissage afin d'observer l'évolution de ceux-ci

FIGURE 6.1 – Cycle de vie de données produit dans le cadre du processus de collecte de données

6.2.2 Processus de collecte des conversations d'emails

En général, pour toute tâche d'analyse de données par des outils statistiques, d'ETL (Extraction, Transformation, Load) ou autres, la première étape consiste à collecter des données, à moins qu'elles soient déjà disponibles. Dans notre cas, comme nous l'avons mentionné précédemment, il n'existe pas de corpus de conversations d'emails disponible en français qui puisse être exploité pour notre problématique. Les conversations d'emails ont été la première source de contenus à partir de laquelle nous avons décidé d'extraire des données, ce parce qu'à Orange les emails sont encore très largement utilisés et contiennent plus de connaissances que les conversations existantes sous d'autres modalités, comme les discussions avec Skype par exemple. Notre objectif principal étant de construire des sous-fils de conversations, les approches proposées pour les conversations d'emails pourront ensuite être également appliquées à d'autres types de CMO dans les entreprises tels que les données de forums internes, les tchats, la messagerie instantanée, etc.

Pour extraire tous les emails de collaborateur dans une entreprise comme Orange, il faut être en conformité avec le RGPD et respecter le secret des correspondances comme nous avons mentionné précédemment. Pour ce faire, l'obtention des consentements de tous les collaborateurs d'Orange et des partenaires externes impliqués dans ces échanges est une étape essentielle. Cette étape est très difficile et prend beaucoup de temps en raison du grand nombre d'employés et de l'organisation matricielle du travail à Orange. Même si de nombreux collaborateurs donnent leur accord, il sera toujours difficile de collecter leurs emails à partir d'un entrepôt de données tel que le serveur Microsoft Exchange en raison de contraintes de sécurité. Le contexte général de nos travaux est centré sur le poste de travail d'un utilisateur et nous avons choisi de collecter les emails directement à partir des ordinateurs des collaborateurs. Ce choix a été guidé par deux raisons principales, à savoir :

— Il est facile d'obtenir le consentement de certains collaborateurs ciblés. Les aspects de sécurité et de confiance sont ainsi facilement abordés car les collaborateurs ont le plein contrôle du contenu extrait de leur ordinateur.

— Il permet également de ne collecter que les emails strictement nécessaires et non ceux dont le collaborateur ne voudrait pas les extraire.

Face aux contraintes du grand nombre de personnes à convaincre, malgré des collaborateurs bien ciblés, l'extraction d'une quantité significative d'emails et de conversations nécessite un nombre important de personnes à contacter. L'une des principales difficultés est que si dans une conversation comprenant de nombreux collaborateurs, un seul d'entre eux ne donne pas son consentement, alors aucun des mails dans lesquels celui-ci est partie prenante ne pourra être extrait. Si une telle situation s'étend à plusieurs autres conversations et emails, l'extraction se soldera tout simplement par un échec car les conversations qui nous intéressent en premier lieu seraient de taille insignifiante et ne pourraient pas nous aider à approcher le démêlage des conversations. Pour faire face à ce problème, nous proposons une approche où nous maximisons le ratio entre le nombre d'emails/conversations à extraire et le nombre de collaborateurs à contacter. Nous distinguons également deux types de collaborateurs qui ont besoin de donner leur accord, à savoir : - *référents*, qui sont des collaborateurs au sein de notre équipe et avec lesquels nous effectuons des manipulations sur leur poste de travail afin de collecter des emails ; - *interlocuteurs*, qui sont des collaborateurs impliqués comme simples parties prenantes des emails collectés ci-dessus.

La précédente approche est divisée en deux passages. Le premier consiste simplement à sélectionner un nombre réduit d'interlocuteurs proches d'un référent tout en maximisant le nombre d'emails associés à ceux-ci. Le résultat de ce premier passage est une liste d'interlocuteurs qui sert de guide lors de la sélection d'interlocuteurs sur les postes de travail d'autres référents. Le deuxième passage se concentre sur l'extraction des emails et des conversations sur les postes de travail des référents. Lors de l'extraction des emails/conversations, nous avons également veillé à extraire les métadonnées telles que les expéditeurs, les destinataires, les identifiants d'emails et de conversations, les dates et les heures des en-têtes des emails qui nous serviront plus tard à l'étape de pseudo-anonymisation et pour conserver les liens réels entre les messages des conversations.

Cependant, avant toute sélection d'interlocuteurs, il y a eu l'étape d'obtention du consentement des collaborateurs concernés. L'objectif ici était de générer pour chaque collaborateur contacté un document expliquant pourquoi nous voulons exploiter certains emails dans lesquels il apparaît comme interlocuteur afin de recueillir sa signature attestant ainsi qu'il nous a donné son consentement. A cause des périodes de confinement lors de la pandémie de COVID-19, nous avons eu recours, de manière systématique, à un systèmes de signature électronique recommandé par le département juridique d'Orange ; les consentements obtenus étaient alors stockés dans un système de fichiers sécurisé.

Pour l'obtention de ces consentements, nous avons adopté une approche de proximité consistant à contacter d'abord les collaborateurs d'une même équipe, puis ceux d'un même projet, enfin ceux d'une même entité, et ainsi de suite.

La collecte d'emails/conversations sur les poste des collaborateurs s'est faite de façon itérative à chaque fois qu'un certain nombre de nouveaux consentements étaient obtenus. Nous avons conçu et développé un outil nommé **OutlookScrapping** qui facilite l'extraction des données en respectant le processus décrit ci-avant. La mise en œuvre et l'utilisation dudit outil sont détaillées dans la section 6.2.3. **OutlookScrapping** produit en sortie des données aux formats JSON et CSV afin de faciliter les traitements ultérieurs. Les données collectées sont envoyées au CTD, soit sous la forme d'un fichier crypté envoyé par courrier électronique, soit au moyen d'un dispositif de stockage sécurisé et crypté.

Nous avons ainsi obtenu le consentement de 122 collaborateurs impliqués dans des projets communs. Grâce à ces 122 consentements, nous avons pu extraire 12k emails sans doublons sur le poste de cinq collaborateurs de la même équipe, y compris le CTD, mais contribuant à une demi douzaine de projets différents. Ces emails incluent des conversations sur la période 2013 à 2021 pour certains collaborateurs ayant occupé différents postes, ceci montre que les données collectées sont riches en informations sur plusieurs années (de 2013 à 2021). Le nombre d'employés chez Orange en 2021 était de 132K et donc on aurait pu collecter un million d'emails avec environ 10k consentements soit environ 8% de tous les employés. Dans la prochaine section nous présentons concrètement comment les données ont été collectées et prétraitées suite à la démarche que nous venons de décrire.

6.2.3 Extraction des emails/conversations et prétraitement

Dans cette section, nous présentons les outils et méthodes utilisés pour la mise en œuvre de *OutlookScrapping*, ensuite nous décrivons ses fonctionnalités afin de terminer sur les différentes phases de prétraitement des données collectées.

A) Collecte de données

La principale application utilisée à Orange pour l'échange d'emails est Microsoft Outlook. Nous avons donc développé une application C# Windows Presentation Foundation (WPF)[1] basée sur *Microsoft Office Interface*[2] qui nous permet d'interagir avec l'application Outlook et également de sauvegarder des fichiers à partir d'Outlook. *OutlookScrapping* est l'outil que nous avons développé pour collecter les contenus de messagerie d'Outlook que nous manipulons ensuite à l'aide d'algorithmes et de scripts afin d'en dégager les interlocuteurs ciblés et les conversations d'emails correspondants. L'interface de cette application, ainsi que ses différentes fonctionnalités sont détaillées en annexe A. Nous avons ainsi pu collecter environ 12k emails sans doublons. Dans ces 12k, on distingue 5105 conversations avec chacune un seul email et le reste sont des conversations avec au moins deux emails. La figure 6.B présente la répartition du nombre d'emails par conversation.

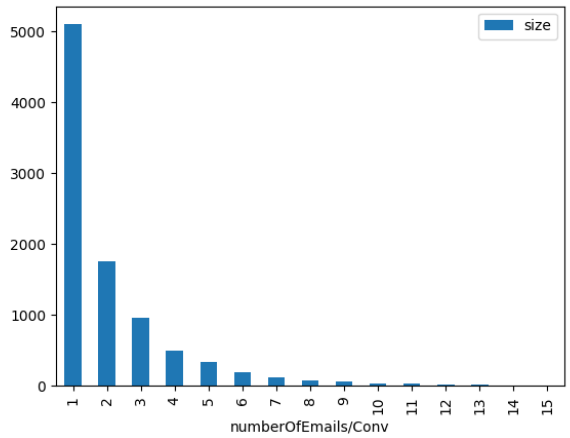


FIGURE 6.B – Répartition du nombre d'emails par conversation dans le corpus constitué à Orange

Après avoir collecté ces données, l'étape suivante consiste à les prétraiter en vue de l'étape de pseudo-anonymisation.

B) Prétraitement des données collectées

Après la collecte des données et suite à une rapide analyse visuelle de celles-ci, nous avons pu constater que les emails ne sont pas assez structurés comme nous le pensions en raison de la diversité des formats et styles d'emails ainsi qu'à cause des multiples variations de structure, de présentation des messages d'utilisateurs (Carvalho and Cohen, 2004), mais aussi en fonction des différentes façons qu'ont les interlocuteurs de répondre à des emails ; alors que certains interlocuteurs d'une conversation répondent à un email directement avant ou après celui-ci, d'autres vont répondre de

1. https://docs.microsoft.com/fr-fr/dotnet/desktop/wpf/overview/?view=netdesktop-5.0
2. https://docs.microsoft.com/fr-fr/dotnet/api/microsoft.office.interop.outlook.application?view=outlook-pia

façon explicite à chaque segment de texte ou phrase directement en-dessous (*Lignes de réponse*) de ceux-ci. Les emails contiennent différentes zones qui peuvent être facilement identifiées par l'être humain. (Estival et al., 2007) identifient cinq catégories de zones dans les emails, à savoir : *Texte de l'auteur, signature, publicité ou avis de non responsabilité ou de propriété, messages cités et lignes de réponse*. Trois ans plus tard, (Carvalho and Cohen, 2004) ont affiné et étendu ces catégories à neuf zones parmi lesquelles deux zones attirent particulièrement notre attention :

— **Zones de conversation citées** comprenant à la fois du contenu *cité* en réponse à des messages précédents dans le même fil de conversation et du contenu *transféré* provenant d'autres conversations,

— **Zones statiques** contenant du contenu qui est réutilisé sans modification dans plusieurs messages électroniques. Ces contenus sont les *signatures, publicité ou avis de non responsabilité ou de propriété* et les *pièces jointes*.

Cette étape de prétraitement des données collectées consiste dans un premier temps à supprimer tous les emails en doublon dans l'ensemble des données collectées. Ensuite, nous avons identifié et dissocié les différents contenus des zones de conversation citées et nous avons également identifié et supprimé le contenu des zones d'information. Enfin, pour être en conformité avec le RGPD et respecter les règles de secret de la correspondance, nous avons commencé par remplacer tout ce qui, dans le corps ou dans les objets des emails, pouvait directement ou indirectement conduire à l'identification d'une personne ayant donné son consentement.

Tout d'abord, nous avons des termes de substitution qui conservent la logique sémantique des phrases. Nous avons donc remplacé aux travers de scripts et d'expressions régulières chaque adresse email, numéro de téléphone, lien hypertexte (url/uri), chemin de dossier ou de fichier et identifiant d'utilisateur respectivement par **EMAIL**, **PHONE**, **URL**, **PATH** et **ID**. Cette substitution correspond à une première étape du processus de pseudo-anonymisation. Et ensuite nous avons utilisé la bibliothèque **Talon** [3] inspirée des travaux de recherche de (Carvalho and Cohen, 2004; Joachims, 2001) pour extraire les signatures des messages et les messages cités. Cette bibliothèque fonctionne assez bien pour les emails en français tant pour l'extraction des messages cités que pour les signatures avec respectivement une moyenne de prédictions d'environ 95% et 70%, malgré le fait que le modèle de cette bibliothèque ait été entraîné sur le corpus ENRON d'emails en anglais. Les auteurs de cette bibliothèque ont proposé la possibilité de réentraîner le modèle de classification avec d'autres données. Lors de l'extraction des signatures, nous avons créé une colonne pour celles-ci dans notre fichier CSV au cas où nous aurions besoin d'utiliser certaines connaissances contenues dans les signatures plus tard dans nos travaux. Afin d'améliorer l'extraction des messages cités, nous avons développé des scripts s'appuyant sur les métadonnées qui accompagnent généralement ces messages. La figure 6.C montre l'écart important entre le nombre de tokens d'un email avec et sans ses messages cités. Ces messages cités sont les contenus d'emails précédents au fur et à mesure que la conversation s'enrichit de nouveaux emails. Les informations triviales mentionnées ci-dessus ont été substituées à l'aide d'expressions régulières incluses dans python [4].

Afin d'avoir un corpus exploitable tout en respectant le RGPD et le secret de la correspondance, la dernière partie de notre traitement est un peu plus difficile que les précédentes car elle implique des

3. https://github.com/mailgun/talon
4. https://docs.python.org/3/library/re.html

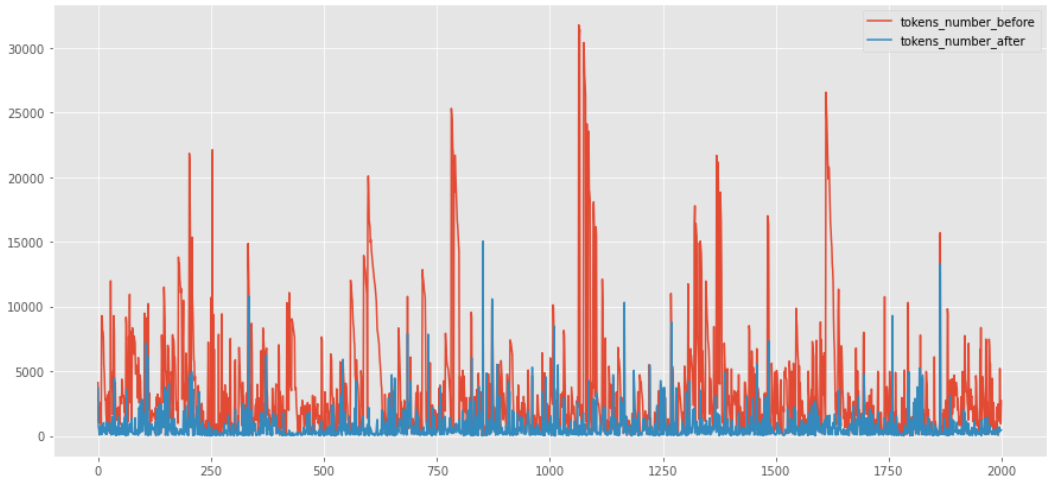


FIGURE 6.C – Nombre de tokens par email avant et après l'extraction des messages cités et des signatures sur un échantillon de 2000 emails

tâches complexes comme **NER** (Named Entities Recognition en anglais, reconnaissance d'entités nommées), la désambiguïsation des noms des collaborateurs, etc. Pour cela, nous avons mené quelques expériences détaillés dans la section 6.2.4.

6.2.4 Pseudo-anonymisation : Méthodes et expériences

Outre les expressions qui permettent d'identifier de façon directe une personne dans le corps des emails, qui sont des chaînes de caractères liées à des collaborateurs, telles que leurs noms, prénoms ou des abréviations de ceux-ci, les contenus dits indirects sont des informations sensibles pour Orange ou appartenant à sa propriété intellectuelle. Les noms de projets ou d'outils, tous types d'identifiants, les noms de groupes et d'entités au sein d'Orange et les chaînes de caractères alphanumériques d'identification unique de conversations et d'emails (fournis par exemple par des applications comme Outlook) font partie de ces données personnelles. Leur identification et leur remplacement ont été réalisés pour se conformer au RGPD et au secret des correspondances. Notons que ce type de mise en conformité a poussé à l'organisation d'un atelier sur le Traitement du Langage Naturel et la pseudo-anonymisation [5]. Cela montre l'importance de rendre les données de recherche librement accessibles. Selon le contexte de l'extraction de données pour la recherche, les données extraites peuvent être disponibles en permanence. C'est le cas par exemple du corpus Enron [6] et du corpus Webis Gmane Email 2019 [7]. Le corpus Enron a été rendu public par la Federal Agency Regulatory Commission à la suite de la faillite de l'ancienne société Enron et une version de ces données a ensuite été achetée par le projet CALO [8], et mise à disposition à des fins de

5. https://aclanthology.org/volumes/W19-65/
6. https://www.kaggle.com/wcukierski/enron-email-dataset
7. https://zenodo.org/record/3766985
8. http://www.ai.sri.com/project/CALO

recherche. Notre cas est différent de celui d'Enron, car il existe aujourd'hui le RGPD et ses règles sont respectées par les entreprises ou organisations concernées.

Dans les données que nous avons collectées, les conversations et emails sont respectivement identifiés par deux chaînes alphanumériques :

— **ConversationID** est l'identifiant d'une conversation. Il est donc le même pour tous les emails de la même conversation,

— **ConversationIndex** est l'identifiant unique de chaque email. Il est constitué de l'identifiant de la conversation à laquelle il appartient (*conversationID*) auquel on ajoute une ou plusieurs autres chaînes de 10 caractères en fonction de la profondeur de l'email dans l'arborescence de la conversation. Cet identifiant réflète donc la relation père-fils puis que le *ConversationIndex* du fils est constitué du *ConversationIndex* du père et d'une sous-chaîne qui lui est propre.

Dans nos travaux, nous avons constaté que ces deux éléments sont toujours présents pour une conversation et ses emails lorsque ceux-ci sont initiés depuis l'outil de messagerie Outlook. Cependant ils sont soit absents ou formatés différemment de ceux d'Outlok selon qu'un email provient d'un autre client ou serveur de messagerie comme Gmail, Mail (sous Mac), Mozilla Thunderbird, etc.

L'approche que nous avons adoptée afin de se conformer au RGPD et au secret de la correspondance a consisté dans un premier temps à remplacer les deux précédentes chaînes de caractères de façon à conserver les emails dans leur conversation respective et toutes les relations père-fils entre les emails et dans un second temps à identifier et à remplacer les chaînes de caractères d'identification directes et indirectes.

L'identification de chaînes directes et indirectes est une problématique générale depuis quelques décennies appelée reconnaissance d'entités nommées (*NER*) dans le domaine du TALN. L'état de l'art pour la reconnaissance d'entités nommées en français a été récemment établi par (Ortiz Suárez et al., 2020) avec des modèles s'appuyant sur les réseaux de neurones de type *Transformers*. Toutefois, certains travaux antérieurs ont été réalisés pour le NER sur des contenus pas ou peu structurés tels que les emails. (Minkov et al., 2005) proposent une méthode basée sur les champs aléatoires conditionnels (Conditional Random Field - CRF en anglais) et combinée à un système à base de règles pour extraire les noms de personnes des emails. (Zhang et al., 2018) utilisent des expressions régulières pour collecter des étiquettes à partir de données bruitées contenant des mentions d'entités nommées pour entraîner un réseau neuronal pour prédire ces étiquettes extraites par les expressions régulières. Ces travaux montrent que les expressions régulières et les systèmes à base de règles restent une approche viable pour la reconnaissance des entités nommées, malgré les bons résultats des modèles récents s'appuyant sur les *Transformers*. Comme nos données collectées ne pouvaient pas être déplacées sur un serveur avec de grandes puissances de calcul afin de finetuner de tels modèles et du fait de la faible taille des données que nous avons annotées, nous avons été contraints d'aborder le problème avec un simple ordinateur avec CPU en combinant plusieurs approches, notamment les CRF, CamemBERT [9] en inférence, les expressions régulières et un système à base de règles, pour atteindre notre objectif d'identification et de substitution des chaînes de caractères directes, indirectes et d'autres sensibles pour Orange afin d'être en conformité avec le RGPD et le secret de la correspondance.

9. https://huggingface.co/Jean-Baptiste/camembert-ner

Nous avons ainsi, pour pseudo-anonymiser nos données, mené quelques expériences en deux étapes décrites ci-dessous : l'annotation de données et la mise en œuvre d'une chaîne de pseudo-anonymisation.

A) Annotation de données

L'élaboration de la vérité de terrain pour les tâches de TALN est une étape nécessaire et correspond à un processus d'annotation généralement coûteux en temps et en argent. Dans le cas du NER, des outils ont été conçus pour accélérer les annotations. Par exemple, il existe :

— **Prodigy** [10], un outil payant d'annotation moderne de création des données d'entraînement pour les modèles d'apprentissage automatique,

— **INCEpTION** [11] une plateforme d'annotation sémantique open source offrant une assistance intelligente et une gestion des connaissances.

Ces outils sont tous deux accessibles via des interfaces web. *Prodigy* quant à lui offre en plus des fonctionnalités en ligne de commande. Ils utilisent tous deux l'*active learning* qui est un concept décrit dans (Ren et al., 2020) comme une méthode visant à sélectionner les échantillons les plus utiles de l'ensemble de données non étiquetées, les fournir à un « oracle » (par exemple, un annotateur humain) pour les étiqueter de manière à réduire le coût de l'étiquetage global autant que possible tout en maintenant de bonnes performances. Comme *Prodigy* n'est pas gratuit et que *INCEpTION* nécessite un temps significatif pour sa prise en main, nous avons décidé d'utiliser un outil accessible via une interface web qui s'appuie sur les CRF et l'*active learning* qui avait été développé par le passé au sein d'Orange dans le cadre d'un stage académique. Dans les 12k emails extraits, nous avons sélectionné 1k emails contenant chacun au moins 150 tokens ceci afin de s'assurer que nous avons assez d'entités représentatives pour chacune des classes à annoter. Trois annotateurs ont effectué l'exercice d'annotation sur 1k emails, chacun avec des étiquettes d'entités nommées que nous avons définies sur la base de ce que nous considérions comme des informations sensibles à substituer dans le corpus d'emails collectés. Ces trois annotateurs font partie des cinq référents auprès de qui nous avons extrait les 12k emails ; ils connaissent bien le contexte de ces emails et sont donc de bons annotateurs pour cette tâche. Le tableau 6.A liste et décrit les étiquettes que nous avons utilisées et qui sont sous le format *BI* avec respectivement *B* pour *begin* qui marque le début (ou permet d'annoter l'unique token d'une entité) d'une chaîne à annoter et *I* pour *intermediate* pour le reste des tokens d'une expression à annoter.

Tags d'entités nommées	**Éléments à annoter**
PERSON (B-, I-)	Prénoms, noms, noms complets et abréviations
ORG (B-, I-)	Entreprises ou organismes connus
SUBGROUP (B-, I-)	Entités et départements à Orange
EMAIL	Adresse emails (au cas où le prétraitement les aurait ignorés)
PHONE (B-, I-)	Numéros de téléphone (au cas où le prétraitement les aurait ignorés)
PROJECT (B-, I-)	Noms des projets à Orange
LOC (B-, I-)	Tout type de lieu (par exemple, adresse, numéro de bureau, salle de réunion)

10. https://prodi.gy
11. https://inception-project.github.io/

ID	Tout identifiant lié à une personne
ROLE (B-, I-)	Fonction ou rôle des personnes (Manager, Responsable de projet, etc.)
UNCERTAIN	Toute entité identifiée qui n'entre pas dans les catégories précédentes

TABLE 6.A – Tags utilisés pour annoter 1k emails

En fonction de ces balises identifiées, un processus d'annotation a été réalisé par trois annotateurs, comme nous l'avons mentionné plus haut. La figure 6.D présente les statistiques des annotations. Nous pouvons constater le très faible taux d'étiquetage *ID*, *ROLE*, *LOC* et *ORG*; ceci est dû à l'étape de prétraitement au cours de laquelle nous avons retiré du contenu des emails les signatures qui contiennent généralement des postes ou fonctions des collaborateurs, des numéros de téléphone, adresses électronique et des adresses.

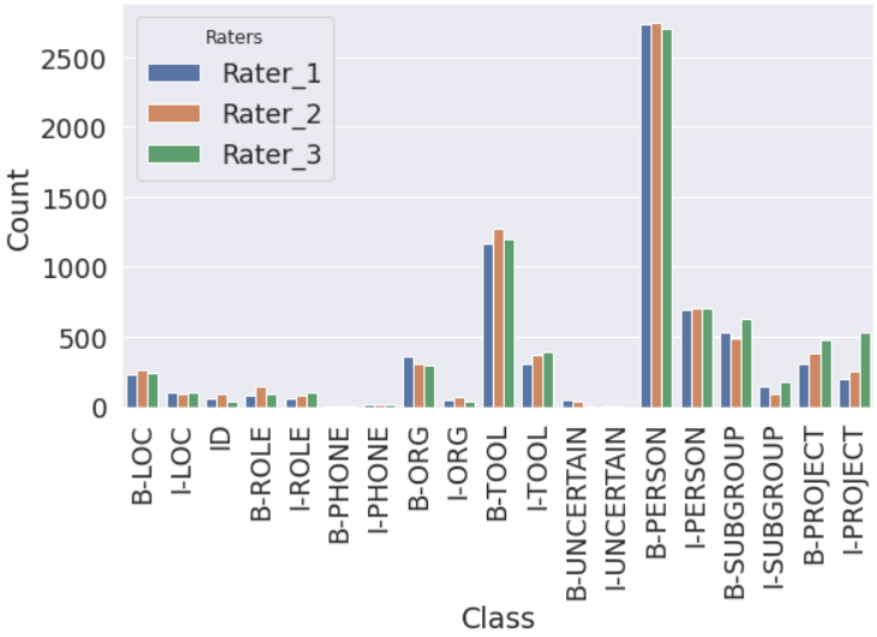


FIGURE 6.D – Statistiques des annotations

L'étiquette *UNCERTAIN* est faible car les évaluateurs connaissent bien le contexte de leurs emails, sinon nous aurions probablement eu un nombre élevé de balises *UNCERTAIN*. Cette connaissance partagée du contexte des emails annotés s'observe également sur plusieurs balises (**LOC**, **ID**, **TOOL**, **PERSON**) qui sont annotés de manière similire par les évaluateurs 1 et 2. Après les personnes et les entités (*SUBGROUP*), les projest(*PROJECT*) et les outils (*TOOL*) ont majoritairement été annotés, ce qui montre une certaine ambiguïté sur les chaînes de caractères liées à ces deux balises. Ceci reflète effectivement la réalité de la difficulté de distinction entre projet et outil, car dans une entreprise, un outil développé dans le cadre d'un projet peut parfois porter le même nom que le projet. Il existe également une ambiguïté entre *SUBGROUP* (qui sont des départements ou des équipes à Orange) et *PROJECT*, car il arrive que le nom d'une équipe soit identique au produit qu'elle développe. Ces disparités montrent que certains concepts de l'entreprise ne sont pas toujours univoques et peuvent faire l'objet de malentendus.

Dans le cadre de l'annotation des données, il existe des mesures permettant d'évaluer les accords entre les différents annotateurs. En général, l'évaluation des accords entre annotateurs se fait à l'aide de la mesure **Kappa de Cohen** (McHugh, 2012) considérée comme une mesure standard dans le domaine biomédical. La mesure Kappa de Cohen n'est pas la plus pertinente pour la *NER*, comme mentionné dans (Hripcsak and Rothschild, 2005). En effet, cette mesure nécessite des cas négatifs (tokens annotés « O ») qui n'existent pas pour la *NER*. En outre, la mesure Kappa de Cohen ne peut pas être utilisée lorsqu'il y a plus de deux annotateurs. Ainsi, comme trois annotateurs ont effectué la tâche sur nos données, nous utilisons le Kappa de Cohen pour évaluer les paires d'annotateurs (R_1, R_2), (R_1, R_3) et (R_2, R_3). Pour évaluer les accords inter-annotateurs (IAA pour Inter-Annotators Agreements) pour plus de deux annotateurs comme dans notre cas, (Zapf et al., 2016) conseillent d'utiliser les mesures **Fleiss's Kappa** et **Krippendorff's alpha** parce ce qu'elles prennent en compte et corrigent les limites du kappa de Cohen. Comme l'ont fait (Brandsen et al., 2020), nous calculons tous les scores IAA dans deux cas avec tous les tokens et uniquement avec les tokens annotés. Le tableau 6.B montre les différents scores calculés, et nous observons que les scores avec tous les tokens sont assez élevés, mais cela est dû au biais des tokens non étiquetés. Les différentes valeurs du Kappa de Cohen sur les tokens annotés sont comprises dans l'intervalle $[0.66 - 0.88]$ qui dénote des accords substantiels et proches de l'idéal (pour ceux supérieurs à 0.80) selon les interprétations de (Viera and Garrett, 2005a), ceci pour trois paires d'annotateurs. Les valeurs calculées pour l'alpha de Krippendorff et le Kappa de Fleiss sont identiques et respectivement égales à 0.85 et 0.71 et dénote des accords substantiels et proches de l'idéal. Sur la base de ces mesures calculées et de leur interprétation, nous pouvons trouver certaines corrélations avec le diagramme de la figure 6.D.

	(R_1,R_2)	(R_1,R_3)	(R_2,R_3)
Cohen's Kappa*	0.8879	0.8450	0.8344
Cohen's Kappa #	0.7832	0.6959	0.6690
Krippendorff's alpha*	0.8554		
Krippendorff's alpha #	0.7155		
Fleiss's Kappa*	0.8554		
Fleiss's Kappa #	0.7158		

TABLE 6.B – Valeurs d'accords Inter-annotateurs sur 1k emails avec les mesures Kappa de Cohen et de Fleiss et la mesure Alpha Krippendorff $R_i estpour for\ Rater_i,\ i \in \{1, 2, 3\}$;
* et # signifient que les mesures sont calculées respectivement avec tous les tokens, inclus ceux marqués « O » d'une part et uniquement les tokens annotés d'autre part

Ces annotations ont permis de constituer un référentiel d'expressions annotées que nous avons combinées avec les résultats de Camembert-ner et un système à base de règles pour pseudo-anonymiser l'ensemble de nos données. Cette combinaison est ce que nous appelons *chaîne de pseudo-anonymisation des données* détaillée dans la prochaine section.

B) Chaîne de pseudo-anonymisation des données

La pseudo-anonymisation des informations sensibles à partir desquelles on peut identifier directement ou indirectement des personnes commence par leur identification par la tâche NER. Pour ce faire, nous avons utilisé *CamemBERT-NER* sur le lot de 1k emails annotés dans l'objectif d'évaluer comment ce modèle se comporte sur les données collectées chez Orange que nous avons annotées (section (A) ci-dessus). La figure 6.E illustre les étapes de notre *chaîne de pseudo-anonymisation des données*

La figure 6.F présente les statistiques du modèle *CamemBERTNER* sur ces 1k emails avec un nombre total de 6952 chaînes identifiées dont 3915 sont classées dans la catégorie divers (MISC). Nous constatons que les résultats donnés par ce modèle sont similaires à ceux obtenus avec des annotations manuelles sur les entités

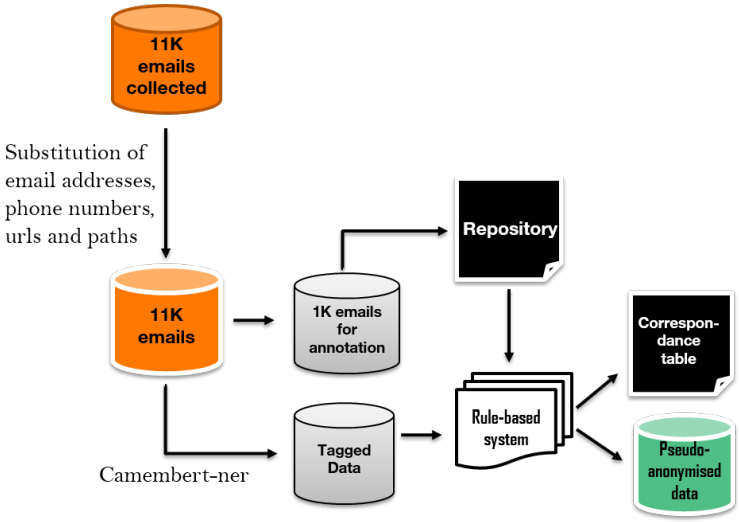


FIGURE 6.E – Chaîne de pseudo-anonymisation des données

personne, *organisation* et *locations*. Il détecte et classe de nombreuses autres entités dans la catégorie *autres* (MISC). Ces résultats sont logiques du fait que *CamemBERT-NER* a été affiné à partir de *CamemBERT*[12] sur l'ensemble de données wikiner-fr (∼ 170 634 phrases) et validé sur des données de emails/chats pour reconnaître des entités personnes, organisations et adresses. On constate aussi qu'il n'arrive pas à bien classifier environ 1500 entités de type personne. Cependant il classifie environ 750 entités comme des organisations.

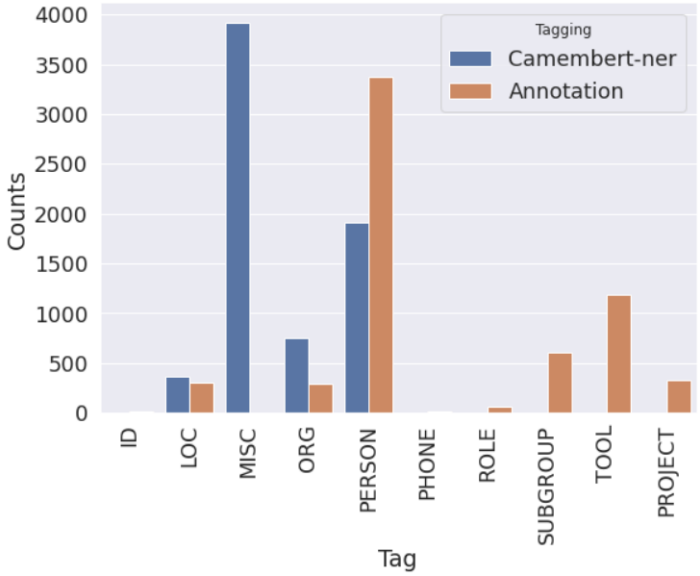


FIGURE 6.F – Nombre d'éléments identifiés et classés par *CamemBERT-NER* et tags identifiés par les annotateurs sur 1K emails

12. camembert-model.fr

À partir de l'étape d'annotation des données (réalisée sur les mêmes 1k emails), il y a 6219 chaînes identifiées qui ont été annotées de manière similaire par les trois annotateurs ou par deux d'entre eux. La différence de 733 entités identifiées par *CamemBERT-NER* par rapport aux annotations manuelles montre sa performance dans l'identification d'éléments qu'il ne peut pas classifier (ceux qu'il range dans la classe *MISC*).

Comme le montre la figure 6.F, le principal défi pour nous va consister à pouvoir répartir les éléments de la classe *MISC* dans les autres classes (*ROLE, SUBGROUP, TOOL, PROJECT*) sur lesquelles CamemBERT-NER n'a pas été finetuné. De même, les organisations qu'il a reconnues sont plus de deux fois supérieures à celles de l'annotation et doivent être reclassées dans d'autres catégories telles que les sous-groupes ou les projets. Des scripts reposant sur un système à base de règles ont été utilisés pour affiner les éléments de la classe *MISC*. Au cours du processus d'annotation, un référentiel a été construit, contenant une liste de chaînes de caractères avec leurs étiquettes respectives (celle du tableau 6.A) sans doublons. Le système à base de règles consiste à comparer les expressions identifiées comme *MISC* par *CamemBERT-NER* avec celles du référentiel d'annotation et à utiliser leur étiquette respective définie dans celui-ci par les annotateurs pour les affiner. Cette étape d'affinage aide le système à base de règles à créer un code sémantique pour la substitution de ces chaînes de caractères à partir du contenu et des sujets des courriels. Les codes créés ressemblent à *user_xxxxxxxx*, *org_xxxxxxxx*, *tool_xxxxxxxx* respectivement pour les collaborateurs identifiés, les organisations, les outils et ce même modèle de codage a été utilisé pour toutes les autres classes d'entités. Lors de la substitution des prénoms ou des noms de famille des collaborateurs, ou des deux dans un email, le système à base de règles vérifie l'expéditeur ou les destinataires de celui-ci afin de conserver le contexte du courrier électronique et du fil de conversation. En cours d'exécution, le système à base de règles construit un nouveau référentiel contenant toutes les paires de chaînes de caractères substituées et leur code respectif. Ce référentiel est appelé **tableau de correspondance** et peut être utilisé ultérieurement pour reconstruire les emails et les sujets sans les codes mais avec les chaînes de caractères qui ont été substituées auparavant. Voici un exemple de paragraphe avant et après la pseudo-anonymisation.

Les espaces de co-working sont plutôt traités dans le ***Research Paper*** *en cours de rédaction par* ***Pierre***, ***Paul***, ***Louise***. *l'étude de* ***Louise*** *(seule) devait porter initialement sur la valeur pour* ***NomEntreprise*** *des tierslieux (fablabs, espaces de coworking) pour ce qui concerne l'apprentissage et la transmission de connaissances à distance. A part le recentrage sur une population interne* ***NomEntreprise***, *l'idée générale est globalement conservée.*

Le paragraphe ci-dessus contient des segments de texte en gras qui sont des noms de collaborateurs et d'entreprises qui ont été remplacés par un code comme dans le texte pseudo-anonymisé ci-dessous.

Les espaces de coworking sont plutôt traités dans le ***misc_55e6a*** *en cours de rédaction par* ***user_8e47d***, ***user_c6f1d***, ***user_5ff59***. *l'étude de* ***user_5ff59*** *(seule) devait porter initialement sur la valeur pour* ***org_252f2a*** *des tierslieux (fablabs, espaces de co-working) pour ce qui concerne l'apprentissage et la transmission de connaissances à distance. A part le recentrage sur une population interne* ***org_252f2a*** *, l'idée générale est globalement conservée.*

6.3 Conclusion

Le processus que nous venons de décrire par plusieurs sous-étapes a permis de collecter des données et de les pseudo-anonymiser, ceci afin d'être conforme au RGPD et de respecter le secret de la correspondance. Toutes ces étapes ont contribué à la production d'un processus de pseudo-anonymisation de données qui a donné lieu à une publication (Tadonfouet Tadjou et al., 2021). Ce même processus peut être utilisé par d'autres entreprises sur leurs données d'emails et ainsi permettre un partage des corpus de données pseudo-anonymisées ainsi générés" pour faciliter la recherche dans certains domaines liés au TALN. Cette chaîne de pseudo-anonymisation est faite de modules qui peuvent être améliorés, notamment le module de *NER* qui peut être remplacé par

un autre modèle affiné sur des données et des catégories d'entités spécifiques à une entreprise tierce et suivant son besoin.

6.4 Autres Corpus

La collecte des données à Orange a pris un temps considérable et a ainsi fortement impacté le planning de nos travaux. La seconde étape dans ce planning consistait à un autre processus d'annotation, mais cette fois orienté sur des segments de texte d'emails qui devaient être classifiés en actes de dialogues d'une part et, d'autre part, la mise en relation de ces segments de texte d'emails de telle façon qu'un segment de texte d'email B réponde à un autre segment de l'email A. Cependant, vu la petite taille de notre corpus d'emails constitué à Orange et la complexité de la tâche d'annotation pour approcher notre problématique de base de constitution de fils de discussions, nous avons opéré le choix d'explorer d'autres corpus. Ce choix a été fait un peu en anticipation pendant la constitution du corpus d'Orange parce que la taille de corpus évoluait très lentement et n'allait probablement pas être suffisante pour la résolution de notre problématique. C'est ainsi que nous avons opté pour l'exploration d'autres corpus de discussions en langue française et par la suite des corpus d'emails et de forum en anglais pour des raisons détaillées dans les prochaines sections. Dans leurs travaux, (Taniguchi et al., 2020) ont annoté en actes de dialogue plus de 2k fils de conversations d'emails extraits 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 à notre connaissance, 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 relations transverses de messages d'emails. Cette absence de corpus avec des annotations spécifiques nous a emmenés à annoter le corpus BC3 pour répondre à notre besoin.

6.4.1 Discussions Wikipédia

Nos travaux s'intéressent à la constitution de sous-fils de conversations à partir de conversations asynchrones en langue française. C'est dans ce sens que nous nous sommes intéressés au corpus de discussions des pages wikipédia en français. Le corpus de discussions des pages de Wikipédia est un corpus disponible en téléchargement libre sur la plateforme ORTOLANG [13] qui est une plateforme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française. Ce corpus a été constitué et rendu disponible au format XML, encodé selon la norme TEI-P5 [14] dans le cadre des travaux de (Ho-Dac, 2021) pour la caractérisation des discussions en ligne.

Ce corpus WikiDiscussion contient 65612 pages de discussion « article » contenant au minimum 2 mots sur les 3,5 millions de pages qui ont été extraites de l'entrepôt de données des pages de discussions de 2015. Les analyses que nous avons effectuées sur ce corpus de discussions Wikipédia n'ont porté que sur un échantillon de 5k pages de discussions, soit environ 7,7% du corpus total. Dans ces 5k pages de discussions, 2865 conversations distinctes ont été extraites. Plusieurs caractéristiques ont ensuite été extraites du corpus, et d'autres ont été calculées. Parmi celles qui ont été extraites, on a : le titre de la page de discussion et la liste des interlocuteurs de ladite

13. www.ortolang.fr
14. https://tei-c.org/guidelines/p5/

page. Concernant les caractéristiques qui ont été calculées, on a : le nombre de posts ou messages par page de discussion (en moyenne 4,45 par page), les tokens les plus fréquents (*article, source, y, non, faire, etc.*), le nombre d'interlocuteurs par page, le nombre d'interlocuteurs anonymes (1.28 en moyenne par page), le nombre de messages envoyés par des bots, etc. Nous avons utilisé un outil de profilage de données [15] sur un échantillon de 1k pages de discussions afin d'y extraire des caractéristiques et informations générales sur ce corpus. Il ressort de ces extractions que les titres de page majoritaires sont les suivants : *Discussions, Avis, Supprimer, Avis non décomptés, Conserver, Fichier proposé à la suppression, liens externes modifiés, Votes, Neutre, Avis divers non décomptés, etc.* Ces titres laissent apparaître que les discussions autour des pages Wikipédia sont fortement à caractère de vote afin de valider ou pas la publication, la suppression ou conservation d'une page (ou article). Ce vote est fait par les interlocuteurs intervenant dans la discussion de ladite page. Une analyse de fréquences de mots en début et fin de discussion (figure 6.7), ainsi que les bigrammes dans les discussions (figure 6.8) mettent en exergue les expressions fréquentes suivantes : *article, supprimer, source, @url (représentant les liens), mettre, référence, section, conserver, etc.* lorsqu'on filtre les *stops-words* ; ce qui laisse transparaître que les discussions sont autour du référencement de certains contenus, de la conservation ou suppression de la page. Cependant dans les bigrammes on retrouve *critères_admissibilité, utiliser_modèle, existence_source, déplacer_supprimer, conserver_fusionner, accentuer_idée, avis_admissibilité, etc.*, ce qui pousse encore à déduire que les discussions tournent autour des publications/conservations/suppressions/mises à jour d'articles ; ceci a du sens parce que les publications d'articles sur Wikipedia doivent avoir des sources fiables et des contenus bien rédigés respectant des modèles prédéfinis. Il est donc tout à fait normal que, dans les discussions, il ressort des votes pour validation ou pas d'articles à publier.

15. ydata-profiling

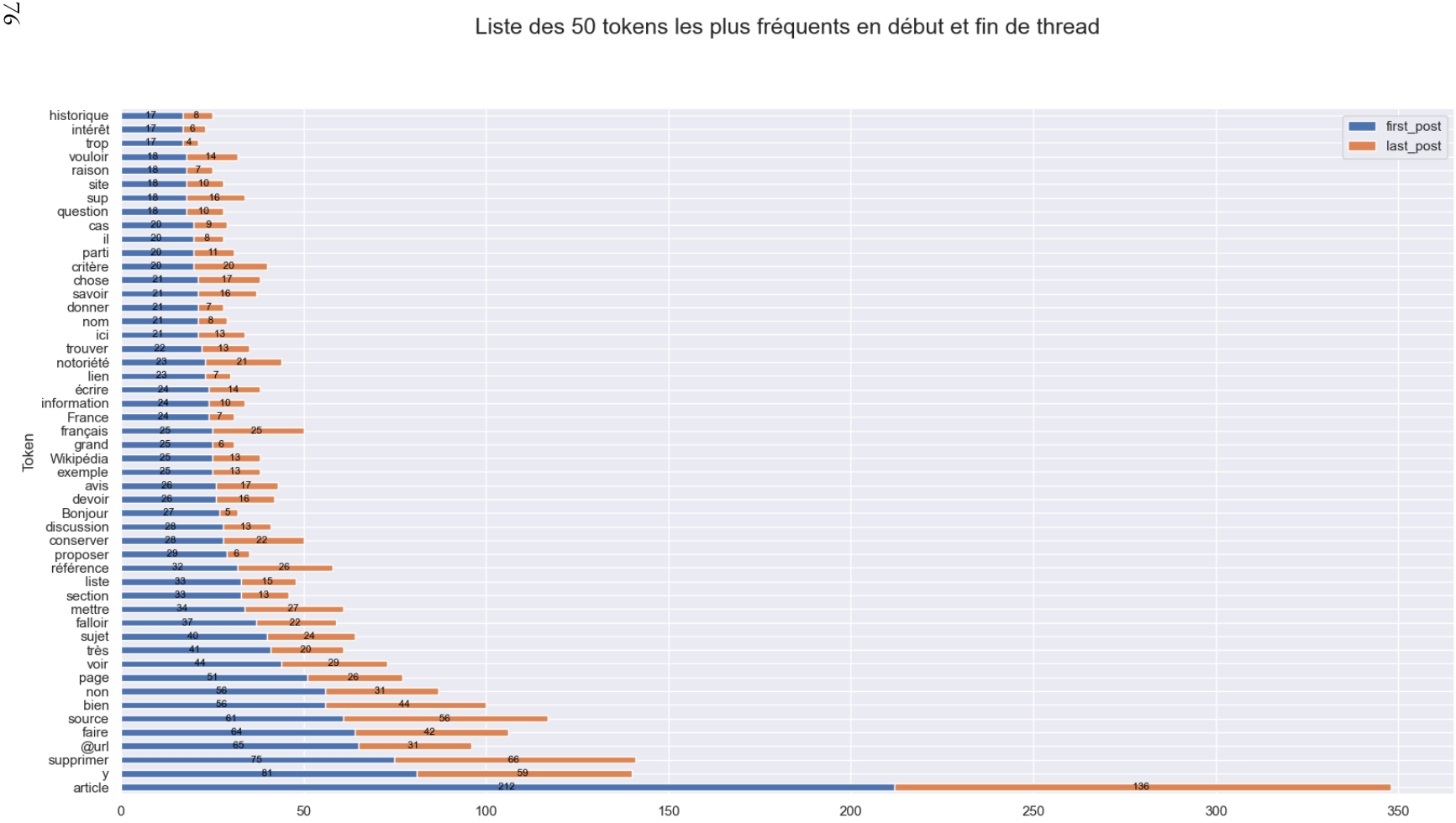


FIGURE 6.7 – Liste des 50 tokens les plus fréquents en début et fin de fil de discussions de pages de Wikipedia

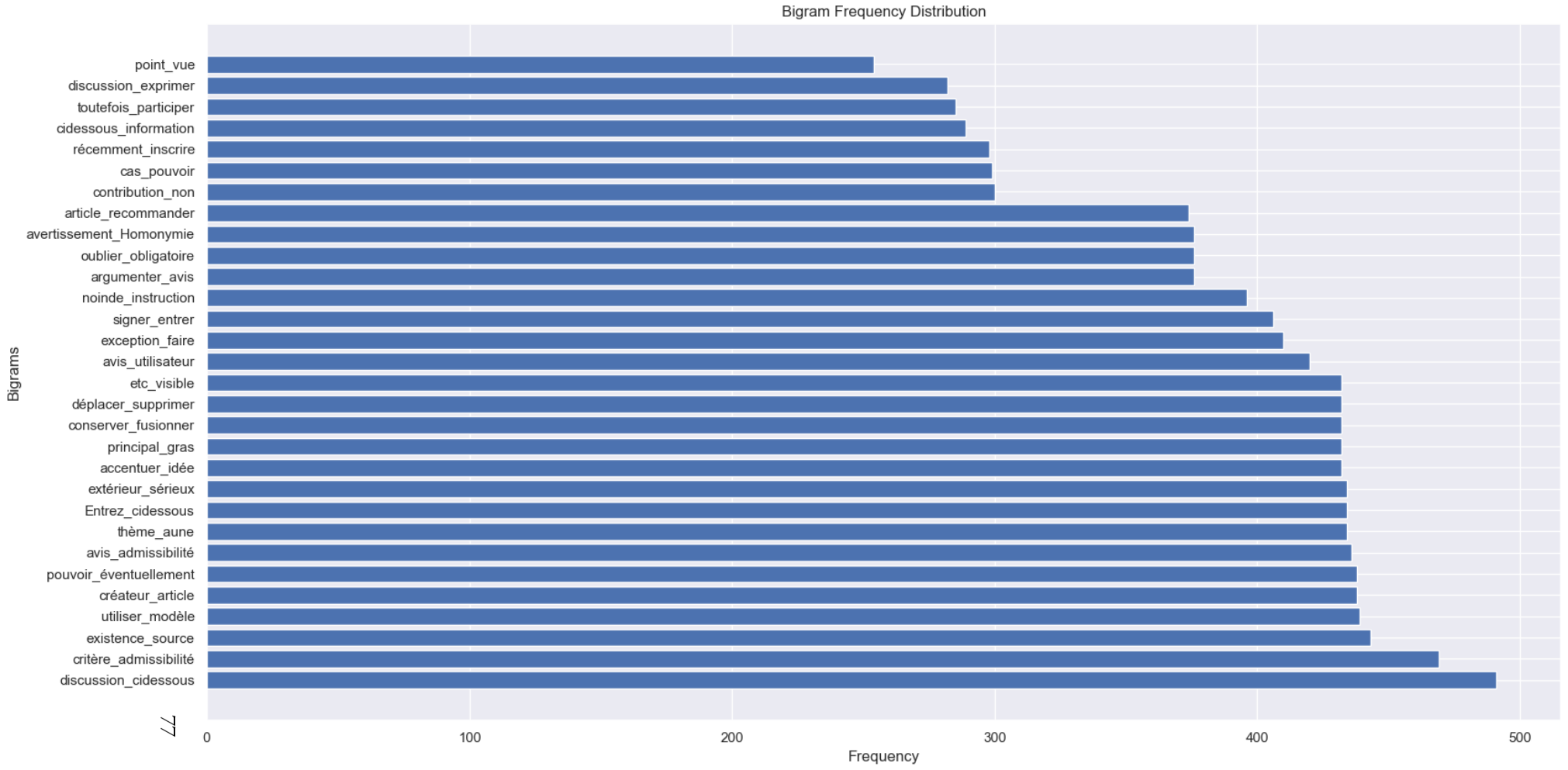


FIGURE 6.8 – Distribution de bigrammes dans les discussions de pages de Wikipedia

Nous avons aussi exploré, en les lisant, quelques discussions d'une vingtaine de pages. Et en plus des aspects de vote décelés, nous avons identifié, mais à très faible fréquence, des *questions/réponses, demande d'actions (portant généralement sur le renseignement de sources ou références), suggestions, acquiescement* qui sont pourtant très intéressants pour aborder notre problématique. Le caractère conversationnel autour du vote des discussions Wikipedia est très peu fréquent, si ce n'est inexistant, dans les conversations d'emails en entreprise. Dans celles-ci, on retrouve généralement des échanges autour du suivi de projets, de demande d'action/d'informations, de partage d'informations, de processus de recrutement, etc. En entreprise, les votes sont en général effectués via des outils spécifiques. Cette caractéristique de vote déduit des discussions des pages Wikipedia nous a emmenés à ne pas analyser davantage, ni utiliser le corpus Wikipedia pour approcher notre problème, mais à nous orienter pour des raisons que nous allons détailler dans les prochaines sections, vers des corpus de conversations d'emails en anglais, BC3 et Enron, de forums tels qu'un extrait de Reddit, et même de transcription de réunion comme MRDA.

6.4.2 BC3

Le corpus BC3 contient des fils de conversations extraits de la liste de diffusion du World Wide Web Consortium (W3C). Les discussions portent sur une variété de sujets tels que l'accessibilité du Web, la planification de réunions pour ne citer que ceux-ci. BC3 se compose de 40 fils de discussion pour 269 emails et 1127 phrases. Il a été initialement construit pour la tâche de résumé de conversations d'emails (Ulrich et al., 2008). Dans le cadre de travaux, des fils de conversations ont été constitués de façon cohérente par trois annotateurs qui ont tout d'abord écrit des résumés de ces fils de conversations, puis ont lié les différentes phrases des emails à celles produites dans les résumés. Dans ces mêmes travaux, les annotateurs ont utilisé les actes de dialogues suivants : *réunion, requête, subjectif, engagement* pour annoter chacune des phrases de chaque email des 40 conversations. (Jeong et al., 2009) ont affiné ceux-ci avec douze nouveaux actes de dialogues (figure 6.3) pour entraîner un classifieur multiclasses des phrases d'emails en actes de dialogues de façon semi-supervisée. Le corpus a ainsi été annoté en actes de dialogue, qui sont des fonctions communicatives intéressantes pour nos travaux puisqu'ils sont un facteur important de compréhension de conversations.

La version du corpus BC3 de (Jeong et al., 2009) est l'un des rares, si ce n'est le seul, corpus de conversations d'emails groupé en conversations indépendantes avec pour chacune, leurs emails segmentés en phrases annotées en acte de dialogues. Nous avons donc choisi d'exploiter ce corpus du fait des différentes couches d'annotations. Les différents travaux que nous avons effectués avec BC3 sont détaillés dans les chapitres suivants :

— 7.5 pour la mise en place d'un référentiel (section 5.3) d'actes de dialogue qui répond mieux à notre problématique et qui s'appuie sur la norme ISO ISO 24617-2 :2020 [16] de gestion des ressources langagières dans le cadre d'annotation sémantique (SemAF) dans sa partie 2 sur les actes de dialogue. Ensuite nous utilisons ledit référentiel pour faire correspondre des actes de dialogue du corpus MRDA (section 6.4.3) et enfin entraîner des modèles de classification d'énoncés ou segments de texte en actes de dialogues.

16. ISO 24617-2:2020

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

TABLE 6.3 – Actes de dialogues du corpus BC3, extrait de (Joty and Mohiuddin, 2018b)

— 8.5 : pour entraîner un classifieur binaire d'appariement d'énoncés ou de segments de texte, ceci après une annotation (effectuée par deux collègues de notre équipe à Orange) qui a consisté à apparier ces segments de texte deux à deux de façon transverses sur les emails de conversations

Vu la petite taille du corpus BC3 qui, comme nous allons le montrer dans ces chapitres, ne permet pas d'avoir des modèles qui permettent une bonne généralisation sur des données qu'ils n'ont jamais vues, nous avons exploré d'autres corpus : Reddit, MRDA et Enron.

6.4.3 MRDA

Le corpus MRDA [17] (*Meeting Recorder Dialogue Act Corpus*) est un corpus de transcriptions d'enregistrements de réunion annotées en actes dialogues avec 3 niveaux de granulatité d'étiquettes : basique (gros grain), général(granularité moyenne) et complet (granularité fine) qui sont respectivement des actes de dialogues de granularité grosse (5 étiquettes), moyenne (12 étiquettes) et fine (52 étiquettes). Ce corpus a été utilisé dans nombre de travaux (Ravi and Kozareva, 2018; Kumar et al., 2018; Raheja and Tetreault, 2019) pour la classification de texte en actes de dialogues, mais aussi pour la modélisation de tours de parole dans les dialogues (He et al., 2021). Nous nous sommes intéressés et avons utilisé ce corpus du fait de sa taille conséquente de 108202 énoncés annotés avec plusieurs (52) actes de dialogue de granularité fine. Lesdits énoncés sont divisés en trous sous-groupes de données : 75067 pour les données d'entraînement, 16702 pour les données de test et 16433 énoncés pour les données de validation.

17. MRDA

Nous avons utilisé ce corpus afin d'entraîner un modèle pour la classification ou reconnaissance d'actes de dialogue sur des énoncés de conversations qui sont dans ce corpus des transcriptions d'enregistrements audio de réunion. Ces transcriptions se rapprochent plus des conversations synchrones, et donc de par les modèles de classification d'énoncés en actes de dialogues que nous avons entraînés avec MRDA, nous avons analysé le transfert de ces modèles sur les conversations asynchrones, plus précisément les conversations d'emails. De par son origine d'échanges audios, les transcriptions de MRDA contiennent de multiples marqueurs de conversations orales tels que *umh, umhumh, you know, so, hummm, etc.* connus sous le nom d'interjections qui sont quasi absentes dans les conversations écrites, surtout dans des emails d'entreprise. Il est fréquent dans MRDA que des énoncés ne soient constitués que de ces interjections, on les retrouve aussi à différentes positions dans des énoncés. Cette présence peut induire une mauvaise compréhension ou bien dégrader la structure sémantique et syntaxique d'un énoncé. La répétition de mots ou d'expressions et l'utilisation très fréquente de certaines formules comme « *you know* » en anglais sont des caractéristiques propres dans les dialogues oraux, qu'on ne retrouve pas nécessairement dans les écrits, excepté les saisies de mots en double. D'autres aspects distinguent les conversations écrites de celles orales comme présentés dans la section 4.3.1.

Nous sommes partis du postulat selon lequel ces caractéristiques propres aux conversations orales et plus précisément des enregistrements de réunion créeront du bruit dans les modèles de classification d'énoncés en acte de dialogues lors de leur utilisation en inférence sur des conversations écrites asynchrones comme les emails. Ce postulat nous a emmenés à filtrer le corpus MRDA en supprimant les énoncés constitués seulement des interjections, mais aussi à supprimer ceux-ci dans les énoncées en plus des mots et expressions doubles, ceci dans le but de nous rapprocher d'une certaine façon de la structure de phrases pouvant être rencontrées dans une conversation écrite.

Dans le chapitre 7.5 sur la classification d'énoncés en actes de dialogue, nous avons utilisé notre propre référentiel 5.3 d'actes de dialogue qui a des correspondances avec ceux de base de MRDA et nous y détaillons nos expériences, ainsi que les résultats et analyses concernant le postulat portant sur les caractéristiques des conversations orales transcrites en énoncés dans MRDA. Bien que les actes de dialogues nous aident dans la résolution de notre problématique, il n'en demeure pas moins que d'autres approches comme l'appariement de segments de texte transverses dans une conversation doivent être explorés et nécessitent l'utilisation de corpus adaptés. Nous avons ainsi porté un intérêt à un sous-ensemble du corpus Reddit présenté dans la prochaine section.

6.4.4 Reddit

Reddit est une plateforme en ligne qui offre à ses utilisateurs un réseau social ainsi qu'un forum de discussion organisé par thématiques, où les internautes peuvent publier des sujets et les classer dans des catégories spécifiques. C'est l'un des plus grands forums de discussion en ligne avec des millions d'utilisateurs actifs dans le monde entier. Ce forum regorge de textes variés qui peuvent être utilisés pour différentes tâches de TALN, telles que l'analyse de sentiment, la classification de textes, la génération de textes et l'extraction d'informations. Différentes variantes de corpus extraits de la plateforme Reddit ont été utilisées pour approcher les problématiques de résumé de texte (Kim et al., 2019), d'analyse et de génération de réponses et questions dans les conversations (Li et al., 2016; Gupta et al., 2022), de détection d'anxiété (Shen and Rudzicz, 2017), de sarcasme (Khodak et al., 2018), mais aussi pour des études de phénomènes impactant la société et comment

les internautes réagissent à ceux-ci (Baumgartner et al., 2020; Monti et al., 2023; Engel et al., 2022), certains de ces phénomènes étant modélisés sous forme de phrases.

Nous avons aussi utilisé une version du corpus Reddit nommée « *Coarse Discourse* » [18] constituée lors des travaux de (Zhang et al., 2017) dans le cadre d'une caractérisation spécifique de discussions de forum. Dans ces travaux, ils ont fait annoter le corpus par trois personnes, en considérant deux niveaux :

— **niveau énoncé :** chacun des annotateurs a classifié avec au moins un acte de dialogue chaque post ou message dans une conversation. Les actes de dialogues suivants sont ceux utilisés pour cette annotation : *Question & Request, Answer, Announcement, Agreement, Appreciation & Positive Reaction, Disagreement, Negative Reaction, Elaboration & FYI, Humor, Other*. Pour évaluer l'accord inter-annotateur sur cette couche d'annotation, l'indicateur Alpha de Krippendorff a été utilisé avec une valeur pondérée sur tous les actes de dialogue égale à 0.645 qui s'interprète comme un accord substantiel. Le tableau ci-dessous récapitule les valeurs du coefficient Alpha de Krippendorff pour chacun des actes de dialogue.

— au **niveau conversation**, chaque annotateur a défini la relation « *répond à* » entre tous les messages d'une même conversation excepté le premier message qui initie celle-ci. Cette mise en relation des commentaires ou posts dans une relation a permis aux auteurs de dresser ci-dessous le tableau de distribution des relations les plus fréquentes entre actes de dialogues.

L'annotation de cette distribution à deux niveaux est la principale raison pour laquelle nous avons décidé d'utiliser ce corpus dans le cadre de nos travaux. Cette variante de Reddit est constituée de 9483 fils de conversation pour 115827 énoncés avec 63573 interlocuteurs. Notons que la distribution officielle du jeu de données fourni avec l'article ne contient que des identifiants de commentaires/publications, et non leur contenu textuel; le jeu de données est également accompagné d'un script permettant de lier les identifiants avec leur texte correspondant en utilisant l'API Reddit. Nous avons utilisé ces scripts et identifiants afin de récupérer les contenus textuels du corpus que nous utiliserons plus tard dans nos expériences. Les actes de dialogue utilisés pour annoter les énoncés de ce corpus ne sont pas exhaustifs et de ce fait ne sont pas assez explicites pour des énoncés qui seraient, par exemple, des suggestions, des instructions, des offres ou des promesses. En plus de cette limite nous avons constaté, après avoir récupéré les contenus des messages ou posts, que certains parmi ceux-ci étaient constitués de plusieurs phrases et ainsi annotés en plusieurs actes de dialogues. Pour ce type de commentaires, il est difficile de pouvoir déterminer quelles sont les phrases ou segments de textes correspondant respectivement aux actes de dialogues, encore faut-il que les actes de dialogues définis par les annotateurs soient les mêmes. Pour remédier à une telle ambiguïté, nous avons filtré des conversations avec une moyenne de phrases par commentaire inférieure ou égale à 2, ce qui a réduit la taille des données que nous avons utilisées.

Une seconde hypothèse que nous posons avec l'utilisation de ce corpus est le transfert de la mise en relation (« *répond à* ») des messages dans les conversations de forum (Reddit) sur les énoncés de conversation d'emails de façon transverse. Pour ce faire, dans le chapitre 8.5, nous entraînons des modèles d'appariement d'énoncés de conversations et y présentons les résultats

18. coarse-discourse

6.4.5 Enron

Le corpus d'emails Enron est l'un des corpus les plus utilisés lorsqu'il est question d'analyse d'emails et de conversations asynchrones. Cependant pour la problématique de reconnaissance d'actes de dialogue dans des conversations écrites, il n'a pas souvent été utilisé, exceptés les travaux de (Taniguchi et al., 2020) dont nous avons fait mention dans la section 6.4 et pour lequel le corpus n'est pas disponible. Nous avons cependant identifié les travaux de (Jamison and Gurevych, 2013) qui ont extrait 70178 fils de conversation du corpus Enron et les ont rendus disponibles. Dans leurs travaux, chaque conversation a été construite avec des emails de la conversation d'origine mais en séparant les messages cités dans chacun de ces emails. Dans leur papier, ils dressent le tableau ci-dessous 6.9 de la distribution des conversations dans leur corpus.

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

Figure 6.9 – Nombre des conversations dans le Corpus de Threads Enron.

Nous avons utilisé cette version du corpus Enron pour entraîner un modèle de prédiction de mail suivant c'est-à-dire étant deux emails A et B, le modèle prédit si l'email B suit A dans une conversation. Nous nous sommes intéressés à cette tâche afin de pouvoir l'appliquer sur notre problème de mise en relation transverse de segments de texte d'emails dans une conversation.

6.5 Conclusion

Dans ce chapitre nous avons présenté les différents corpus que nous avons utilisés pendant nos travaux de thèse pour notre problématique de constitution de sous-fils de conversations cohérents à partir de conversations issues d'outils de communication et de conversation en entreprise.

Nous avons tout d'abord détaillé un processus méthodologique de constitution de corpus en entreprise, qui peut être appliqué par une entreprise ou organisation tierce pour la constitution

d'un corpus spécifique. Ledit processus contient différents aspects et étapes qui ont permis la création d'un pipeline de pseudo-anonymisation des données. Un tel pipeline est une réponse aux contraintes induites par le RGPD et la conformité au secret de la correspondance. Le processus que nous avons décrit se compose de plusieurs étapes : premièrement, nous avons sollicité nos collaborateurs afin d'obtenir leur accord de consentement; deuxièmement, nous avons collecté des emails à partir des boîtes de messagerie Outlook, ces emails ont ensuite été prétraités; la troisième étape a porté sur l'annotation manuelle et la reconnaissance d'entités nommées; et enfin nous avons procédé à la pseudo-anonymisation des données. Au cours de la deuxième étape, nous avons développé un outil appelé **OutlookScrapping** qui nous a permis de collecter 12k emails sur les postes de travail de 5 collaborateurs parmi les 122 collaborateurs qui ont donné leur accord pour l'utilisation de leurs données personnelles. Ce processus de constitution de corpus et de pseudo-anonymisation a donné lieu à une publication (Tadonfouet Tadjou et al., 2021) à l'atelier étudiant de la conférence RANLP 2021. La constitution de ce corpus a nécessité beaucoup de temps, et son annotation pour approcher notre problématique prendrait davantage de temps. Nous avons ainsi fait le choix d'explorer d'autres corpus déjà annotés afin de pleinement débuter la résolution de notre problématique.

Dans la seconde partie de ce chapitre, nous avons présenté d'autres corpus que nous avons explorés pour certains et exploités pour les autres. Le corpus de discussion de pages Wikipédia en français fait partie des premiers corpus que nous avons analysés et nous avons décidé de l'abandonner au profit d'autres, parce que les résultats de nos analyses ont montré que ses conversations étaient très fortement liées à des processus de vote, ce qui est très distant des thématiques que l'on peut rencontrer dans des conversations d'emails en entreprise. Nous avons alors sélectionné deux corpus d'emails, BC3 et Enron, le corpus de discussions du forum REddit et, enfin, MRDA, un corpus de transcriptions d'enregistrement de réunion. Ces différents corpus ont été choisis parce qu'ils étaient annotés pour Reddit en relation « *répond à* » entre les messages des conversations et pour les autres en actes de dialogues. Ces différents prismes d'annotation sont très importants dans l'analyse conversationnelle et ainsi pour notre problématique. Dans les prochains chapitres nous présentons les tâches de classification d'énoncés de conversation en actes de dialogues, de prédiction d'email suivant et d'appariement d'énoncés dans une conversation qui concourent à la résolution de notre problématique.

7 Reconnaissance d'Actes de Dialogue (ADs) dans les CMO

7.1 Introduction

Un acte de dialogue (AD) est un élément fondamental de la structure d'un échange verbal ou écrit entre deux personnes ou plus. Il s'agit d'une unité de sens qui permet de représenter les différentes intentions ou actions communicatives des interlocuteurs. Chaque AD peut être considéré comme une étape ou une unité de base dans le processus de communication.

Nous nous intéressons aux ADs dans nos travaux parce que les conversations d'emails professionnelles, en plus de contenir des connaissances dans différents domaines, sont des communications entre deux ou plusieurs interlocuteurs cherchant en général à atteindre un objectif. Ces communications peuvent donc être considérées comme des dialogues dans lesquels certains énoncés d'emails sont des suites ou « *follow-up* » d'autres énoncés d'emails les précédant. Par exemple dans un email A, deux énoncés ou segments de texte qui ont pour ADs respectifs « question » et « suggestion » peuvent avoir dans un email B ou C de la même conversation des énoncés qui leur répondent respectivement et qui sont de type « réponse » et « approbation ». Ceci montre l'importance des ADs dans la résolution de notre problématique de constitution de sous-fils de conversations d'emails.

Les ADs sont généralement classifiés selon la théorie des actes de langage développée par le philosophe du langage John Searle. Selon cette théorie, chaque AD peut être catégorisé en fonction de son intention communicative, c'est-à-dire de ce que le locuteur cherche à accomplir en prononçant ou écrivant une phrase dans une conversation. Par exemple, certains ADs peuvent être des actes assertifs (exprimer une croyance ou une affirmation), des actes directifs (donner un ordre ou une directive), des actes expressifs (exprimer des émotions) ou des actes déclaratifs (déclarer quelque chose officiellement). Dans le chapitre 5, nous avons détaillé l'évolution de la théorie des actes de langage, ainsi que les différentes taxonomies et schémas d'annotation les plus utilisés pour labelliser des énoncés de conversations écrites ou parlées.

Dans ce chapitre, nous présentons les expériences effectuées en termes de classification d'énoncés de conversations en ADs, mais avant de détailler ces expériences, nous présentons les corpus utilisés.

7.2 Corpus utilisés

Il existe quelques travaux d'identification d'ADs dans les conversations d'emails comme ceux de (Taniguchi et al., 2020) qui ont annoté en ADs plus de 2k fils de conversations du corpus Enron avec deux granularités différentes : 35k phrases annotées avec une granularité fine et 6k emails

annotés avec une granularité moins fine. Cependant ce corpus d'emails annotés n'est pas disponible et à notre connaissance, excepté les travaux de (Jeong et al., 2009) pour la tâche de classification des phrases d'emails en ADs ; à notre connaissance, il n'existe pas de corpus d'emails finement annoté en ADs et en relations transverses entre phrases d'emails, ce que nous avons effectué avec le corpus BC3.

Dans le cadre de leurs travaux de création de résumé d'emails, (Ulrich et al., 2008) ont annoté en ADs les phrases d'emails de *BC3* avec les actes suivants : *Propose, Request, Commit, Agreement/Disagreement, Meeting et Subjective*. (Jeong et al., 2009) ont fait correspondre ces ADs à une liste de 12 actes listés dans la table 6.3 qui ne font pas partie de la norme *ISO 24617-2*. Ce remapping a été fait par deux annotateurs avec un accord inter-annotateur égal à $\mathbf{0.79}$. Le fort pourcentage de l'AD *Statement* nous a amené à inspecter ce corpus BC3 et à constater que cette classe peut être décomposée en d'autres ADs de la norme. Nous avons ainsi construit un référentiel (cf. section 5.3) d'AD extrait des fonctions communicatives de la norme *ISO 24617-2*. Nous avons ensuite utilisé ce nouveau référentiel pour ré-annoter BC3 pour approcher la problématique de classification en ADs d'énoncés d'emails. Néanmoins avec le corpus BC3 est sa petite taille, 40 conversations pour 269 emails et 1127 phrases. Ce manque de données nous a orienté vers un autre corpus annoté en ADs , à savoir le corpus de transcription d'enregistrements de réunions *MRDA*

MRDA est à la base un corpus d'échanges oraux lors de réunions et on y retrouve donc (ou en particulier) des interjections qui n'existent pas dans des échanges écrits. En plus d'être des énoncés, ces marqueurs de communications orales se retrouvent dans des énoncés et pourraient constituer du bruit dans la prédiction des ADs, notamment lors du transfert sur des conversations écrites. Ainsi, nous avons nettoyé *MRDA* en supprimant ces marqueurs de paroles lorsque retrouvés dans les énoncés, le plus souvent en début ou à la fin de ceux-ci. Les détails de nettoyage effectué sur *MRDA* sont présentés dans la section 6.4.3.

MRDA et *BC3* sont au final les 2 corpus que nous utilisons pour la classification d'énoncés en ADs.

7.2.1 Annotation et statistiques

En annexe B, les détails du corpus MRDA avec ses différentes catégories (basique, générale et fine) d'ADs sont présentés tels que décrits dans les travaux de (Shriberg et al., 2004). Ces détails sont extraits de la page *github* du corpus. [1]. Dans la section 5.3, nous détaillons les correspondances que nous avons effectuées sur les ADs de MRDA pour constituer notre référentiel. La figure 7.1 illustre la répartition des ADs de notre référentiel sur le corpus MRDA. Notons que lors du processus de correspondance, certains énoncés, en majorité les interjections, ont été supprimés. Suite à ces suppressions, la nouvelle version de MRDA nommée **Filtered_MRDA** avec nos ADs a une taille La figure 7.1 illustre la répartition des ADs de notre référentiel sur le corpus MRDA. Notons que lors du processus de correspondance, certains énoncés, en majorité les interjections, ont été supprimés. Suite à ces suppressions, la nouvelle version de MRDA nommée *Filtered_MRDA* avec nos ADs a une taille réduite de 50% en moins par rapport au corpus d'origine. Les données de *Filtered_MRDA* sont réparties comme suit : 36.960 de données d'entraînement, 8.021 de données

1. Corpus MRDA

de validation et 7.964 de données de test. Les données de *Filtered_MRDA* sont réparties comme suit : 36.960 en données d'entraînement, 8.021 en données de validation et 7.964 en données de test.

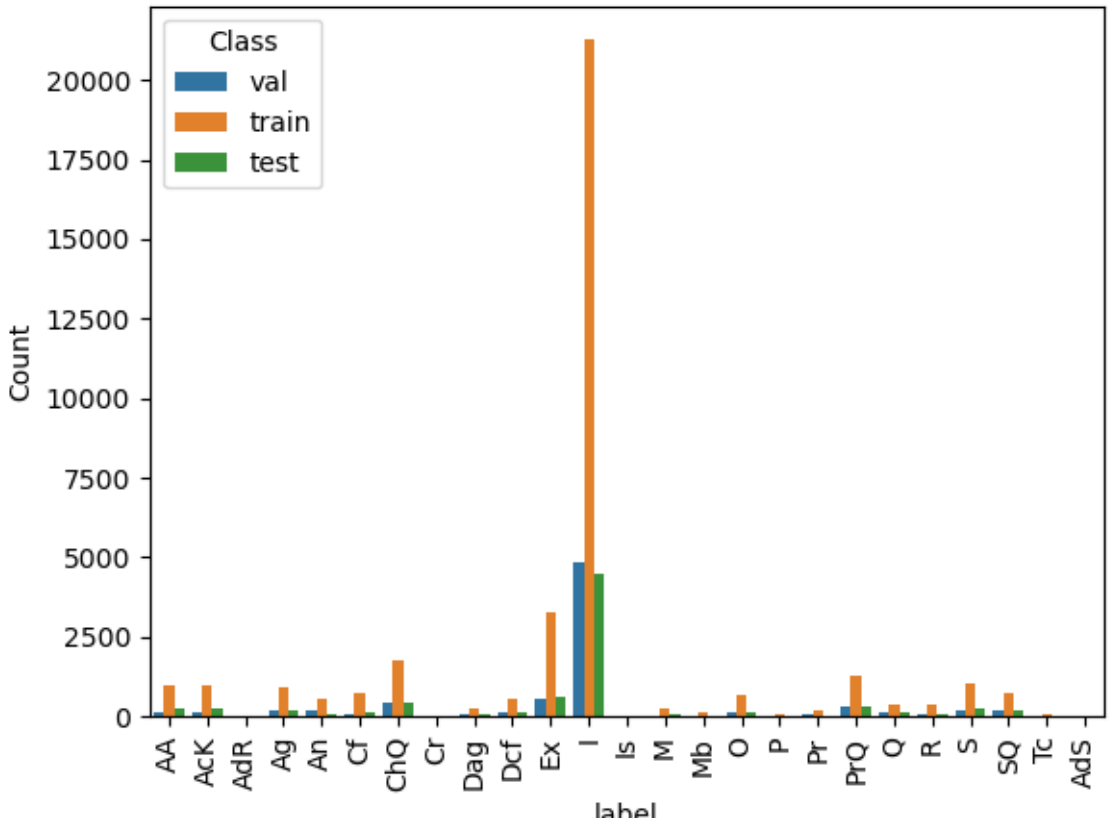


FIGURE 7.1 – Distribution des ADs de notre référentiel sur le corpus MRDA

7.3 Classification en ADs

7.3.1 Architecture de nos modèles

A. [BERT + Bi-LSTM] + Att

Nous avons affiné *BERT* (**Bidirectional Encoder Representations from Transformers**) (Devlin et al., 2018) avec différentes stratégies pour notre tâche de classification d'énoncés en ADs ou d'identification d'Ads. Nous avons choisi cette approche d'affinage de *BERT*

parce que ce modèle de langage pré-entraîné a montré son efficacité sur les tâches en aval. Cette efficacité est due à la richesse des représentations contextuelles produites par BERT. Notre architecture utilise comme base le modèle pré-entraîné BERT avec ses couches gelées auquel nous ajoutons une couche *BiLSTM* (**Bi-directional Long Short Term Memory**), constituant ainsi la base de notre architecture. Cette architecture de base **BERT+BiLSTM** a été utilisée dans de nombreux travaux pour la résolution de certaines tâches dans le domaine du Traitement Automatique de Langage (TAL). Par exemple (Li et al., 2019b) l'utilisent pour extraire davantage des caractéristiques pour certaines tâches de classification automatique de texte ; (Liu et al., 2019) s'appuient aussi sur cette architecture de base pour l'identification des intentions.

Pour certains modèles que nous avons entraînés, nous rajoutons une couche d'**attention** qui permet de pondérer l'importance des informations contextuelles produites par la couche **Bi-LSTM** et de se concentrer sur les parties les plus pertinentes de la séquence à classifier. Elle permet aussi d'extraire des informations sur les relations entre les mots et d'identifier les mots-clés ou les parties importantes de la séquence. L'utilisation de cette couche d'attention a montré son efficacité dans les travaux de (Aboutaleb et al., 2021) et (Xu et al., 2023) respectivement pour les tâches de similarité sémantique et de construction de graphes de connaissances.

Introduits dans cette architecture, nos énoncés à classifier deviennent des représentations vectorielles que nous passons à notre dernière couche ou couche de sortie qui est généralement une couche dense suivie d'une fonction d'activation appropriée selon la tâche qui est traitée. Nous traitons une tâche de classification multiclasses pour laquelle la fonction **Softmax** est la plus appropriée. La couche dense produit des activations pour chaque classe possible, et la fonction **Softmax** convertit ces activations en probabilités pour chaque classe.

7.3.2 Protocole d'entraînement de nos modèles

BERT+BiLSTM est notre modèle de base auquel nous avons ajouté ou pas une couche d'attention. Dans nos expériences, on le nomme **default**.

— **Données**

Les inputs passés à ces modèles sont de deux ordres : dans un premier temps les énoncés à classifier sont pris indépendamment les uns des autres et, dans un second temps, un contexte est ajouté à l'énoncé à classifier. Ce contexte n'est rien d'autre que le précédent énoncé dans les transcriptions. Les modèles qui utilisent les entrées avec contexte ont leur nom préfixé par **Ctx**. Certains modèles entraînés ont comme entrée des énoncés regroupés, c'est-à-dire que deux ou plusieurs énoncés consécutifs d'un même interlocuteur qui ont le même AD sont regroupés en un seul énoncé. Ce regroupement donne lieu à des variantes de nos modèles dont les noms commencent par **Grp**.

Comme mentionné à la section 7.2.1, nous utilisons une version filtrée de *MRDA* nommée **Filtered_MRDA** qui a une taille réduite de moitié par rapport à celle du corpus d'origine. Les données de *Filtered_MRDA* sont réparties comme suit : 36.960 données d'entraînement, 8.021 données de validation et 7.964 données de test.

— **Variantes de nos modèles**
L'ajout d'une couche d'attention sur notre architecture de base (**default**) donne un nouveau modèle appelé **Att**. La variante nommée par exemple **Gpr+Ctx** correspond au modèle sans couche d'attention avec, comme entrées, les énoncés groupés avec la prise en compte du contexte. Les noms **Grp** et **Ctx** désignent les modèles avec l'architecture de base avec respectivement, comme entrées, les énoncés groupés et ceux avec leurs contextes respectifs.

— **Hyper-paramètres**
La couche **BiLSTM** ajoutée à *BERT* est constituée de **64** unités ou neurones, avec une taille de plongements de mots de 128. Comme optimiseur, nous avons utilisé l'algorithme Adam (*adaptive moment estimation*) (Kingma and Ba, 2017) qui est une extension de la descente de gradient stochastique qui est largement utilisée pour l'entraînement des modèles de tâches de vision par ordinateur ou l'entraînement du langage naturel. Nous utilisons l'entropie croisée catégorielle (**categorical_crossentropy**)

$$\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

comme fonction de perte, qui est aussi couramment utilisée dans les tâches de classification multiclasses en apprentissage automatique. Elle est spécifiquement conçue pour mesurer la différence entre la distribution de probabilité prédite par un modèle et la distribution de probabilité réelle des étiquettes de classe.

B. Utilisation des CRF.
Les **CRF** (**Conditional Random Fields** (Kumar et al., 2017; Joty and Mohiuddin, 2018a), champs aléatoires conditionnels) sont généralement utilisés pour modéliser les dépendances entre des étiquettes consécutives dans une séquence, en tenant compte du contexte et en améliorant les performances globales d'étiquetage de séquence. Nous avons mené d'autres expériences de classification d'énoncés du corpus MRDA en nos actes de dialogue en rajoutant une couche CRF après une couche *Bi-LSTM*. Celle-ci prend en entrées des encodages d'énoncés issus de deux types de représentations :

— des plongements issus de modèles pré-entraînés à l'instar de *BERT* de la même façon que dans notre architecture de base *BERT+Bi-LSTM* précédemment décrite.

— des plongements issus de **ConceptNet Numberbatch** (Speer and Lowry-Duda, 2017) :
ConceptNet est un réseau sémantique open source qui représente des mots (et des séquences courtes de mots communément vus ensemble) comme des nœuds et des arêtes qui représentent les relations entre ces nœuds. *ConceptNet Numberbatch* est un type de plongement de mots enrichis qui résulte de la combinaison de *ConceptNet* avec d'autres représentations de mots tels que **word2Vec** ou **Glove** via le procédé de **Retrofitting**. Ce dernier consiste à améliorer des plongements de mots en incorporant des connaissances ou des informations externes dans les plongements existants. Nous avons utilisé *ConceptNet Numberbatch* sur les données de *MRDA* en nous appuyant

sur les travaux de Jonas Scholz [2]. La figure 7.2 illustre l'architecture utilisée avec la prise en compte des CRF, cette architecture prend en entrées N énoncés u_i. Nous avons effectué nos expériences avec différentes valeurs de $N = \{5, 10, 50, 100\}$. Dans cette

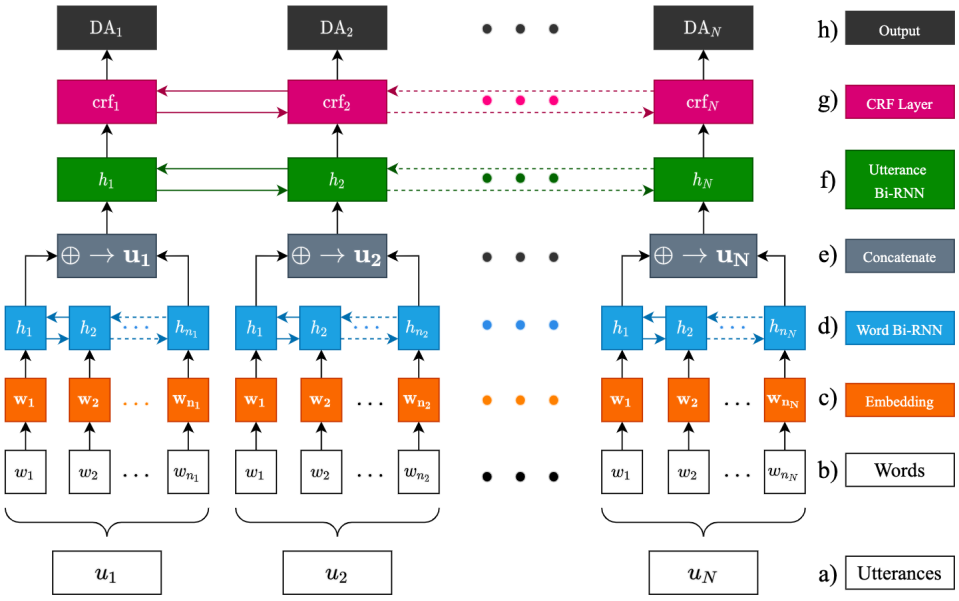


FIGURE 7.2 – Architecture utilisée pour l'identification des ADs (DA_i (Dialogue Acts)) avec une couche CRF, extrait des travaux Jonas Scholz.

architecture, les couches *c), d) et e)* sont remplacées par les plongements de mots de **BERT** lorsque nous les utilisons pour la représentation de nos mots, sinon les couches restent les mêmes qu'avec l'utilisation de *ConceptNet Numberbatch*(*CoNNb*).

Pendant l'entraînement, la log-vraisemblance négative (formule 7.1) est la fonction de perte calculée sur la base des séquences d'étiquettes prédites par rapport aux séquences d'étiquettes réelles. L'objectif est de minimiser cette perte, ce qui maximise essentiellement la probabilité des vraies séquences d'étiquettes compte tenu des données d'entrée et des paramètres du modèle.

$$l(\theta) = -\sum_{i=1}^{n} \Big(y_i \, log \, \hat{y}_{\theta,i} \; + \; (1 \; - \; y_i) \, log \, (1 \; - \; \hat{y}_{\theta,i}) \Big) \tag{7.1}$$

La log-vraisemblance négative tient compte à la fois des probabilités d'étiquettes prédites à partir de la couche CRF et des scores de transition entre les étiquettes. Elle pénalise les prédictions d'étiquettes incorrectes et attribue des valeurs de perte plus élevées pour les séquences d'étiquettes moins probables. Cette fonction de perte est un choix courant lors de l'utilisation d'une couche CRF comme couche finale au-dessus d'une architecture *BERT* pour les tâches d'étiquetage de séquence.

2. State of the Art Natural Language Processing Methods For Conversation Analysis

Actes de dialogue
Nous avons entraîné nos modèles avec trois niveaux de granularité d'ADs :

— **fine** avec 24 ADs : *['AA', 'S', 'AcK', 'Hy', 'Cr', 'P', 'O', 'R', 'Dcf', 'Q', 'Ex', 'An', 'Dag', 'Cf', 'AdR', 'Misc', 'ChQ', 'PrQ', 'SQ', 'Ag', 'I', 'Pr', 'Is', 'AdS']*

— **moyenne** avec 18 ADs : *['AdR', 'R', 'S', 'Q', 'AcK', 'Inform', 'Dag', 'Pr', 'P', 'An', 'Ex', 'O', 'Ag', 'Misc', 'Hy', 'AdS', 'Is', 'AA']*

— **basique** avec 9 ADs :*['Inform', 'Ex', 'Offer', 'Feedback', 'Hy', 'Q', 'R', 'S', 'P']*

Ces différents niveaux suivent en majorité la structure arborescente des fonctions communicatives de la norme *ISO 24617-2*. Dans Le tableau 5.2, on retrouve les noms complets de chacun de ces ADs.

Lors d'une première phase dans nos expériences, 6 modèles avec pour base **BERT+BiLSTM** (sans les CRF) ont été entraînés avec différentes variantes : avec ou sans couche d'attention, des énoncés en entrée groupés ou pas et avec ou sans contexte. Dans une seconde phase incluant les CRF, nous utilisons une approche de recherche aléatoire de la librairie *Weights & Biases* qui permet de configurer différents paramètres et hyperparamètres et de suivre les hyperparamètres, métriques système et prédictions de nos modèles afin de les comparer en direct. Elle permet aussi de facilement partager les résultats. Cette librairie a ainsi permis d'avoir différentes configurations de nos modèles avec des couches rajoutées (BiLSTM, attention et CRF), des données : énoncés simples, groupés ou pas, avec ou sans contexte et les 3 niveaux de nos ADs. Notons que lors de cette seconde phase, seuls les plongements de mots de BERT ont été utilisés.

Pour l'implémentation de nos expériences, **Tensorflow** a été utilisé avec une version de base de *BERT*[3] sans casse téléchargée depuis le Tensorflow hub. Lors de l'entraînement de nos modèles, nous avons mis en place l'arrêt précoce (*Early Stopping*) qui est une forme de régularisation utilisée pour éviter le sur-apprentissage. Nous l'avons configuré sur la performance de modèle sur le paramètre *monitor* et le paramètre *patience* est fixé à 3, qui est le nombre d'epochs après lesquelles l'entraînement sera arrêté sans amélioration de performance.

7.4 Résultats, Analyses et Évaluations

Comme mentionné précédemment, deux phases constituent nos expériences.

7.4.1 Phase 1

Au cours de celle-ci, nous entraînons 6 modèles comme décrit dans la section 7.3.2 de protocole d'entraînement. La figure 7.3 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 **Att** qui a comme base **BERT+BiLSTM** plus une couche d'auto-attention présente une meilleure performance au bout de 2 epochs, contrairement aux autres combinaisons.

3. https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4

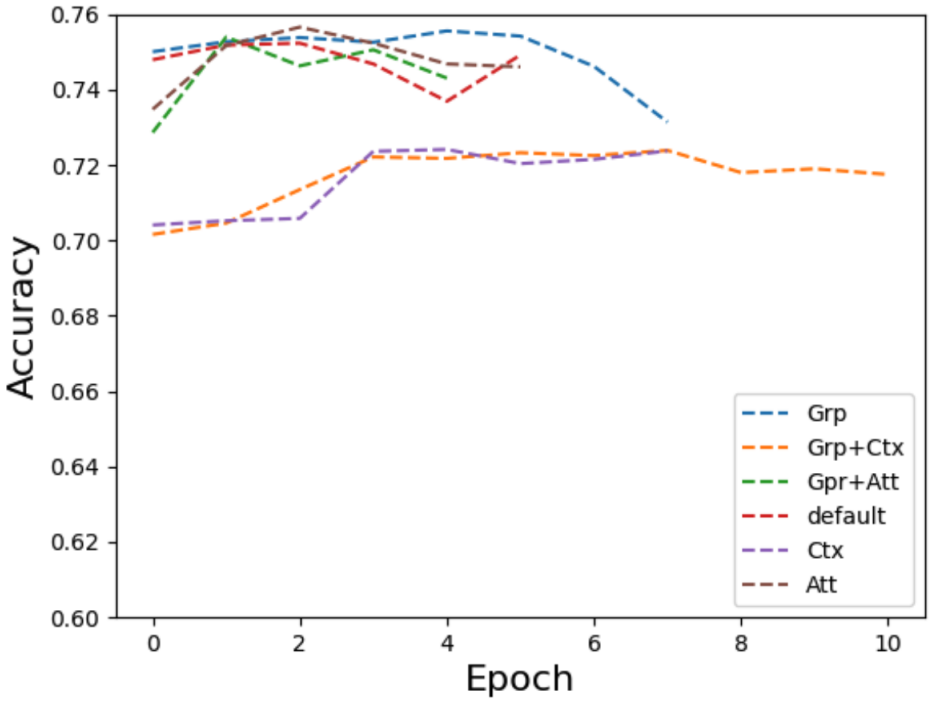


FIGURE 7.3 – Performances pour la classification en actes de dialogue avec les ADs de granularité moyenne

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 intuitivement 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 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 pas toujours de similarité sémantique ou ne partagent aucune information commune avec les énoncés à classifier. Ceci peut expliquer la différence d'environ 4 points entre **Att** et **Ctx**. Cette différence de points semble signifier que l'ajout de contexte constitue plutôt du bruit pour le modèle.

La même interprétation peut être formulée avec **Grp+Ctx**. Cependant le modèle **Grp** sans attention, mais avec des entrées groupées est le second modèle avec une meilleure performance (à l'epoch 4) dans cette première phase.

Les performances des deux meilleurs modèles dont **Att** avec une couche d'attention et **Grp** poussent à déduire d'une part que la couche d'attention dans le modèle permet en effet d'amplifier l'importance de certains tokens (le but l'attention en soit) et ainsi permet d'avoir un meilleur résultat. D'autre part, le fait de grouper les énoncés (avec la stratégie décrite en section 7.3.2) ajoute aussi plus de connaissances aux énoncés à classifier. Ce regroupement d'énoncés réduit d'environ 25% les données utilisées pour l'entraînement et peut être considéré comme étant la bonne approche de

rajout de contexte. Vu la différence de moins d'un point entre les deux modèles, on pourrait déduire que cette stratégie d'augmentation de contexte (regroupement) apporte autant d'informations sur les tokens au niveau de la couche *BiLSTM* que ce que fournit la couche d'attention sur les représentations obtenues de la couche *BiLSTM* des simples énoncés.

Au vu de ces analyses, on peut faire l'hypothèse que les entrées groupées avec une couche d'attention **Grp+Att** auraient de meilleures performances que chacune des deux approches prises séparément. À notre surprise ce n'est pas le cas au regard de la figure 7.3 listant les performances de nos modèles. En plus à la deuxième epoch, **Grp+Att** sous-performe respectivement environ un point et 0.60 de moins de performances que **Att** et **Grp**. Bien que ces différences de performance soient minimes, on pourrait associer cela à la taille du corpus qui est diminuée de 25% lorsqu'on adopte la stratégie de regroupement. Cette hypothèse est d'autant plus plausible lorsqu'on voit que le modèle par défaut **BERT+BiLSTM** à l'epoch 2 a une meilleure performance d'environ 0.50 point par rapport à **Grp+Att**.

La même figure 7.3 montre que le modèle **Att** atteint son optimal après deux epochs contre 4 pour le modèle **Grp**, on en déduit ainsi que la couche d'attention permet une généralisation rapide du modèle.

Ces modèles entraînés lors de cette première phase ont tous comme base **BERT+BiLSTM** et pour les entraîner, nous n'avons pas utilisé des approches de recherche de meilleurs hyperparamètres. De plus, bien que l'ajout de la couche *BiLSTM* lors de l'affinage des modèles *BERT* permette d'avoir de meilleurs résultats dans la littérature pour différentes tâches de classification, nous avons voulu vérifier cela. Nous avons exploré l'utilisation d'autres types de plongements de mots à l'instar de *ConceptNet Numberbatch*, ainsi que l'impact de nos différentes granularités d'ADs et l'utilisation des CRF. Ce sont ces différentes raisons qui nous ont poussées à la deuxième phase d'entraînement de nos modèles.

7.4.2 Phase 2

Cette seconde phase d'entraînement de nos modèles a la particularité de construire de façon aléatoire, à l'aide de la librairie *wandb* [4], différents paramètres et hyperparamètres. D'une part les hyperparamètres portent sur les valeurs du taux d'apprentissage, du nombre d'epochs, du *dropout*, de la taille des *batchs*, du nombre de neurones dans les couches cachées (si la couche *BiLSTM est utilisée*). Et d'autre part, les paramètres quant à eux ont des valeurs booléennes (*True or False*) sur l'ajout de contexte aux énoncés, l'utilisation d'une couche *BiLSTM*, d'attention, et du niveau de granularité des ADs.

Dans cette seconde phase, nos expériences portent sur deux meta-types de modèles : ceux avec en sortie une couche CRF (figure 7.2) et les autres avec une couche dense en sortie dont chaque neurone a la fonction *Softmax* de distribution des probabilités suivant les ADs à prédire. Notons qu'avec une couche CRF en sortie, un input de notre modèle est une séquence d'énoncés de taille n qui a reçu différentes valeurs allant de 5 à 50 lors de nos expérimentations. Ci-dessous les résultats de ces modèles :

4. *Weights & Biases*

A. Avec CRF

Dans ces modèles avec une couche CRF en sortie, nous utilisons deux types de plongements de mots : ceux des modèles pré-entraînés à l'instar de *BERT* et les autres de *ConceptNet Numberbatch (CoNNb)*.

La table 7.1 présente les paramètres et hyperparamètres qui ont un impact fort en termes d'importance et de corrélation sur la performance des modèles. Ces résultats proviennent de 55 modèles différents entraînés avec des paramètres et hyperparamètres dont les valeurs ont été choisies de façon aléatoire sur des plages de valeurs que nous avons définies en amont, ceci par le biais de *Wandb*.

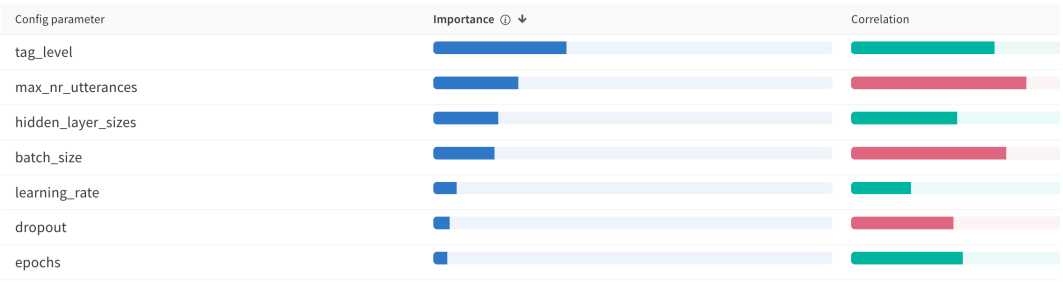


TABLE 7.1 – Influence (importance & corrélation) des paramètres et hyperparamètres sur la performance des modèles

Deux mesures dont l'**importance** et la **corrélation** sont évaluées en fonction d'autres configurations, ces métriques sont calculées par *Wandb*. La corrélation est une fonction linéaire entre l'hyperparamètre ou paramètre et la précision de nos modèles sur les données de validation lors de l'entraînement. Ainsi, une corrélation élevée signifie que lorsque l'hyperparamètre a une valeur plus élevée, la métrique a également des valeurs plus élevées et vice versa. L'importance est une métrique calculée via un entraînement de forêt aléatoire avec les hyperparamètres et paramètres comme entrées et qui permet de produire en sortie des valeurs d'importance des entités pour la forêt aléatoire.

Dans cette table 7.1, le paramètre *tag_level* a la plus haute importance et la meilleure corrélation sur la performance des modèles. Comme mentionné plus haut, nous avons trois niveaux de granularité d'ADs et chacun de ces niveaux a un nombre d'ADs (cf. section 7.3.2). Ce paramètre est suivi par *max_nr_utterances* (qui est le nombre d'énoncés constituants un input passé à nos modèles) et le nombre d'epochs respectivement pour l'importance et la corrélation sur la performance des modèles.

La figure 7.4 illustre l'impact du niveau de granularité des ADs, du type de plongements de mots et du nombre d'énoncés par inputs. Concernant la granularité des ADs, moins on a d'ADs à prédire, meilleures sont les prédictions de nos modèles(sur les données de validation lors de l'entraînement). Statistiquement, ce postulat fait sens du fait que sur la couche dense de sortie de nos modèles, une distribution de probabilité est appliquée en fonction du nombre d'ADs à prédire. Sur notre figure, le niveau de granularité basique (*tag_level_2*) avec 9 ADs a une valeur maximale de précision égale à 0.74 contre une valeur minimale de 0.68.

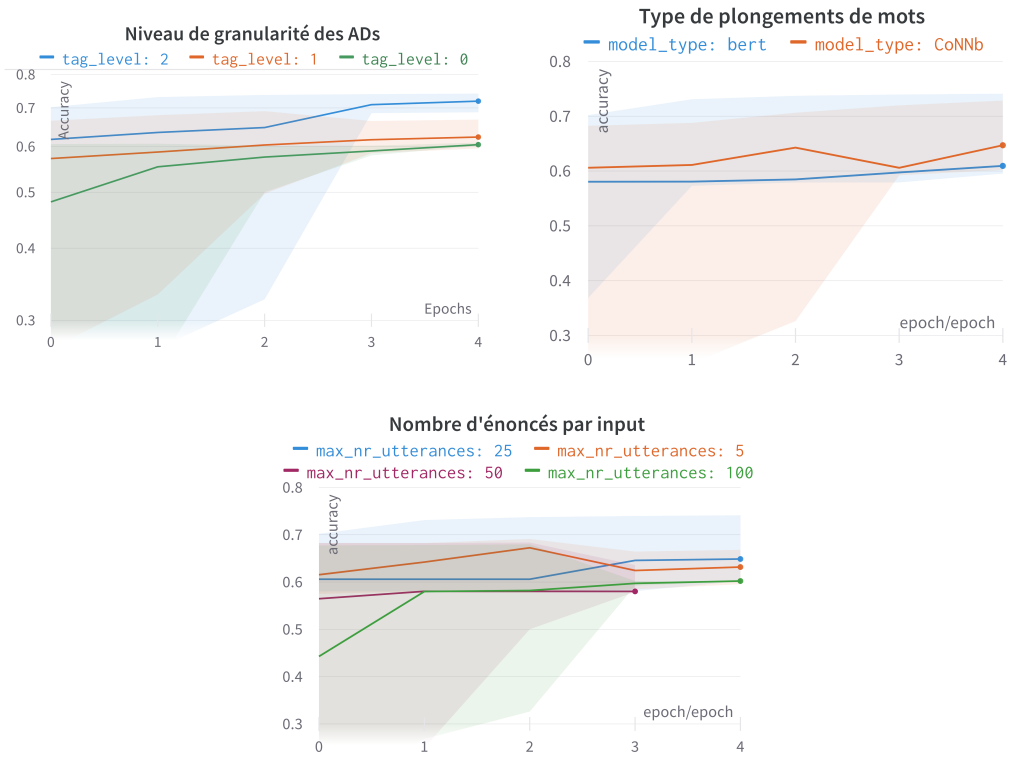


FIGURE 7.4 – Performance des modèles en fonction des 3 principales configurations énumérées précédemment

De même les niveaux *tag_level_1* (avec 18 ADs) et *tag_level_0* (24 ADs) ont respectivement des valeurs de précision minimale/maximale égales à $0.59/0.66$ et $0.60/0.61$. Une seconde tâche, comme l'appariement d'énoncés que nous abordons au chapitre 8 et qui s'appuierait fortement sur les prédictions des ADs fins utilisera le modèle obtenu avec une précision de 0.61, ce qui induirait déjà 39% d'erreur pour cette nouvelle tâche.

Le second graphique de la figure 7.4 met en avant le fait que les plongements de mots de *BERT* donnent une meilleure prédiction que ceux de *CoNNb* avec un gain d'environ 2 points. Le troisième et dernier graphique relatif au nombre d'énoncés par input fait ressortir que la meilleure précision de 0.74 est obtenue avec un nombre n d'énoncés par input égal 25. Viennent ensuite $n = 5$ et $n = 50$ avec respectivement des valeurs maximales de précision égales 0.66 et 0.63. On en déduit une différence d'au moins 10 points avec $n = 25$. D'autres expériences ont donné une valeur maximale de précision toujours égale 0.74 avec $n = 10$, des plongements de mots *BERT* et le niveau d'ADs basique (*tag_level_2*).

De toutes ces analyses, il ressort que la meilleure précision obtenue avec une couche CRF en sortie de nos modèles est égale à 0.74. Qu'en est-il des modèles avec une simple couche dense en sortie qui ne s'appuie pas sur une quelconque dépendance avec un élément précédent dans l'input (comme le fait une couche CRF en sortie), c'est-à-dire des modèles avec un

seul énoncé par input, un peu comme dans la première phase décrite précédemment. Nous répondons à cette question dans les résultats et analyses de la prochaine sous-section des modèles dits « sans CRF ».

B. Sans CRF

Dans ces expérimentations de modèles « sans CRF » en sortie, nous utilisons un seul type de plongement de mots à savoir ceux de *BERT* parce qu'ils ont donné de meilleurs résultats par rapport à ceux de *CoNNb*.

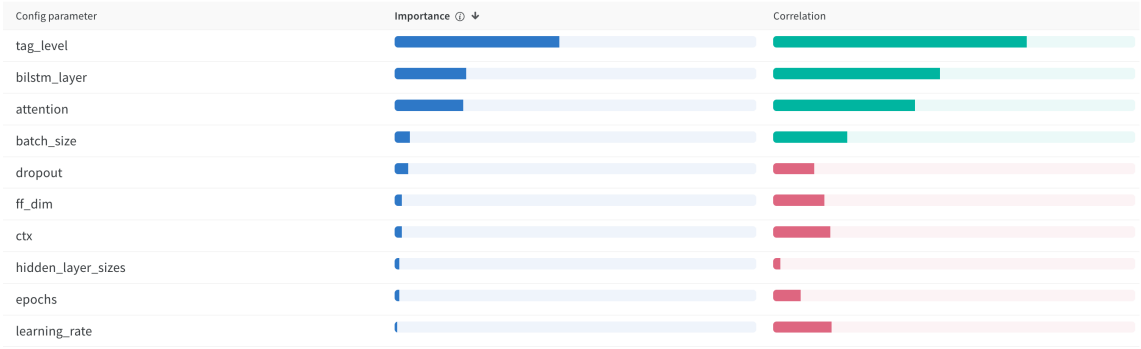


TABLE 7.2 – Influence (importance & corrélation) des paramètres et hyperparamètres sur la performance des modèles « sans CRF »

La table 7.2 présente comme dans la précédente sous-section, l'importance et la corrélation de certains paramètres (*tag_level, bilstm_layer, attention, ff_dim, ctx*) et hyperparamètres (*batch_size, dropout, hidden_layers_size, epochs, learning_rate.*) sur la performance de nos modèles. Il en ressort une fois de plus que la granularité des ADs a un impact important sur les performances, suivi des couches *Bi_LSTM* et d'attention et enfin la taille des batchs pour les 4 premiers hyper/paramètres (hyper/paramètres ici regroupent les hyperparamètres des modèles et des configurations ou paramètres que nous définissons pour avoir différentes variantes de nos modèles). Dans le même ordre d'idée, ces derniers sont fortement corrélés à la performance des modèles. Nous avons analysé plus haut le pourquoi de l'impact des 3 groupes de granularité. *Wandb* a construit 69 modèles différents les uns des autres en combinant aléatoirement les hyper/paramètres.

Dans la phase 1 (cf. section 7.4.1) de nos expérimentations, nous avons présenté l'importance de la couche d'attention pour cette tâche de classification d'énoncés en ADs. Cependant dans cette première phase, la couche *Bi-LSTM* était systématiquement ajoutée aux modèles. Ici nous entraînons des modèles pour certains sans ces couches et pour d'autres avec. Au vu de la table 7.2, il va de soi que ces couches améliorent la performance des modèles.

Dans la figure 7.5, le premier graphique présente sur les epochs (max 4) choisies aléatoirement par *Wandb* les précisions des 4 meilleurs modèles. Force est de constater que ceux-ci ont tous en commun le niveau de granularité basique (*tag_level_2*) avec 9 ADs. Sur ce graphique, les légendes précisent l'ajout ou non des couches aux modèles et les tailles de batchs. Il en ressort que ce dernier hyperparamètre compris entre 16 à 64, bien qu'il soit important, n'en

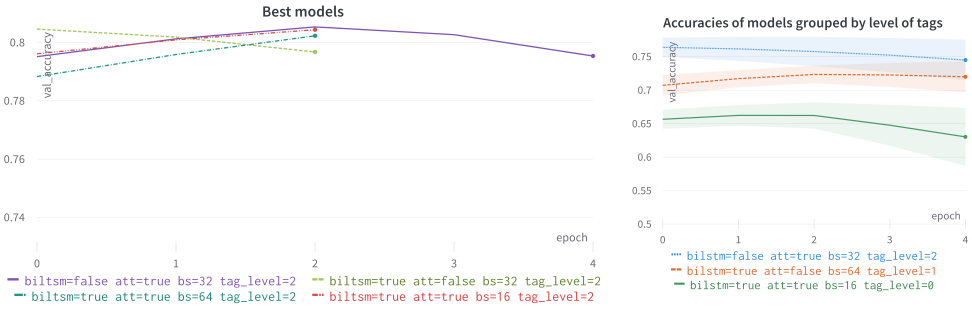


FIGURE 7.5 – Performances des 4 meilleurs modèles et en fonction des niveaux de granularité

est pas le plus intéressant. On peut cependant noter que parmi ces 4 modèles, il y a en deux qui ont une même taille de batchs égale à 32.
Concernant les couches *Bi-LSTM* et d'attention et dans cet ordre, elles ont été ajoutées sur deux de ces modèles qui ont des tailles de batchs de 16 et 64 avec une précision maximale identique à deux décimales près de .80 et ont une différence de points très faible de 0.0010. De ces 4 modèles, le premier ne possède que la couche d'attention et pas la couche *Bi_LSTM*. Il a une précision max de 0.81 et donc une différence d'un point avec les deux précédents et a été entraîné sur ces deux epochs en plus. On peut déduire que la couche d'attention seule apporte plus de valeurs aux modèles et permet de mieux mettre en avant ou d'encoder les représentations nécessaires que la couche *Bi_LSTM* ; et ce pour la tâche que nous étudions dans ce chapitre. Cette déduction se confirme avec la différence de 3 points en moins entre le premier des modèles et le dernier qui a une précision de 0.78 et dans lequel seule la couche *Bi_LSTM* a été ajoutée.

Dans le deuxième graphique de la figure 7.5, les performances des 66 modèles entraînés sont exposées de façon groupée suivant le niveau de granularité des ADs. Les valeurs sont agrégées sur les moyennes et les intervalles en fonction de l'erreur type. Comme nous le savons déjà, moins on a d'ADs à prédire, meilleurs sont les modèles entraînés ; ceci se confirme sur ce second graphique.

La table 7.3 récapitule les précisions des différents modèles entraînés avec ou sans CRF.

	Avec CRF			Sans CRF		
	Niveau de granularité d'ADs			Niveau de granularité d'ADs		
	0 (24 ADs)	1 (18 ADs)	2 (24 ADs)	0 (24 ADs)	1 (18ADs)	2 (9 ADS)
Précision	0.60 ± 0.004	0.63 ± 0.39	0.71 ± 0.02	0.63 ± 0.13	0.72 ± 0.06	**0.74 ± 0.07**

TABLE 7.3 – Récapitulatif des performances des modèles entraînés

7.5 Conclusion

Dans ce chapitre, nous avons présenté les différentes expériences effectuées dans le cadre de la classification d'énoncés filtrés du corpus MRDA en actes de dialogues. Dans ces expériences,

plusieurs paramètres et hyperparamètres ont été pris en compte, entre autres le type de plongements (*BERT* ou *CoNNb*), la prise en compte du contexte, le regroupement des énoncés, les combinaisons des différentes couches (*BiLSTM*, *Attention*, *CRF*) ajoutées aux architectures de nos modèles, la taille des unités dans les couches cachées, les 3 niveaux de granularité des actes de dialogues, etc.

Il ressort de ces expériences l'impact en termes d'importance et de corrélation des hyper/paramètres sur les performances des modèles. Nos analyses ont montré, par exemple que l'utilisation de la couche d'attention seule (en lieu et place de la couche *BiLSTM* ou couplée à celle-ci) permet d'obtenir de meilleurs résultats. De même l'utilisation des CRF en sortie de modèles, bien que s'appuyant sur les séquences précédentes pour classifier la courante, ne donne pas de meilleurs résultats qu'une simple couche dense en sortie avec des distributions de probabilités. D'autres analyses sont présentées dans le chapitre.

Nous nous sommes attardés sur cette tâche de classification d'énoncés en actes de dialogue (CLEADs), parce que l'identification de ces derniers de façon fine (au niveau des phrases par exemple) dans des emails d'une conversation permet à priori de connaître les points en suspens dans celle-ci et les contributions en termes d'actions prises par les collaborateurs pour l'avancement d'un projet. Dans le prochain chapitre, nous montrons comment ces ADs sont utilisables comme un levier majeur.

8 Appariement de segments de texte ou d'énoncés

8.1 Introduction

Les emails, les chats et les échanges dans des forums font partie de ces CMO décrits au chapitre 4 et sont des canaux d'échanges utilisés dans des entreprises via des outils de communication et de collaboration tels que *Skype, Outlook, Teams, Slack, etc.*. Les contenus provenant de ces outils renferment des connaissances considérables, mais leur manque de structuration limite leur utilisation pour en extraire leur plein potentiel. Un problème général qui découle de la nécessité d'une meilleure compréhension et d'une extraction des connaissances de ces contenus, notamment dans le contexte des emails, est la constitution de sous-fils de conversations d'emails.

Un fil de conversation dans un corpus d'emails est formellement défini comme un ensemble d'emails échangés sur un même sujet entre le même groupe de personnes via des actions de réponse ou de transfert (Erera and 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.

— Arborescent : 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, soit la structure 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 trois prismes différents. Tout d'abord, l'algorithme de Zawinski [1] aborde le problème en s'appuyant uniquement sur les métadonnées 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. Cependant ils ne permettent pas d'identifier aisément ou clairement l'essence des informations contenues dans une conversation. Aussi ces approches ne permettent pas, par exemple de facilement suivre l'évolution d'une conversation, ni de savoir quelles sont les principales actions

1. message threading

menées par les interlocuteurs dans de telles conversations d'emails. Ces actions fortement liées 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'importance des actes de dialogue (ADs). L'évolution d'une conversation peut par exemple répondre aux questions suivantes : est-ce que les questions posées en amont dans la conversation ont eu des réponses 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 aux insuffisances susmentionnées, constituent les éléments de motivation des travaux décrits dans ce chapitre. 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** entre ces segments de texte qui peuvent être des phrases. Une paire transverse d'énoncés est l'équivalent dans une conversion asynchrone a une paire d'énoncés consécutifs dans deux tours de paroles dans une conversation synchrone. Elle s'apparente à la notion de paire adjacente dans les dialogues. La figure 1.1 met en avant une conversation professionnelle avec ses métadonnées, ses emails, ainsi que des appariements de segments de texte qui sont ceux en surbrillance de même couleur pris deux à deux. Chaque ensemble des segments de texte avec la même couleur de surbrillance constitue un sous-fils de conversation.

L'approche de constitution de sous-fils de conversation d'emails que nous adoptons repose principalement sur l'appariement de deux énoncés pris de façon transverse sur les emails de ladite conversation. Cette approche repose sur l'utilisation du référentiel d'ADs (cf. section 5.3) qui a permis d'annoter en actes de dialogues des segments de texte du corpus d'emails BC3 (cf. section 6.4.2) que nous avons utilisé. Cette approche est constituée de deux tâches principales : la tâche de classification d'énoncés ou segments de texte extraits d'emails en ADs et l'appariement de ces segments de texte de façon transverse. L'implémentation de notre approche s'est faite dans un premier temps en mode pipeline avec ces deux tâches et, dans un second temps, nous entraînons un modèle multitâche.

Dans la suite du chapitre, nous exposons tout à d'abord les hypothèses inhérentes à la problématique d'appariement d'énoncés (AE) avant de formaliser. Ensuite les corpus et les protocoles utilisés dans nos expériences pour approcher cette problématique sont détaillés ; nous présentons et analysons les résultats de ces expériences. Enfin nous évaluons les modèles issus de nos expériences avant de conclure. Dans la suite du document, nous désignons respectivement par **CLEADs** et **AE** les tâches de CLEADs et d'AE.

8.2 Méthodologie et Formalisation

8.2.1 Hypothèse et méthode

Une conversation est constituée d'au moins deux emails, avec au moins deux interlocuteurs, chacun de ces emails aborde au moins un sujet et contient au moins un segment de texte qui est une phrase ou une combinaison de plusieurs phrases. Pour mettre en relation certaines phrases courtes d'une conversation d'un email B avec ceux d'un 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 espérer un meilleur score de similarité sémantique avec les phrases d'un email précédent. Une phrase courte 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 dans un email 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 texte d'emails différents 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 via l'appariement d'énnoncés (AE) est d'identifier des paires de segments de texte qui sont liées 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 aussi de les compléter avec les ADs 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 qui vont ainsi faciliter la lecture et la compréhension.

8.2.2 Formalisation du problème

Nous formalisons notre problématique en définissant les termes conversation, email, paire positive et négative, score de cohérence comme ci-dessous :

Une conversation avec n emails : $\mathcal{C} = [E_a, E_b, E_c, \cdots, E_n]$

Chaque email E_x contient m segments de texte $E_x = [s_x^1, s_x^2, ..., s_x^m]$

L'ensemble des paires d'énoncés $\mathcal{P} = \mathcal{P}^+ \cup \mathcal{P}^-$ avec $\mathcal{P}^+$ et $\mathcal{P}^-$ respectivement pour les paires positives et négatives.

$$\mathcal{P} = \mathcal{P}^+ \cup \mathcal{P}^- = \{(e_a^i, e_b^j)^+, (e_d^k, e_c^l)^+, \cdots, (e_y^m, e_z^n)^+\} \cup \{(e_a^i, e_b^j)^-, (e_d^k, e_c^l)^-, \cdots, (e_y^m, e_z^n)^-\}$$

avec $(a, b, c, \cdots, x, y, z) \in \mathbb{N} \mid a < b,\ b < c,\ \cdots y < z$ qui sont identifiants des emails et $(i, j, k, l, m, n) \in \mathbb{N}$

Chaque énoncé $e_a^i,\ e_b^j, \cdots$ d'emails respectifs $E_a,\ E_b, \cdots$ possède chacun un acte de dialogue $d_a^i,\ d_b^j$. $[d \oplus e]_a^i$, $[d \oplus e]_b^j$ sont des concatenations de chacun de ces énoncés avec leur acte de dialogue respectif

Les scores de cohérence $\mathcal{SC}$ des paires positives et négatives d'énoncés qui s'appuient sur les similarités sémantiques de ceux-ci sont tels que :

$$\mathcal{SC}(e_a^i, e_b^j)^+ \ \leq\ \mathcal{SC}([d \oplus e]_a^i, [d \oplus e]_b^j)^+$$

$\mathcal{SC}(e_a^i, e_b^j)^- \approx \mathcal{SC}([d \oplus e]_a^i, [d \oplus e]_b^j)^- \approx \iota$ (ι désigne une similarité faible)

8.3 Protocoles D'entraînement de nos modèles

Nous entraînons nos modèles pour les deux tâches mentionnées plus haut : la CLEADs et l'AE. Pour chacune de ces tâches nous avons affiné un modèle pré-entraîné BERT. Pour la seconde tâche, nous avons utilisé la prédiction de phrase suivante (en anglais *Next Sentence Prediction - NSP*) qui est l'une des tâches sur lesquelles les modèles *BERT* ont préalablement été pré-entraînés.

8.3.1 Corpus et annotations

Pour la tâche d'AE, nous avons besoin d'un corpus de conversations d'emails dont les énoncés sont annotés en ADs et sont mis en relation de façon transverse sur les emails de chacune des conversations dudit corpus. 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 pour nos travaux, excepté les travaux de (Jeong et al., 2009) pour la tâche de CLEADs. D'une manière générale, il n'existe pas de corpus d'emails finement annoté en ADs avec des relations transverses entre les énoncés ou phrases d'emails. Nous avons ainsi utilisé le corpus d'emails BC3 (Ulrich et al., 2008) et le corpus « *Coarse Discourse Sequence Corpus* (CDSC) [2] » constitué lors des travaux de (Zhang et al., 2017), une version affinée du corpus de base du forum Reddit. Dans la suite de ce document et dans nos expériences, « Reddit » est utilisé pour désigner ce corpus.

A. Corpus BC3

Le corpus BC3 est constitué de seulement 40 conversations pour 261 emails et 1127 phrases. Il 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 CLEADs des phrases d'emails et de forums avec des approches semisupervisées. Ils ont fait réannoter les phrases d'emails de BC3 avec douze ADs 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 trop imprécis : en effet plusieurs phrases d'emails annotées comme *"statement"* (Cf. figure 6.3) sont pour nous des suggestions ou des élaborations.

Pour cette raison, nous avons décidé de réannoter le corpus BC3 en ADs et en relation entre phrases d'emails dans une conversation. Les ADs utilisés proviennent du référentiel d'ADs que nous avons spécialement établi (section 5.3) pour notre besoin sur les conversations d'emails et aussi pour des raisons de clarté. 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 and Garrett, 2005b) a une valeur de 0.47 pour les annotations en actes de dialogue, cette valeur s'interprète comme un accord modéré entre les deux annotateurs.

2. Sous version de Reddit

Concernant les appariements de segments de texte sur ces 20 conversations, les deux annotateurs ont respectivement identifié 289 et 237 relations entre les segments dans lesdites conversations avec une intersection de 107 relations, soit environ 25% de toutes les relations trouvées. Ces disparités dans les annotations tant au niveau des ADs que de la mise en relation des énoncés, mettent en exergue la difficulté de la tâche d'AE transverses effectuée par des humains, même en s'appuyant sur des ADs. Cependant pour entraîner nos modèles d'AE, 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, donc potentiellement proches thématiquement. Dans nos expériences, BC3 est utilisé pour référencer le total de 614 paires.

B. Corpus Reddit

Le corpus Reddit a été 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. Nous avons analysé ce corpus et constaté que bon nombre des messages possèdent plusieurs phrases, chaque message a pourtant été annoté avec un seul AD. Ceci pose un problème d'identification exacte de la phrase qui porte l'AD en question. Face à ce problème nous avons opté pour un filtrage des données de Reddit et avons extrait que les messages avec un maximum de 3 phrases [3].

Les ADs utilisés pour annoter les messages de ce corpus dans le cadre des travaux de (Zhang et al., 2017) sont les suivants : « *Question & Request, Answer, Announcement, Agreement, Appreciation & Positive Reaction, Disagreement, Negative Reaction, Elaboration & FYI, Humor, Other* ». Dans cette liste de 10 ADs, nous constatons une ambiguïté notamment au niveau des questions et des requêtes qui sont sujettes à la même interprétation dans ces travaux comme une « demande d'information ». Pourtant, dans le référentiel 5.3 que nous avons défini et qui s'appuie sur la norme ISO 24617-2, les questions et requêtes correspondent respectivement à des « demandes d'information » et des « demandes d'action ». Nous constatons aussi que certains actes de dialogues *CDSC* ne figurent pas dans notre référentiel à l'instar de « Announcement, Negative Reaction, Humor, Other », mais nous pouvons respectivement faire correspondre « Negative Reaction » et « Announcement », à un désaccord ou une réfutation et un partage d'information.

Les paires positives extraites de ce corpus ne sont rien d'autre que les relations *reply-to* définies lors de l'annotation dans le cadre des travaux (Zhang et al., 2017). Les paires négatives quant à elles, pour ce corpus *CDSC*, sont constituées de la même façon que celles de *BC3*, c'est-à-dire qu'elles sont constituées d'énoncés faisant partie de la même conversation et qui n'ont pas été annotés en relation *reply-to*.

La table 8.1 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 $BC3_{map}$.

3. plus généralement, cela soulève la question de la granularité du support des actes de dialogues et des relations transverses

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 8.1 – Distribution des données utilisées (PP : Paires positives, PN : Paires négatives)

8.3.2 Stratégies d'entraînement des modèles

Pour notre tâche d'AE, comme mentionné plus haut, nous approchons cette tâche comme celle de la prédiction de la phrase suivante utilisée lors du pré-entraînement des modèles comme *BERT* qui n'est rien d'autre qu'un classifieur binaire et se rapproche de la tâche d'implication (*entailment* en anglais). Dans notre cas, puisqu'il est question de conversation, on peut qualifier cette tâche de *follow-up sentence*. Dans nos expériences, nous entraînons nos modèles avec des stratégies différentes au niveau des entrées de ces modèles et des types de ceux-ci.

A. ADs et encodage d'inputs

La première stratégie s'articule autour de l'utilisation des ADs en combinaison avec les énoncés des paires à classifier. Les classifieurs prennent trois entrées : le premier énoncé, le second énoncé et le label positif (0) ou négatif(1) de la paire d'énoncés. Nous affinons l'un de nos modèles avec les inputs tels que nous venons de les présenter. Pour les autres modèles, nous utilisons les ADs soit avec la désignation exacte de ceux-ci : « suggestion, question, approbation, etc. » soit, nous les encodons avec des tokens spéciaux comme ci-dessous :

```
{
    'explanation':'[EPNT]', 'suggestion':'[SGET]', 'politeness':'[POLI]',

    'question':'[QSTI]', 'hypothesys': '[HPTS]', 'inform': '[IFRM]',

    'request':'[RQET]', 'answer':'[ANSR]', 'agreement':'[AGMT]',

    'disagreement': '[DGMT]', 'address a request':'[ARQET]',

    'offer':'[OFER]', 'promise':'[PMSE]', 'address a suggestion':'[ASGT]',

    'instruct':'[ISTR]', 'assessment/appreciation':'[ASMT]',

    'disconfirm/disapprove':'[DSCF]', 'other':'[OTHR]', 'elaboration':'[EPNT]',

    'humor':'[HMOR]', 'negativereaction':'[NGRT]', 'announcement':'[IFRM]',

    'confirm':'[COFR]', 'topic change':'[TPCH]'
}
```

Listing 8.1 – Actes de dialogue (avec une granularité fine) et leurs tokens d'encodage respectifs

Paires	énoncés	label
1	[RQET]→ Je te laisse regarder pour épurer cette liste. [ARQET] → Ok, je regarde çA.	PP
2	[QSTI] → Quelle est ta recommandation quant au nettoyage ou non des inputs (via liste de stopwords) avant de les soumettre à SuperOutil? [QSTI] → Vous retirez les stopwords en apprentissage et en prédiction ?	PN
3	question → Vous retirez les stopwords en apprentissage et en prédiction ? answer → Jusqu'à présent, on retire les stopwords à la fois en apprentissage et en prédiction.	PP
4	inform → Avant d'utiliser SuperOutil, nous « nettoyons » les inputs (sujet + contenu de mails) en retirant des stopwords. inform → Actuellement, nous travaillons sur des n-grams de longueur maxi = 5.	PN

TABLE 8.2 – Exemples de paires d'énoncés avec leurs labels respectifs : PP pour les paires positives et PN pour les négatives

La table 8.2 illustre l'ajout des ADs aux énoncés avant que l'ensemble ne soit encodé en représentations (qui vont servir d'entrées aux modèles à affiner) à l'aide des tokenizeurs des modèles pré-entraînés *BERT*. Comme dans le chapitre 7 sur la CLEADs, nous utilisons aussi les différentes granularités des ADs dans la tâche d'AE.

B. Stratégies d'entraînement

Nous entraînons nos modèles selon deux principales stratégies :

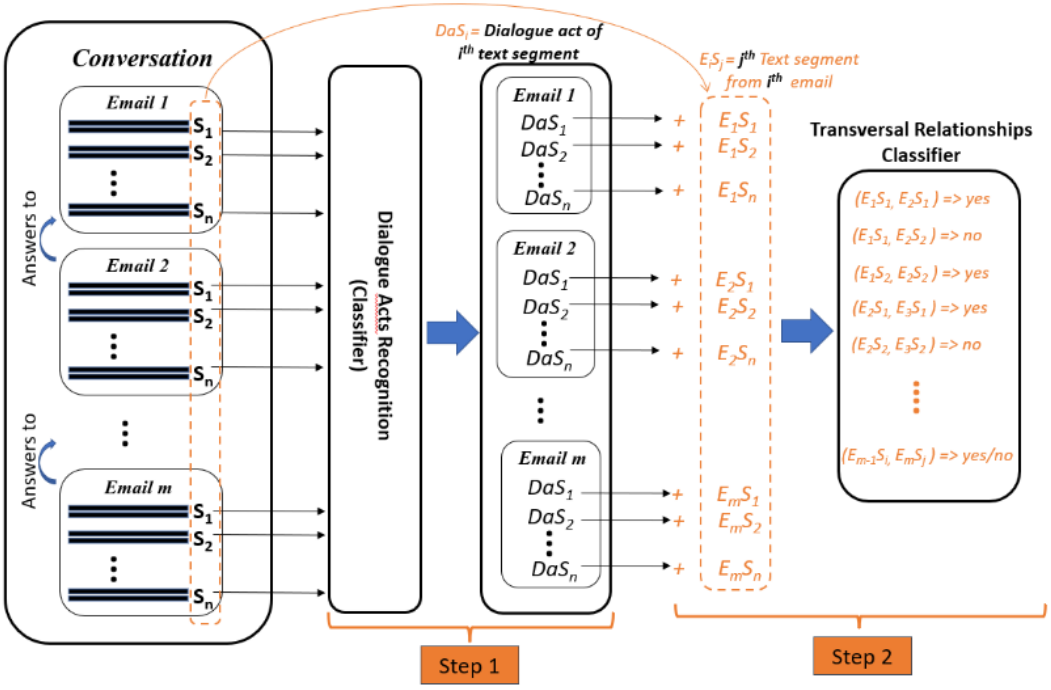


FIGURE 8.1 – Processus en deux étapes pour l'appariement de segments de texte

a) **Mode « Pipeline »**

Ce mode **pipeline** enchaîne deux étapes : la première consiste à classifier les énoncés en ADs avec le meilleur des modèles du chapitre 7 et dans la seconde étape, nous utilisons les ADs prédits ou « gold » pour entraîner les modèles d'AE. Dans cette

seconde étape, nous affinons directement **BERT** pour certains de nos modèles et pour les autres, nous utilisons le framework *SetFit (Efficient Few-shot Learning with Sentence Transformers)* [4] qui est aussi un affinage de *BERT* et qui s'appuie fortement sur l'apprentissage contrastif. Ce type d'apprentissage en général tente d'apprendre un espace de caractéristiques afin de combiner/rassembler des points qui sont liés et écarter des points qui ne sont pas liés. D'après les résultats de (Tunstall et al., 2022), *SetFit* donne des résultats intéressants avec peu de données et avec un temps d'affinage relativement petit. La petite taille de nos données rentre à priori bien dans le cadre d'utilisation de *SetFit*.

La figure 8.1 illustre les deux étapes de notre pipeline dont l'objectif est de faire ressortir les paires de segments positives (annotées "yes" dans le schéma). Dans ce mode « pipeline « , nous entraînons différents modèles ci-dessous avec pour chacun le (ou les) corpus sur lequel (lesquels) il a été entraîné :

— **AE** : AE sans les actes de dialogues, ce modèle s'appuie uniquement sur la similarité sémantique,

— **AE+ADG** : AE avec les Actes de Dialogues Gold, entraîné de façon indépendante avec Reddit, Reddit+BC3 et Reddit+BC3_map,

— **AE+ADP** : AE avec les Actes de Dialogues Prédits en utilisant notre modèle de CLEADs. Ce modèle est entraîné de façon indépendante avec Reddit, Reddit+BC3. Les ADs utilisés reflètent la réalité dans laquelle nos modèles seront utilisés, car les annotations en actes de dialogue seront faites de façon automatique et non manuellement (GOLD) comme dans le modèle AE+ADG. D'autre part un tel modèle est susceptible d'apprendre à corriger certaines erreurs faites lors de la reconnaissance des ADs.

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 si il y a 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 leurs données de test respectives 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.

Tous les modèles entraînés avec *Setfit* utilisent 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].

Pour conclure sur ce mode « pipeline », il est à noter que nous affinons *Setfit* pour la première expérience avec le niveau de granularité fin pour les ADs quand ceux-ci

4. SetFit - Efficient Few-shot Learning with Sentence Transformers
5. https ://optuna.org/

sont ajoutés à leur énoncé respectif et pour la seconde, la version de base sans casse de *BERT* est affinée avec comme tâche, celle de la prédiction de la phrase suivante (en anglais *Next Sentence Prediction-NSP*) qui s'apparente fortement à notre tâche d'AE d'emails. Pour cet affinage, nous utilisons *BertForNextSentencePrediction* [6] qui est une architecture spécifique s'appuyant sur *BERT* avec, pour entrées, les énoncés d'emails et leurs ADs sélectionnés à chaque fois sur l'un des trois niveaux de granularité (cf. section 7.3.2) de ceux-ci.

b) **Modèle joint**
Bien que le mode pipeline soit une première approche intéressante, il n'en demeure pas moins qu'elle a cet inconvénient de se dérouler en deux étapes nécessitant davantage d'opérations entre la première et la seconde étape. De plus, i les AD ont un impact sur l'AE, la réciproque n'est-elle pas envisageable ? Pour répondre à cette question et réduire les coûts en opération de l'approche pipeline, nous optons pour l'entraînement d'un modèle multitâche (bi-tâches dans notre cas). Ce type d'approche a montré son efficacité dans divers domaines et pour plusieurs tâches à l'instar de (Plaza-Del-Arco et al., 2021) qui s'appuient sur l'analyse des sentiments pour mieux détecter les discours de haine, de même (Mehmood et al., 2019) utilisent cette approche pour la reconnaissance d'entités nommées biomédicales.

Nous entraînons un **modèle joint** (cf. 8.2) qui est un empilement de deux couches : la première classifie les énoncés en ADs et la seconde qui utilise les ADs prédits pour l'AE. Pour chacune de ces tâches, nous utilisons deux modèles prédéfinis de la bibliothèque *Tranformers* de *HuggingFace* : **BertForNextSentencePrediction** pour la tâche d'AE et **BertForSequenceClassification** pour la CLEADs.

Les deux modèles de la pile partagent les mêmes entrées encodées avec le tokenizeur *BERT*. La couche de CLEADs prend en entrée deux vecteurs représentant les deux énoncés dont on veut prédire s'il existe une quelconque relation entre eux. Les vecteurs utilisés ici sont ceux du token **[CLS]** qui accumulent toutes les informations des tokens qui le suivent dans l'énoncé. **[CLS]** est ajouté au début de chaque énoncé (il est possible de choisir de ne pas l'inclure) au début d'un texte lors de la tokenization. De la même façon le token **[SEP]** est ajouté entre les deux énoncés. Ces deux représentations vectorielles sont donc passées séparément à la première couche de notre modèle joint qui prédit leurs ADs respectifs et ceux-ci sont encodés en tokens spéciaux qui sont ensuite insérés au début (après **[CLS]**) respectivement de chacun des deux vecteurs. Enfin ces deux nouveaux vecteurs sont concaténés pour former un vecteur qui sert d'entrée à la seconde couche de notre modèle joint qui va prédire si oui ou non les deux énoncés sont en relation. La figure 8.2 illustre ces différentes étapes d'entraînement de notre modèle joint.

En plus des vecteurs que partagent les couches de notre modèle, il y a la fonction de perte $\mathcal{L}$ qui est la somme des deux fonctions de pertes de chacune des couches. Ces fonctions de perte se calculent avec des indicateurs d'entropie croisée. Plus précisément l'entropie croisée binaire (*Binary Cross Entropy : **BCE***) pour l'AE et la perte

6. BertForNextSentencePrediction

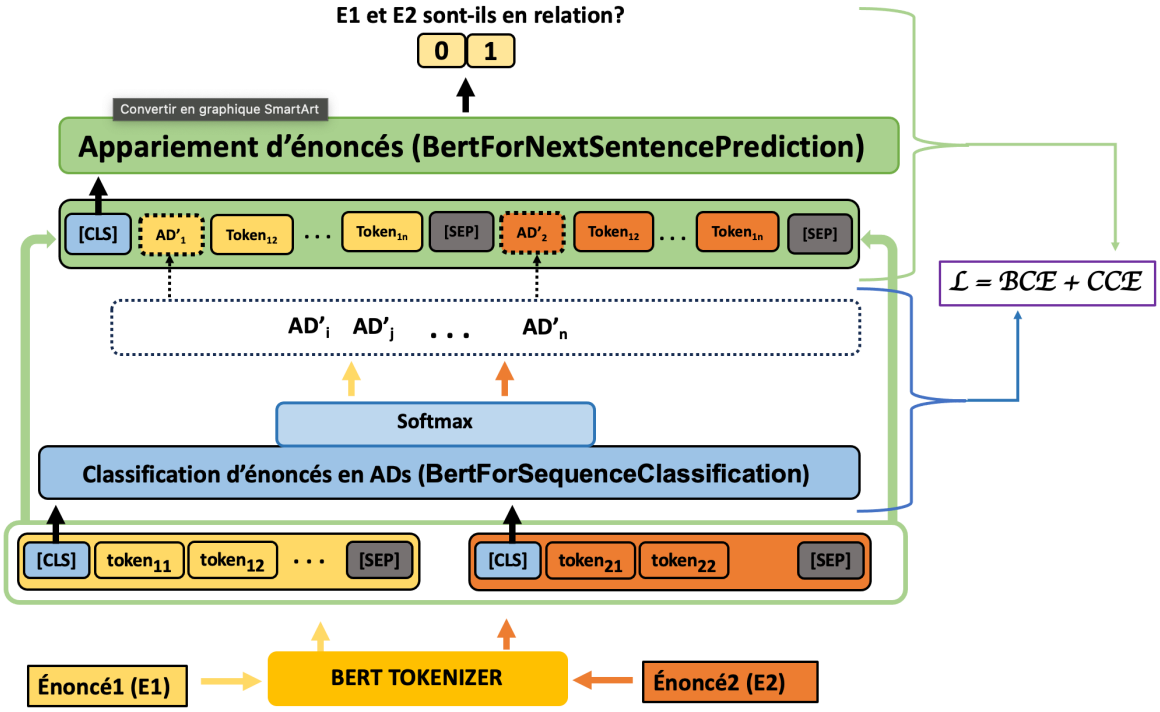


FIGURE 8.2 – Architecture de notre modèle joint

d'entropie croisée catégorielle (*Categorical Cross Entropy Loss* : ***CCE***). Ces fonctions de perte sont respectivement précédées par les fonctions d'activation **Sigmoïd** (8.4) et **Softmax** (8.5). Les paramètres des modèles de nos couches sont mis à jour lors de la rétropropagation au cours de laquelle sont calculés les gradients de cette fonction de perte par rapport aux paramètres des modèles. Ces gradients sont utilisés pour mettre à jour les valeurs de ces paramètres, ce processus de mise à jour est rendu optimal par des optimiseurs. Nous utilisons une version améliorée *AdamW* (Loshchilov and Hutter (2017)) de l'algorithme *Adam* avec un taux d'apprentissage aléatoirement choisi au départ sur une liste de valeurs.

$$\mathcal{L} = \mathcal{BCE} + \mathcal{CCE} \tag{8.1}$$

$$\mathcal{BCE} = -\frac{1}{N}\sum_{n=1}^{N}[y_n log\widehat{y_n} + (1-y_n)log(1-\widehat{y_n})] \tag{8.2}$$

$$\mathcal{CCE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{c,n} log\widehat{y_{c,n}} \tag{8.3}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{8.4}$$

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K \tag{8.5}$$

Lors des entraînements de nos modèles, nous choisissons aléatoirement d'une part des hyperparamètres (taille de *batchs*, taux d'apprentissage et le nombre d'epochs) et d'autre part des paramètres pour :

— spécifier le niveau de granularité des ADs : fin, moyen et basique,

— savoir si on encode les ADs ou pas,

— fixer ou pas un nombre identique de paires négatives et positives,

— spécifier le corpus utilisé : *Reddit* ou *Reddit+BC3* qui sont de petites tailles (cf. 8.1) pour nos tâches. Nous n'utilisons pas BC3 seul comme dans le mode pipeline.

Le choix de ces paramètres s'effectue avec l'API python de *Wandb*[7] comme nous l'avons fait dans le chapitre 7.

Les ADs utilisés ici sont de deux niveaux respectivement avec 16 ADs pour le niveau « 0 » et 9 ADs pour le « 1 ». Certains ADs du listing 8.1 ne sont pas utilisés avec les entrées pour ce modèle joint parce que ne correspondant à aucun des énoncés tant dans le corpus Reddit que dans BC3.

8.4 Résultats et Analyses

8.4.1 Mode « Pipeline »

A. Affinage avec *Setfit*

Dans ce mode « pipeline », et comme nous allons présenter les résultats ainsi que leurs analyses respectives sous deux prismes : le premier porte sur la table 8.3 qui récapitule les scores F1 des différents modèles affinés avec *Setfit* . 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 8.1).

7. Weights & Biases

Données d'entraînements / Données de test	AE			AE + ADP			AE + ADG			
	BC3	Reddit	Reddit+BC3	BC3	Reddit	Reddit + BC3	BC3	Reddit	Reddit + BC3	Reddit+BC3_map
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 8.3 – Résultats des modèles d'AE

Les scores des différents modèles entraînés sur le seul corpus BC3 sont biaisés du fait de la petitesse de BC3. En plus nous avons relevé du sur-ajustement lors de l'entraînement ces modèles.

Sans supprise, 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 donne respectivement des scores de 0.78 et 0.75 sur les données de test de BC3 et Reddit. Cependant, AE+ADP entraîné aussi sur 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'AE dans les conversations. Ainsi, plus performant sera notre modèle de Classification d'énoncés en ADs, meilleurs seront les résultats de nos modèles d'appariement.

Dans nos expériences, nous avons établi une correspondance des actes de dialogues de BC3 vers ceux de Reddit formant ainsi BC3_map. Nous entraînons 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 5.3 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 des 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 étant prédits comme des questions ou des annonces (*inform*). Ci-dessous quelques faux négatifs (FN) et faux positifs(FP) prédits par le modèle AE+ADP.

(a) FN : *Inform : It is not done but you will get the idea. <–> Assessment- it's a good piece of work.*

(b) FN : *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 ?*

(c) FN : *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]*

(d) 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.*

(e) 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.*

(f) 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 a, b, d et f. 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 c, 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 lesquels 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 du processus d'annotations.

B. Affinage simple de *BERT*

Comme mentionné dans la section 8.3.2, en plus de *Setfit* dont nous venons de présenter et analyser les résultats, nous avons affiné *BERT*.

Dans cette expérience, les ADs prédits s'obtiennent par inférence avec le modèle de CLEADs entraîné sur MRDA avec les ADs de granularité fine (24 ADs). Cependant, nous constatons que l'inférence ne produit que 16 ADs que nous considérons comme fin et on obtient respectivement 10 ADS et 16 ADs en les projetant sur les niveaux de granularité basique et moyen. Cette projection a pour objectif d'observer l'impact du niveau de granularité sur l'AE.

```
AD_fin = { 'ex': '[EPNT]', 's': '[SGET]', 'p': '[POLI]', 'hy': '[HPTS]',
           'inform': '[IFRM]','i': '[IFRM]', 'r': '[RQET]', 'ag': '[AGMT]',
           'dag': '[DGMT]', 'adr': '[ARQET]','o': '[OFER]', 'pr': '[PMSE]',
           'ads': '[ASGT]', 'is': '[ISTR]', 'aa': '[ASMT]', 'dcf': '[DSCF]',
           'other': '[OTHR]', 'question': '[QSTI]', 'answer': '[ANSR]',
           'appreciation': '[ASMT]', 'elaboration': '[EPNT]', 'humor': '[HMOR]',
           'agreement': '[AGMT]', 'negative reaction': '[NGRT]', 'tc': '[TCGH]',
           'disagreement': '[DGMT]', 'announcement': '[IFRM]', 'cf': '[COFR]',
         }
AD_basique= { '[EPNT]': ['explanation', 'elaboration'], '[SGET]': 'Suggestion',
              '[POLI]':'Politeness', '[QSTI]':'question', '[HPTS]':'Hypothesis',
              '[FBCK](Feedback)':['appreciation', 'assessment'],
              '[RQET]': ['request', 'instruct],
              '[OFER]':['Offer', 'address a suggestion', 'address a request', '
                  promise'],
```

```
15              '[IFRM]': ['inform', 'other', 'humour', 'agreement', 'answer', '
                    negative reaction', 'disagreement', 'announcement', 'confirm',
                     'disconfirm', 'topic change']
16              }
```

Listing 8.2 – Les deux niveaux de granularité des Actes de dialogue

On retrouve également 16 ADs « gold » pour les ADs moyens et fins sur le corpus Reddit+BC3 et 9 ADs « gold » sur Reddit seul. Nous ne considérons finalement que deux niveaux de granularité : basique et fin, à cause du nombre d'occurences identiques des ADs prédits et « gold » pour les niveaux de granularité fin et moyen. Le listing 8.2 présente les ADs associés aux deux niveaux de granularités retenus.

Données d'entraînement	AE	AE+ADP		AE+ADG	
		basique	fin	basique	fin
Reddit	0.58	0.57	0.60	0.71	0,72
Reddit+BC3	0.61	0.63	0,65	0.70	0.72

TABLE 8.4 – (#) rappelle le nombre d'ADs par niveau de granularité. * et ** spécifie ce nombre respectivement sur Reddit et Reddit+BC3

Nous avons affiné des modèles sans ADs (dénoté AE), avec les ADs prédits (AE+ADP) et « gold » (AE+ADG). Contrairement à ce qui est fait avec *Setfit* où nous utilisons seulement les ADs de niveau de granularité fin, ici avec *BERT* nous utilisons les deux niveaux de granularité des ADs : basique et fin. Dans la table 8.4, le modèle **AE** entraîné respectivement sur Reddit et Reddit+BC3 a une performance égale à $\mathbf{0.58}$ et $\mathbf{0.61}$, on peut en déduire que l'ajout du corpus BC3 sur Reddit permet un gain de **3** points. Le même modèle avec *Setfit* sur Reddit a une performance de $\mathbf{0.69}$, on note une différence de **11** points.

Dans la table 8.4, le modèle AE+ADP a des performances égales à $\mathbf{0.57}$ et $\mathbf{0.60}$ respectivement pour les granularités « basique » et » fine » sur Reddit, parallèlement sur Reddit+BC3, on a $\mathbf{0.57}$ et $\mathbf{0.60}$. On note une fois de plus, l'apport du corpus BC3 avec des différences de 6 et 5 points respectivement avec les ADs basiques et fins. On note également un gain de 3 points sur le Reddit lorsqu'on passe des ADs basiques aux fins, sur Reddit+BC3 cette différence est de 2 points.

Le modèle AE+ADG a des scores entre $\mathbf{0.70}$ et $\mathbf{0.72}$. Pour ce modèle, on relève une perte d'un point avec l'ajout de BC3 sur le corpus Reddit, ce qui est contraire à l'hypothèse de gain de points par l'ajout de données. Cette contraction peut avoir comme première explication le fait que les ADs « golg » de Reddit ont des ambiguïtés avec les nôtres, notamment au niveau des questions et requêtes qui sont bien distinctes dans nos ADs, et ont la même connotation dans les ADs de Reddit. En plus les paires d'énoncés qui ont pour premier énoncé les questions sont très fortement représentées dans un corpus de forum comme Reddit. avec les ADs fins AE+ADG a un score identique égle à $\mathbf{0.72}$ sur Reddit et Reddit+BC3, là où une fois de plus on se attendu un score meilleur sur Reddit+BC3. Les ADs fins sur Reddit et Reddit+BC3 ont un nombre d'étiquettes respectif de 10 et 16. Et donc cette différence de nombre des ADs serait probablemt le facteur de réduction des performances de AE+ADG

sur REddit+BC3, parce qu'avec les ADs prédits (nombre identique d'ADs sur les deux corpus) on note un gain de 5 points lorsqu'on passe de Reddit et Reddit+BC3.

Sur cette table 8.4, on observe une évolution progressive des performances des modèles sans les ADs, ceux avec les ADs prédits et les ADs « gold ». On a des gains de 2 et 12 points respectivement de AE à AE+ADP et de AE+ADP à AE+ADG si l'on considère la granularité fine des ADs. Cette observation montre l'apport qu'ont les ADs dans la tâche d'AE de conversations d'emails (BC3) ou de forum (Reddit). Elle prouve également que ces appariements sont d'autant meilleurs si en amont le classifieur en ADs l'est aussi. Notons tout de même que la meilleure performance avec l'affinage de *BERT* sur Reddit égale à **0.60** est inférieure de 7 points à celle de *Setfit* égale à **0.67**. Cette différence de points peut se justifier par l'apprentissage contrastif implémenté dans *Setfit*. Toutefois, il est intéressant de noter que chacun des affinages de *BERT* prend en moyenne 30 minutes là où ceux avec *Setfit* mettent environ 3h et 30 minutes.

8.4.2 Modèle Joint

Nous avons entraîné le modèle joint qui empile deux couches avec différents paramètres et hyperparamètres comme mentionné dans la section 8.3.2. La figure 8.3 présente l'importance des paramètres de haut en bas respectivement pour la tâche de CLEADs et d'AE. Pour la CLEADs, on aurait attendu le niveau de granularité des ADs en première position, mais il occupe la seconde place juste après l'hyperparamètre de taille de batch et enfin vient le paramètre sur la parité ou pas des paires positives et négatives, c'est ce dernier paramètre qui a le meilleur score de corrélation pour cette tâche. Contrairement à la position précédente du paramètre sur la granularité des ADs pour CLEADs, il est en première position pour les AE suivie de la taille des batchs. Cette première position alerte d'ores et déjà sur l'importance des ADs pour les AE.

La figure 8.4 montre les performances des modèles de CLEADs et d'AE groupées en fonction de corpus (première colonne) et en fonction du niveau de granularité (seconde colonne). Les bandes colorées sur les graphiques représentent pour chacun des éléments de regroupement, les intervalles des performances. Les meilleures performances sur le corpus Reddit pour la CLEADs et l'AE sont respectivement égales à **0.84** et **0.87** avec une inégalité au niveau de paires d'énoncés et le premier niveau de granularité (9 ADs). Pour Reddit+BC3 on a dans le même ordre **0.77** et **0.80** avec aussi une inégalité au niveau des paires et le second niveau de granularité (16 ADs).

La deuxième colonne de la figure montre l'impact des ADs sur les performances des modèles, autant pour la CLEADs, moins on a ADs à prédire meilleurs seront les prédictions et le contraire devrait s'appliquer sur l'AE parce que mieux sont détaillés les ADs meilleurs sont les AE. Cependant, cette seconde hypothèse ne se justifie pas au regard du premier graphique de la seconde colonne. Une observation des résultats du modèle *Setfit* pour l'AE avec les ADs prédits sur le corpus Reddit fait ressortir une performance de **0.67** contre des performances autour de **0.80** qui est même aussi supérieur à l'affinage de *Setfit* pour l'AE avec les ADs « gold ». Cette observation démontre les améliorations que nous obtenons avec notre modèle joint. Ces améliorations de performances seront confirmées ou infirmées dans le prochain chapitre 9 dans lequel nous évaluons nos modèles.

La mise en œuvre de notre modèle joint a été motivée par la question de savoir si l'apprentissage des AE aurait un impact sur la CLEADs, sachant que la réciprocité a été démontrée par l'analyse des résultats de notre approche en pipeline. La deuxième colonne de la figure 8.5 montre des

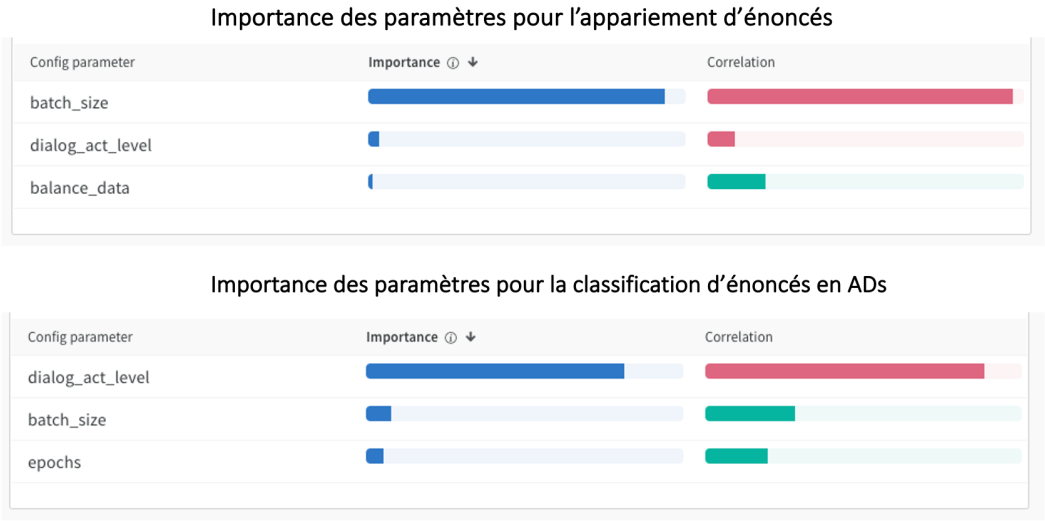


FIGURE 8.3 – Importance des paramètres/hyperparamètres pour les deux tâches entraînés dans notre modèle joint

FIGURE 8.4 – Performances des modèles de CLEADs et d'AE groupé en fonction de corpus (première colonne) et en fonction du niveau de granularité (seconde colonne)

droites qui ont une pente ascendante dont l'angle augmente très légèrement au fur à mesure que les valeurs de l'axe des abscisses évoluent de façon très lente également. Ceci démontre bien que les performances des couches de notre modèle joint ont une influence les unes sur les autres. Et

FIGURE 8.5 – Performances des modèles de CLEADs et d'AE groupé en fonction de la parité ou de l'inégalité des paires d'énoncés (première colonne) et relations entre ces performances

donc une réponse favorable est retournée à la question de savoir si les AE auraient un impact sur la CLEADs. La première colonne dans la même figure 8.5 s'intéresse à la parité ou pas des paires d'énoncés à passer dans nos modèles. Pour la tâche d'AE, on voit très clairement en haut de cette colonne que la parité en bleu a un impact sur les performances des modèles avec une différence d'environ un point en moyenne par rapport aux modèles entraînés sans rééquilibrer des paires positives et négatives. Cette observation n'est pas la même pour la tâche de CLEADs où la parité des paires ne donne pas de meilleures performances.

Dans notre première approche en pipeline, l'utilisation des ADs prédits correspond à ce que nous faisons avec notre première couche dans le modèle joint. Notons tout de même que, pendant que les performances des modèles du mode pipeline avec les ADs prédits ont des performances maximales autour **0.70** selon que l'on utilise *Setfit* ou **BERT**, avec notre modèle joint, nous avons des perfomaces maximales de **0.80** soit une différence moyenne d'environ 10 points. Ce qui démontre tout l'intérêt de cette approche qui met en moyenne une heure pour son exécution.

8.5 Conclusion

Ce chapitre a présenté les expériences effectuées pour la tâche d'AE. Elles ont principalement porté sur deux approches : la première qui est un pipeline et la seconde un modèle joint multitâches. Dans chacune de ces approches, deux calculs sont effectués de façon consécutive : un pour les ADs et un autre pour l'AE.

Le mode pipeline comme son nom l'indique se déroule en deux étapes (cf. 8.1) : la première consiste à utiliser le modèle avec des meilleures performances issu du chapitre 7 pour classifier

les énoncés en ADs et lors de la seconde les ADs prédits sont utilisés pour entraîner un modèle d'AE. Dans ce même mode, pour attester de l'impact des ADs, au lieu d'utiliser ceux prédits, les ADs ou « gold » ont été utilisés directement dans la seconde phase. Les résultats de cette approche d'AE ont montré l'intérêt de l'utilisation des ADs. L'analyse de ces résultats 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 lors de la première étape. Une des insuffisances dans ce mode pipeline se situe au niveau des correspondances entre les différents niveaux de granularités des ADS qui ont été constitués manuellement et sans évaluations. Entraîner des modèles de correspondances d'ADs serait une piste pour palier à cette insuffisance.

La seconde façon d'approcher notre problème d'AE a été d'entraîner un modèle joint en empilant une première couche qui se charge de prédire les actes et une seconde qui prédit si deux énoncés passés au modèle joint sont en relation ou pas. Lors de l'entraînement de ce modèle, les plongements des énoncés et une somme des fonctions de perte respectives des modèles de CLEADs et d'AE sont partagés entre les deux couches pour une amélioration simultanée de leurs performances. Malgré les petites tailles des donnés passés en Input à ce modèle, nous avons obtenu des résultats qui sont meilleurs que ceux obtenus avec les mêmes données dans notre mode pipeline. À noter en plus que le modèle utilé pour la CLEADs a été entraîné en amont sur le corpus MRDA qui a une taille de données d'entraînement de 36 968 qui est d'environ 5 fois la taille de Reddit+BC3 utilisé dans notre modèle joint. Et donc une piste des améliorations de ce modèle serait par exemple d'utiliser respectivement les corpus MRDA et REddit+BC3 pour nos tâches de CLEADs et d'AE. Une seconde piste serait par exemple d'utiliser une fonction de perte pondéré (au lieu d'une simple somme) avec des paramètres de pondérations qui seront appris pour minimiser davantage la fonction de perte et ainsi améliorer les performances de notre modèle. On pourrait aussi envisager de rajouter des couches partagées d'attention par exemple puisqu'on sait que ce type de couche améliore les performances des modèles de CLEADs.

Dans le prochain chapitre, nous évaluons les différents modèles de classification d'énoncés en actes de dialogues et d'appariement d'énoncés dont les expériences ont été présentées dans ce chapitre qui se termine ici.

9 Évaluation et Analyses des résultats

Dans ce chapitre, nous évaluons les modèles que nous avons entraînés pour les tâches de classification d'énoncés en ADs (CLEADs) et d'appariement d'énoncés (AE).

9.1 Classification d'énoncés en ADs : CLEADs

Lors de nos expérimentations pour la CLEADs, nous avons entraîné différents modèles avec des architectures différentes avec ou sans des couches d'attention, des couches *Bi-LSTM* et en sortie des couches CRF ou une simple couche dense avec la fonction d'activation *Softmax*. L'analyse des résultats (cf.figure 9.1) de nos expérimentations pour cette tâche a fait ressortir, trois modèles distincts les uns des autres qui performent chacun sur des données de tests respectivemenst sur les trois niveaux d'actes de dialogues.

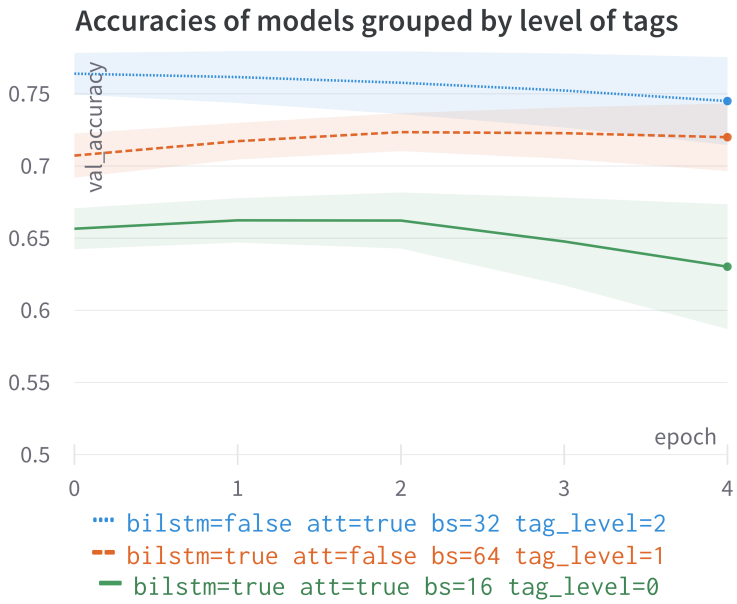
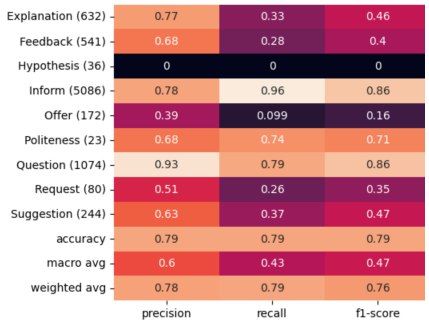


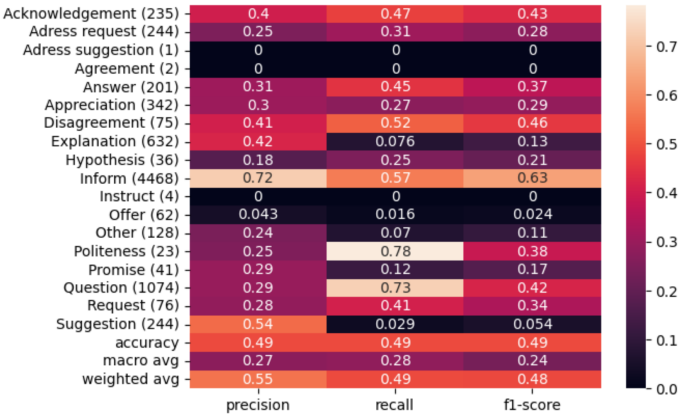
Figure 9.1 – Performances de CLEADs sur les 3 niveaux de granularité

Sur cette figure 9.1, les 3 lignes du bas indiquent l'architecture de chacun de ces modèles qui sont tous des affinages d'un même modèle pré-entraîné *BERT*. Les niveaux de granularités sont de 0, 1 2 respectivement pour les ADs fins, moyens et basiques. Nous désignons respectivement par **Att_2**, **Bi_LSTM_1** et **Bi_LSTM_Att_0** les modèles qui performent le mieux sur chacun des niveaux de granularité des ADs précisé dans ces désignations. Dans ces dernières, les expressions *Att* (couche d'attention) et *Bi_LSTM* font référence aux couches ajoutées sur le modèle pré-entraîné. Ce sont ces trois modèles que nous évaluons ici.

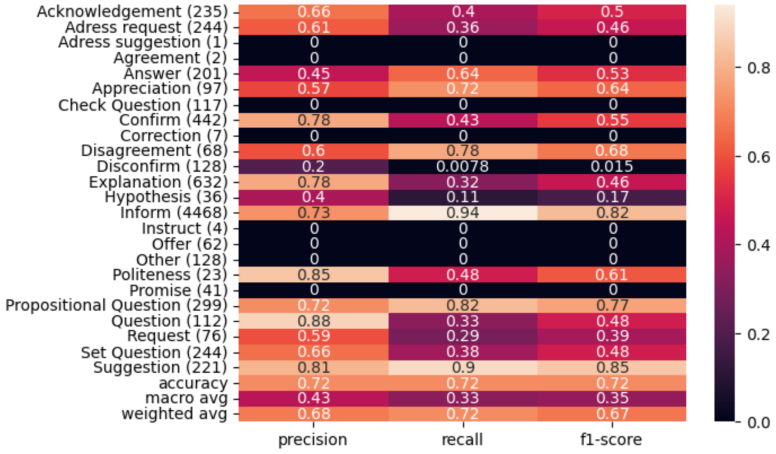
Pour leur évaluation, on utilise les métriques de précision, rappel et f1-score sur des données d'évaluation (avec 7 964 énoncés) qui ont été produites en suivant le même protocole de filtrage sur le corpus de base MRDA pour l'obtention des données d'entraînement et de test. Nous évaluons chacun des modèles sur ces données d'évaluation avec les différents niveaux d'ADs.



(a) Modèle **Att_2**



(b) Modèle **Bi_LSTM_1**



(c) Modèle **Bi_LSTM-Att_0**

FIGURE 9.2 – Rapport de classification pour la tâche CLEADs sur les trois niveaux de granularité

Sur ces graphiques de la figure 9.2, nous constatons que les performances des modèles sont d'autant meilleures que la granularité des ADs est moins fine, ce qui est tout à fait logique car plus la granularité est fine, plus on a d'ADs et moins les probabilités de distribution sur ces ADs sont élevées. Certains scores sont très faibles ou nuls, ils correspondent à des ADs spécifiques (*hypothèse, offre, feedback, accord, prise en compte de suggestion ou de requête, instruction*) que nous avons spécifiés en fonction de nos besoins d'AE. Ces faibles scores sont corrélés à la très faible représentation de ces ADs dans le corpus MRDA que nous avons utilisé pour la CLEADs. Il serait intéressant dans nos futurs travaux de trouver des données annotées en ces ADs qui sont d'une très grande importance pour l'appariement d'énoncés dans les conversations asynchrones, exactement comme la relation *question-réponse*.

9.2 Appariemment d'Énoncés (AE)

Pour évaluer notre modèle d'AE, nous définissons deux baselines qui utilisent les métadonnées dont principalement la relation *reply-to* entre les emails d'une conversation. D'une part nous utilisons BM25 (Robertson and 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, nous utilisons un système de similarité sémantique basé sur des modèles neuronaux. Pour ce second système, nous utilisons SentenceTransformers (Reimers and Gurevych, 2019). On constate que

	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 9.1 – Comparaison de notre modèle avec nos baselines

notre approche performe mieux que les baselines BM25 et Sentence-BERT avec des différences de points respectivement de 16 et 9 sur ces baselines. Ceci montre que le simple calcul de similarité entre deux énoncés d'une conversation d'emails, y compris leurs représentations contextuelles issues des modèles pré-entrainés (Sentence-BERT), n'est pas suffisant pour établir une relation transverse entre les deux énoncés en question. À contrario, notre approche a un taux d'erreur d'environ 26% dans l'établissement de ce type de relation entre les énoncés d'emails du corpus de test de BC3. La petitesse de taille de nos données de test ne nous permet pas d'analyser combien notre approche pourrait être robuste. Cette robustesse, avant même d'être analysée, devrait être prendre corpus avec des volumes de données beaucoup plus importants que ceux utilisés pour entraîner nos modèles.

9.2.1 Évaluation sur des emails d'Orange

Nous avons mené des expériences pour la tâche d'appariement d'énoncés sur des corpus d'emails (BC3) et de forum (Reddit) en anglais. Mais nous n'avons pas oublié qu'au départ notre problématique porte sur des emails en langue française et précisément sur ceux de l'entreprise Orange. Nous n'avons pas pu exploiter le corpus d'emails collectés chez Orange parce que son annotation aurait

pris beaucoup de temps. Toutefois, nous avons constitué un très petit corpus de 26 paires d'énoncés positives et 23 paires négatives, pour un total de 49 paires d'énoncés extraits d'une dizaine de conversations d'emails. La figure 9.3 présente des exemples de ces paires d'énoncés. Il est important de noter que cette taille de donnée est insignifiante et constitue un biais dans les analyses que nous effectuons.

Pour évaluer notre approche de modèle joint sur ces données d'Orange, nous traduisons en langue française les données des corpus BC3 et Reddit avec l'API de *DeepL* [1] que nous avons ensuite utilisée pour entraîner notre modèle en affinant *CamemBERT*. Nous nommons **CamemBERT-joint** le modèle affiné avec les données en français et les deux niveaux de granularité du listing 8.2. **CamemBERT-joint** a une performance de $\mathbf{0.57}$ et $\mathbf{0.47}$ avec respectivement les niveaux de granularité basique et fin. Ces performances montrent, contrairement à l'hypothèse selon laquelle plus la granularité d'ADs est fine, meilleures sont les performances de la couche d'AE : hypothèse justifiée dans les expériences avec les modèles pipeline comme le montre la table 8.4. De même dans la section 8.4.2, nous montrons qu'avec notre modèle joint, il existe une relation linéaire ascendante entre les performances des couches pour les tâches de CLEADs et d'AE. Une première explication de la non justification de cette hypothèse dans le cadre des données traduites en français serait liée aux biais introduits lors du processus de traduction automatique. Il est fort probable que certaines phrases auraient perdu leur sens dialogique lors de cette phase de traduction.

Une seconde explication se trouverait dans le fait qu'un modèle de classification a généralement de meilleures performances selon qu'il a un nombre réduit de catégories à prédire. Et dans nos expériences, on a plus d'ADs dans la granularité fine que dans la basique. La couche de CLEADs avec le niveau de granularité basique a ainsi de meilleures performances qu'avec le niveau de granularité fin. Et vu qu'on entraîne un modèle joint dont les paramètres sont mis à jour en fonction de la somme de deux fonctions de perte pour la CLEADs et l'AE, on peut déduire une relation entre le niveau de granularité des ADs et les performances de notre modèle joint *CamemBERT-joint*. Dans cette relation, moins fins sont les ADs, meilleurs sont les performances du modèle. Ce qui est contraire à l'hypothèse justifiée par nos expériences et analyses pour les données en anglais. On observe qu'il faut probablement être fin sur les actes qui sont utiles pour l'AE et plus grossier ailleur. Cette seconde interprétation laisse penser que le biais viendrait des traductions automatiques.

1. https://www.deepl.com/en/docs-api

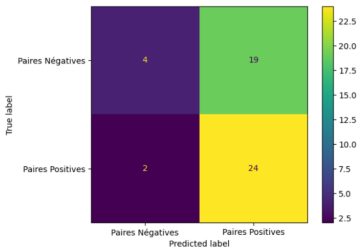
	Conv_ids	Utterance_1	Utterance_2	Class
1	1	SENIOR, Comme convenu, je ferai du télétravail cet après-midi.	OK	1
2	2	En analysant de près mon fichier d` export, la base de données et les inputs de Tool1, voici la liste des bugs que j` ai identifiés sur les distances sémantiques :	Et ben , ce n` est pas gagné tout ça	1
3	2	En analysant de près mon fichier d` export, la base de données et les inputs de Tool1, voici la liste des bugs que j` ai identifiés sur les distances sémantiques :	J` hésite à initier un fichier pour recenser tous ces bugs ce qui aurait l` avantage de pouvoir le partager et de concaténer toutes nos remontées.	1
4	2	En analysant de près mon fichier d` export, la base de données et les inputs de Tool1234, voici la liste des bugs que j` ai identifiés sur les distances sémantiques :	Sinon toujours pas de news d` OAB de mon côté.	0
5	2	En analysant de près mon fichier d` export, la base de données et les inputs de Tool1, voici la liste des bugs que j` ai identifiés sur les distances sémantiques :	Cependant pour ton information, j` ai appris ce matin que USER1 avait fini par obtenir la contribution de Ludo sur la partie Tool2 (voire sur la partie search) jusqu` au mois d` août et à plein temps.. Ce serait le temps qu` ils trouvent une autre personne pour le remplacer à plus longs termes.	0
6	2	En analysant de près mon fichier d` export, la base de données et les inputs de Tool1, voici la liste des bugs que j` ai identifiés sur les distances sémantiques :	PS : merci pour cette analyse précise :)	1
7	2	Et ben , ce n` est pas gagné tout ça ...	Pour les distances sémantiques, ce n` est pas surprenant car cette partie des spécifications était difficile à comprendre, et il n` y avait pas d` exemples.	1
8	2	Mais malheureusement pas de news pour notre plugin Tool3 :(.	Pour les distances sémantiques, ce n` est pas surprenant car cette partie des spécifications était difficile à comprendre, et il n` y avait pas d` exemples.	0
9	2	Tu prends des congés de ton côté ?	Je suis donc en train d` essayer d` expliquer ça via un algo.	0
10	3	Pour info	De mon point de vue, j` espère que cette nouvelle période (avril) va être conservée dans les années suivantes pour le Salon de la Recherche ; on ne pouvait pas faire tellement pire que début décembre !!	0
11	4	Hello, Voici le jeu de données que j` ai réalisé pour des évaluations.	Concernant l'objectif visé, je suggère ceci : « Quels sont, selon vous, les messages qui ont un rapport direct ou indirect avec l` objectif visé : "production d` un article scientifique pour CONF 2018" » ?	0
12	3	Pour info	Ok.	1
13	4	Concernant l'objectif visé, je suggère ceci : « Quels sont, selon vous, les messages qui ont un rapport direct ou indirect avec l` objectif visé : "production d` un article scientifique pour CONF 2018" » ?	Suite à notre discussion de ce matin, voici le fichier avec tes modifications suggérés et les interlocuteurs pseudoanonymisés.	0
14	5	Idéalement, il faudrait qu` on puisse en discuter demain soir afin que j` apporte les modifications nécessaires et que je crée la DI sur anaqua.	Quelques compléments dans le doc.	1
15	8	Voici les slides de résultats manuelles, les analyses sont en commentaires (et sinon en mode animation c` est plus pratique pour comprendre certaines diapo).	J` avoue que sans l` explication de texte j` ai un peu de mal... entre les GS-IeCTs, IeCTs, IaCTs, GP :) ...	1

FIGURE 9.3 – Quelques exemples d'énoncés construits avec la classe 1 pour les paires positives et 0 pour les négatives

La table 9.2 présente les rapports de classification et matrices de confusion de *CamemBERT-joint* sur des données de test d'Orange. De façon globale, on voit que *CamemBERT-joint* est plus apte à prédire des paires d'énoncés comme des paires positives. Sur les rapports de classification, on constate des valeurs de rappel intéressantes pour les paires positives. Le rappel en question est d'autant plus intéressant (0.92) avec les actes de dialogues basiques.

CamemBERT-Joint avec les ADs basiques

	precision	recall	f1-score	support
0	0.67	0.17	0.28	23
1	0.56	0.92	0.70	26
accuracy			0.57	49
macro avg	0.61	0.55	0.49	49
weighted avg	0.61	0.57	0.50	49



CamemBERT-Joint avec les ADs fins

	precision	recall	f1-score	support
0	0.40	0.26	0.32	23
1	0.50	0.65	0.57	26
accuracy			0.47	49
macro avg	0.45	0.46	0.44	49
weighted avg	0.45	0.47	0.45	49

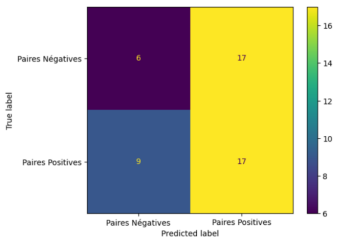


TABLE 9.2 – Rapport de classification et matrice de confusion de *CamemBERT-joint* sur des données de test d'Orange

Bien que la figure 9.2 donne des informations sur comment le modèle se comporte sur des données d'Orange, il est intéressant de regarder de façon approfondie des exemples de faux positifs et faux négatifs pour les deux variantes de *CamemBERT-joint* dont l'un est entraîné avec les ADs basiques, et l'autre avec des ADs fins.

Ci-dessous quelques exemples de paires d'énoncés avec de faux positifs (FP), faux négatifs (FN), vrais négatifs (VN) et vrais positifs (VP) extraits des résultats de *CamemBERT-joint* avec des ADs basiques, avec entre paranthèse devant chaque énoncé son AD prédit. Il est important de noter ici que les énoncés de chaque paire sont extraits d'une même conversation.

(a) (FP) : *(requête)* Dites-moi ce que vous en pensez :) <-> (*question*) Sinon j'ai du mal à comprendre pourquoi tu n'appliques pas les mêmes seuils pour identifier les liens à étudier.

(b) (FP) : *(réponse)* À relire et à valider avant dépôt dans Tool4 <-> *(requête)* Anna, Legrand et moi avons complété le mémoire simplifié (voir doc ci-joint)."

(c) (VP) : *(offre)* Je vous laisse le relire tranquillou demain. Idéalement, il faudrait qu'on puisse en discuter demain soir afin que j'apporte les modifications nécessaires et que je crée la DI sur anaqua. <-> *(réponse)* J'ai relu jusqu'au bas de la page 3.

(d) (VP) : *(question)* Sinon j'ai du mal à comprendre pourquoi tu n'appliques pas les mêmes seuils pour identifier les liens à étudier. <-> *(feedback)* En espérant que ce soit un peu plus claire, je reste dispo. et sinon on en reparle lundi ;).

(e) (FN) : *(suggestion)* En analysant de près mon fichier d'export, la base de données et les inputs de Tool1234, voici la liste des bugs que j'ai identifiés sur les distances sémantiques : <-> *(explication)* Sinon toujours pas de news d'OAB de mon côté.

(f) (FN) : *(politesse)* Tu prends des congés de ton côté ? <-> *(offre)* Je suis donc en train d'essayer d'expliquer ça via un algo.

(g) (VN) : *(question)* "Concernant l'objectif visé, je suggère ceci : « Quels sont, selon vous, les messages qui ont un rapport direct ou indirect avec l'objectif visé : production d'un article scientifique pour CONF 2018" » ? <-> *(answer)* Suite à notre discussion de ce matin, voici le fichier avec tes modifications suggérés et les interlocuteurs pseudoanonymisés.

(h) (VN) : *(suggestion)* "Comme discuté au daily, voici la proposition de calcul pour l'approche 1 basée sur l'historique des invitations envoyées à une personne ." <-> *(suggestion)* Ce point ne sera peut-être pas traité dans le sprint courant, car il faut pour cela développer le mode au fil de l'eau.

Sur ce petit échantillon, on peut remarquer que pour certains énoncés, les ADs prédits reflètent la réalité des intentions de ces énoncés. Ces le cas dans (a), (c), (d), (g) et (h). Malgré de bonnes prédictions d'ADs en (a), le modèle s'est trompé sur l'AE des énoncés, probablement à cause d'une absence de contexte, parce le premier énoncé dans (a) est bel et bien une requête concernant quelque chose dans un segment précédent qui aurait dû faire partie de la dite requête et peut-être que le modèle se serait comporté autrement. Le même constat peut être fait pour (b). La lecture des exemples (c) et (d), qui sont de vrais positifs, ne laisse par contre pas transparaître un manque d'information ou de contexte qui limiterait fortement leur compréhension.

(e) et (f) sont aussi des exemples avec des énoncés qui nécessitent de penser à des contextes qui amélioreraient leur compréhension, à part peut-être le second membre de (f). Sur ces deux exemples, il ressort clairement des dissimilarités sémantiques entre les énoncés de chacune des paires. Pour ces deux exemples, les ADs prédits pour les énoncés sont très loin de la réalité, en (f) une question est prédite comme étant un formule de politesse et un partage d'informations comme étant une offre. De même en (e), les énoncés de partage d'information sont respectivement considérés comme une suggestion et une élaboration ou explication par *CamemBERT-joint*. L'élaboration est peut-être prédite pour le second énoncé de (e) à cause la conjonction » Sinon « qui, en général, introduit de la contradiction lorsqu'on élabore un raisonnement. Au vu des dissimilarités sémantiques mentionnées précédemment sur lesquelles *CamemBERT-joint* aurait dû s'appuyer pour bien classer ces paires d'énoncés, il a pris en compte les ADs prédits qui ne sont pas adaptés dans ces exemples et qui ont fort probablement constitué du bruit pour un bon AE.

Contrairement à (e) et (f) où les ADs prédits ne sont pas adaptés, (g) et (h) ont leurs ADs d'énoncés plutôt proches de la réalité, à part le second énoncé de (g) qui est un partage d'information. Dans (g) et (h), il est difficile même pour un humain d'établir une sorte de dissimilarité sans avoir le contexte général de la conversation qui a permis de constituer ces paires. (g) et (h) ont des énoncés de taille (en termes de nombre de mots) un peu plus consistant que ceux de (f) et le second de (e). Cette taille d'énoncés aurait peut-être un impact sur la prédiction des ADs et même l'AE, car plus on a de mots dans une phrase, plus on a du contexte et mieux on peut prédire les ADs adaptés et les AE. Il serait intéressant d'explorer cette hypothèse sur un plus grand corpus d'énoncés extraits de conversation d'emails. (g) et (h) sont des vrais négatifs, on peut ici déduire que la couche d'AE dans *CamemBERT-joint* a tiré profit des ADs prédits et proches de la réalité, mais aussi de la dissimilarité

sémantique, même si celle-ci n'est pas apparente pour un humain, pour bien classer les paires de ces exemples.

Au vu des rapports de classification et des matrices de la figure 9.2, on constate que *CamemBERT-joint* avec les actes de dialogues fins (14 ADs) a de moins performances par rapport à la version avec moins d'ADs. Ceci est principalement dû au nombre de classes d'ADs à prédire qui est plus important (on passe de 9 ADs à 14) et a un impact sur la couche les performances de la couche de classification en ADs qui va se propager jusque dans la couche d'AE qui s'appuie fortement sur les ADs prédits. De même, la sous représentativité de certaines classes d'ADs dans les données d'entraînement est aussi un des facteurs qui contribue à ces mauvaises perfomances de *CamemBERT-joint* avec les actes de dialogues fins. Et pourtant, la finesse des ADs d'énoncés permet un meilleur AE de ceux-ci.

Vu la petitesse de nos données d'entraînement et de test, nous sommes limités dans les analyses que nous pouvons effectuer pour en tirer des conclusions fortes et définitives. Toutefois, certains aspects comme la prise en compte de contexte, des données équilibrées en fonction des ADs utilisés sont des facteurs à fortement considérer, en plus de corpus plus volumineux, si l'on veut améliorer les performances et la robustesse de tels modèles.

9.3 Conclusion

Les évaluations présentées dans ce chapitre portent sur les deux tâches de classification d'énoncés de conversations en ADs (CLEADs) et d'appariement de ces mêmes énoncés. Pour la CLEADs, nous présentons les rapports de classification de nos trois meilleurs modèles, chacun en fonction de son niveau de granularité d'ADs les mieux adaptés pour notre principale tâche d'AE. Pour les niveaux de granularité moyen et fin, on note de faibles performances proches parfois de zéro pour certains ADs. Ces performances sont principalement dues à la faible représentation des ADs en question dans les données d'entraînement. Pour améliorer ces performances, il conviendrait d'équilibrer les corpus en fonction des différents ADs avec des techniques d'augmentation de données, de sous ou de sur-échantillonnage des données.

Pour la seconde tâche d'AE, nous avons évalué nos modèles spécifiquement sur un petit corpus de 49 paires d'énoncés (26 positives et 23 négatives) que nous avons constitués à partir d'une dizaine de conversations extraites du corpus d'Orange. Le modèle joint *CamemBERT-joint* est celui que nous évaluons avec des données d'Orange. *CamemBERT-joint* est entraîné avec des données de Reddit et BC3, traduites de façon automatique de l'anglais vers le français. Les performances de *CamemBERT-joint* semblent être meilleures pour identifier les paires d'énoncés positives d'une part et d'autre part avec le niveau de granularité d'ADs basiques. Ceci ne pourrait se confirmer qu'avec des modèles plus robustes parce qu'entraînés sur des plus gros volumes de données.

Quatrième partie

Conclusion et perspectives

10 Conclusion et Perspectives

10.1 Conclusion

Dans ce manuscrit, nous nous intéressons à la constitution de sous-fils de conversation asynchrone en complétant des approches standards à base de métadonnées et similarité par l'exploitation d'actes de dialogue en particulier pour apparier des énoncés qui à son tour s'appuie sur l'identification des actes dialogue. Lorsque des suites d'appariements sont identifiées dans une conversation avec des énoncés en commun, elles constituent ainsi des sous-fils de conversation avec des structures linéaires. Entre l'introduction générale dans laquelle nous avons posé toutes les bases pour nos travaux et cette conclusion, nous avons présenté successivement les travaux connexes à notre problématique. Ensuite nous avons exposé ce que sont les Communications Médiées par Ordinateur (CMO), ainsi que les concepts d'analyse de discursive liées à ces CMO, en mettant cependant l'accent sur l'évolution des actes de dialogue. Enfin nous présentons les différents corpus que nous avons utilisés pour nos travaux et les différentes expériences menées avec leurs résultats et analyses respectifs.

Travaux connexes

La constitution de sous-fils de conversations comme nous l'avons présenté dans le chapitre 1 a été abordée sous différents angles avec des approches qui s'appuient sur les métadonnées comme les sujets d'email d'une part et d'autre part sur les contenus de conversations, et puis comme nous des travaux qui prennent en compte ces deux sources d'information. Les travaux de Dans la littérature, c'est la problématique de reconstruction de fils qui consiste à mettre ensemble des contenus d'emails qui appartiendraient à une même conversation ou qui traiteraient d'un même sujet d'où les approches de clustering. L'algorithme de Zawinski est l'un des algorithmes les plus populaires pour la construction de fils de conversation d'emails et est basé sur des informations provenant de métadonnées. Ces reconstructions de fils de conversation sont aussi abordées pour retrouver la structure arborescente ou linéraire originale d'une conversation. Le démêlage de conversation est aussi une problématique lié aux conversations et qui comme la reconstruction de fils de conversation reste à un niveau macro ou semi-macro pour des parties de contenus d'emails sont regroupées entre elles parce qu'elles partagent les mêmes thématiques ou sous-thématiques. L'identification de ces dernières reste une problématique sur laquelle nombre de chercheur continue à travailler. Pour nous la constitution de sous-fils de conversations reste à un niveau très fin interne à une seule conversation dans laquelle des segments très courts comme des phrases sont mis en relation (de façon transverse sur les emails). Ces relations transverses peuvent capturer de la similarité sémantique mais généralisent aux cas où la similarité n'est pas flagrante mais où on a un enchaînement naturel en terme d'actes de dialogue.

CMO et Analyse discursive

Communications médiées par ordinateur : Ici, une analyse approfondie des communications médiées par ordinateur est faite, en mettant en évidence les avantages et les inconvénients des CMO synchrones et asynchrones. Les caractéristiques linguistiques, syntaxiques et paralinguistiques des emails et des forums sont également examinés, car ces types de CMO sont largement utilisés et analysés dans le cadre des travaux présentés.

Analyse discursive : L'analyse discursive est abordée, en commençant par les premières taxonomies d'actes de langage jusqu'à la norme ISO 24617-4, qui permettent d'annoter les conversations écrites, parlées ou transcrites. Ces taxonomies et normes permettent de mieux comprendre et capturer les différentes caractéristiques linguistiques des conversations à travers les différentes couches, dimensions et relations de dépendance entre elles. Ces travaux ont ouvert la voie à la conception de systèmes automatiques de dialogue et de communication intelligente.

Bien que de nombreux travaux aient été réalisés dans le domaine de l'analyse discursive des conversations, répondre de manière satisfaisante à certaines problématiques dans ce domaine reste un défi. Cependant, les aspects abordés dans ce chapitre, en particulier la taxonomie des actes de dialogue, seront extrêmement utiles pour annoter les corpus et catégoriser les segments de texte d'emails en actes de dialogue. Nous nous sommes fortement appuyés dans cette thèse sur ces actes de dialogue pour résoudre la problématique d'appariement d'énoncés.

Corpus et expériences

Constitution de corpus en entreprise : Le chapitre présente un processus méthodologique de constitution de corpus en entreprise, incluant des étapes telles que la collecte d'emails, l'annotation manuelle et la pseudo-anonymisation des données. Ce processus a été utilisé pour créer un corpus chez Orange, mais en raison de contraintes de temps et des questions de confidentialité des données, d'autres corpus déjà annotés ont finalement été explorés pour résoudre notre problématique.

Exploitation de corpus annotés : Différents corpus ont été explorés et exploités pour la résolution de notre problématique. Cela inclut le corpus de discussion de pages Wikipédia en français, qui a été abandonné en raison de son focus trop marqué vers les processus de vote, qui diffèrent des conversations d'emails en entreprise. Les corpus d'emails BC3 et Enron, le corpus de discussions du forum Reddit, et le corpus MRDA de transcriptions d'enregistrements de réunions ont été sélectionnés pour leurs annotations pertinentes pour l'analyse conversationnelle.

Classification d'énoncés en ADs et Appariement d'énoncés : Dans cette partie, des expériences sont menées pour la classification d'énoncés en actes de dialogues et l'appariement d'énoncés. Différents paramètres et hyperparamètres ont été pris en compte. Les résultats ont montré l'importance et la corrélation de ces paramètres sur les performances des modèles. L'utilisation de la couche d'attention seule s'est avérée plus performante, et l'utilisation des CRF en sortie n'a pas amélioré les résultats. Les conclusions ont mis en évidence l'importance des actes de dialogues pour l'appariement des énoncés et l'identification des points en suspens dans les emails d'entreprise. Deux approches, l'une en pipeline et l'autre en modèle joint multitâche, ont été explorées pour la tâche d'appariement d'énoncés.

Ces résumés ouvrent la voie à de potentielles améliorations et pistes de recherche futures, comme l'utilisation de modèles de correspondances d'actes de dialogues et l'exploration de l'utilisation de corpus complémentaires pour les différentes tâches.

10.2 Perspectives

10.2.1 Ingénierie de requêtes et Larges modèles de langages

Suite à une expérimentation préliminaire favorable effectuée lors de nos travaux qui a consisté à segmenter des contenus d'emails en unités cohérentes. Ces unités ont ensuite été étiquetées en actes de dialogues. Toutes ces tâches ont été effectués à l'aide de l"ingénierie de requêtes (*prompt engineering*) avec les larges modèles de langages existants tels GPT-3.5-turbo (Brown et al., 2020),*InstructGPT (Davinci)*. Il serait intéressant d'explorer davantage ces approches pour constituer, par exemple à partir d'Enron (Klimt and Yang, 2004b), un corpus de taille conséquente finement annoté en ADs afin d'améliorer les résultats de nos modèles. La tâche d'appariement d'énoncés pourrait également être examinée avec cette nouvelle technique d'ingénierie de requêtes (*prompt engineering*). L'exploitation de ce corpus d'une taille conséquente permettrait d'évaluer de façon fine la robustesse de notre approche pour la tâche d'établissement d'une relation transverse entre les énoncés d'emails dans une conversation.

10.2.2 Cas d'utilisation

Comme nous avons mentionné en introduction de ce manuscrit, les CMO en général et les emails en particulier présentent des avantages et des inconvénients. Et comme inconvénients de l'utilisation des emails, nous avons relevé la surcharge d'informations, la difficulté d'accès à la bonne information de façon rapide et efficiente, pour ne citer que ceux-ci. Les approches d'identification d'actes de dialogues et d'appariement d'énoncés que nous proposons dans nos travaux permettraient si intégré dans un client mail de palier partiellement à ces inconvénients. Ces approches donneraient la possibilité à un collaborateur de facilement suivre ses contributions dans l'évolution d'un projet via les actes de dialogues identifiés, ces derniers pourraient aussi constituer des listes de tâches à effectuer (*todo list*) pour des professionnels. Par le biais de l'appariement des énoncés de façon transverse sur des emails dans une conversation, celle-ci pourrait être présentée au travers d'agents conversationnels (modulo les questions de confidentialité). Aussi la possibilité de visualisation des threads de discussion avec une indication de statut (clos, demande en attente, facilitant ainsi l'accès à l'information. De même la recherche d'informations serait facilitée via l'ajout des ADs comme facettes de recherche et les recherches *full-text* pourraient être dirigées sur des segments de texte d'amorce, de milieu ou de fin de dialogue.

10.2.3 Déclinaisons des travaux

Ces travaux que nous avons menés peuvent aussi s'appliquer à des conversations synchrones. Les transcriptions de réunions en sont des exemples dans lesquelles plusieurs interlocuteurs interviennent sur des tours de paroles de façon instantanée, et donc l'appariement des énoncés dans de telles transcriptions aiderait à mieux les organiser et ainsi faire ressortir des sous-fils de conversations qui seraient des suites de paires d'énoncés liées les unes aux autres. Ces sous-fils de conversations peuvent aussi se décliner comme des clustering de tours de paroles. La tâche d'appariement d'énoncés peut aussi être vue comme une tâche de sélection de réponse avec ici les questions et réponses qui n'en sont pas dans le sens strict du terme, mais plutôt de simples énoncés. On peut même aller

plus loin à la considérer comme une tâche d'implication (*entailment*) entre les deux énoncés à apparier.

Cinquième partie

Annexes

A OutlookScrapping : Outil de collecte de données et ses fonctionnalités

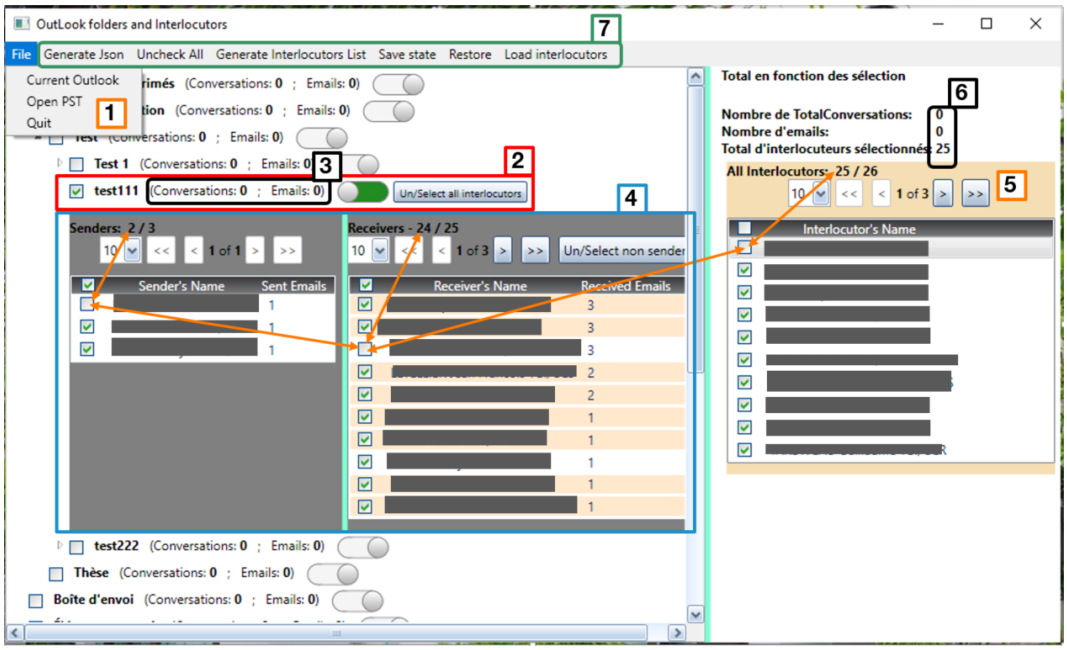


FIGURE A.1 – Un aperçu de *OutlookScrapping*, l'outil développé pour la collecte des métadonnées et des conversations d'emails

La figure A.1 montre l'interface visuelle de *OutlookScrapping* ; ses différentes fonctionnalités sont détaillées ci-dessous :

1. Le chiffre 1 encadré de couleur orange montre les trois sous-menus de l'onglet *File*. L'objectif principal ici pour les deux premiers éléments est de parcourir les dossiers du contexte d'Outlook et de les répertorier sous forme d'arborescence avec des dossiers imbriqués (grande colonne à gauche de la figure A.1) exactement comme dans l'application de messagerie Outlook :

    a) **Current Outlook** permet de spécifier que les données à traiter sont celles de l'application de messagerie Outlook en cours.

   b) **Open PST** affiche une boîte de dialogue dans laquelle on peut sélectionner un fichier de sauvegarde Outlook et le charger en tant que contexte par défaut d'Outlook.
   c) **Quit** : comme son nom l'indique, il permet de quitter l'application
2. Une fois la liste des dossiers chargée, le collaborateur référent peut sélectionner le dossier dans lequel il souhaite collecter les données. Ainsi, s'il a un dossier contenant des données privées, il ne le sélectionnera probablement pas. Une fois les dossiers sélectionnés, un bouton *slider* est tiré et un bouton *Un/Select all interlocutors* apparaît (encadré en rouge (*partie 2*) sur la figure A.1) et le contenu du dossier est déroulé (*partie 4*). Le bouton *slider* masque le contenu d'un dossier pour rendre l'interface de l'application moins touffue au cas où plusieurs dossiers sont sélectionnés. Le bouton *Un/Select all interlocutors* sélectionne ou désélectionne tous les interlocuteurs dans le contenu du dossier sélectionné et cela se répercute sur tous les autres dossiers sélectionnés.
3. Lorsqu'un dossier est sélectionné, des calculs sont effectués en arrière-plan pour mettre à jour le nombre d'emails et de conversations qu'il contient (**partie 3**). Ces calculs permettent également de mettre à jour les informations générales (**partie 6**).
4. La **partie 4** de la figure A.1 représente le contenu d'un dossier sélectionné divisé en deux colonnes. La colonne de gauche (*Senders*) contient les émetteurs sélectionnés (*Sender's Name*) qui ont envoyé au moins un email et le nombre d'emails que chacun d'eux a envoyés (*Sent Emails*). La colonne de droite (*Receivers*) liste les destinataires (*Receiver's Name*) des emails échangés et, devant chacun d'eux, le nombre d'emails auxquels ils ont participé en tant que destinataires (*Received Emails*). Ce sont ces chiffres qui nous permettent de calculer le nombre d'emails et de conversations par dossier sélectionné et de façon globale. La pagination et le tri ont été implémentés partout où il y a une liste d'interlocuteurs afin de faciliter certaines manipulations. Il est également possible de sélectionner ou désélectionner tous les interlocuteurs de chacune de ces listes d'expéditeurs et de destinataires.
5. Les flèches en couleur orange de la figure A.1 montrent comment est mis à jour le nombre total d'interlocuteurs sélectionnés ou non au sein d'un dossier et de façon globale (**partie 6**) si plusieurs dossiers sont sélectionnés. Toute opération de sélection ou de dé-sélection d'un interlocuteur induit des mises à jour non seulement sur les informations du dossier dans lequel la manipulation a eu lieu mais aussi dans les autres dossiers sélectionnés. Ceci en respectant le paradigme de sélection ou de dé-sélection.
6. La **partie 5** de la figure énumère tous les interlocuteurs impliqués en tant qu'émetteurs ou destinataires d'emails, ce qui permet de sélectionner facilement les collaborateurs qui ont donné leur consentement sans avoir à parcourir chaque dossier et ses sous-dossiers. Il permet également à un référent de trouver et de désélectionner facilement un interlocuteur avec lequel il a eu une conversation privée, de sorte que cette conversation ne soit pas collectée.
7. La **partie 7** est la barre d'onglets, chaque onglet ayant une fonction spécifique :
   a) **Generate Json** est l'onglet principal de l'application car c'est celui qui, après toutes les manipulations décrites ci-dessus, permet de générer deux fichiers (JSON et CSV). Chacun d'entre eux contient des informations sur tous les emails collectés. Sender, senderId, receivers, receicersID conversationIndex, conversationID, HTMLBody, EmailBody, subject, sent date and received date sont toutes les informations que nous collectons pour chaque email.

b) **Uncheck All** permet de désélectionner tous les dossiers en les décochant.

c) **Generate interlocutors List** génère une liste contenant tous les interlocuteurs sélectionnés. Cette liste peut être transférée et réutilisée par un autre référent à l'aide de l'onglet *Load interlocuteurs* pour accélérer les manipulations.

d) **Save state** ; certaines boîtes aux lettres contiennent parfois trop d'emails avec différents dossiers et interlocuteurs. Lorsqu'une telle boîte de messagerie est chargée dans notre outil, il est possible d'avoir beaucoup de manipulations de dé/sélection et cela prend beaucoup de temps. Cette fonctionnalité **Save state** permet de sauvegarder l'état de manipulation d'un contexte d'emails Outlook et de le restaurer plus tard avec l'onglet **Restore**, afin d'effectuer des manipulations en plusieurs fois pour un meilleur résultat de collecte de données.

B Statistiques du corpus MRDA et ses actes de dialogues

B.1 Actes de dialogue

B.1.1 Étiquettes basiques

Dialogue Act	Labels	Count	%	Train Count	Train %	Test Count	Test %	Val Count	Val %
Statement	S	64233	59.36	45099	60.08	9571	57.30	9563	58.19
BackChannel	B	14620	13.51	10265	13.67	2152	12.88	2203	13.41
Disruption	D	14548	13.45	9739	12.97	2339	14.00	2470	15.03
FloorGrabber	F	7818	7.23	5324	7.09	1409	8.44	1085	6.60
Question	Q	6983	6.45	4640	6.18	1231	7.37	1112	6.77

Table B.1 – MRDA : Distribution des étiquettes de base

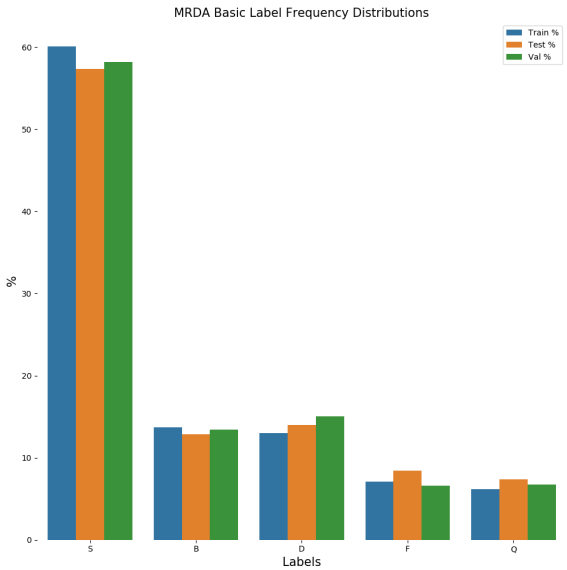


Figure B.1 – MRDA : Distribution des étiquettes basiques

B.1.2 Étiquettes générales

Dialogue Act	Labels	Count	%	Train Count	Train %	Test Count	Test %	Val Count	Val %
Statement	s	69873	64.58	48952	65.21	10472	62.70	10449	63.59
Continuer	b	15167	14.02	10606	14.13	2219	13.29	2342	14.25
Floor Holder	fh	8362	7.73	5617	7.48	1520	9.10	1225	7.45
Yes-No-question	qy	4986	4.61	3310	4.41	870	5.21	806	4.90
Interrupted/Abandoned/Uninterpretable	%	3103	2.87	2171	2.89	492	2.95	440	2.68
Floor Grabber	fg	3092	2.86	2076	2.77	489	2.93	527	3.21
Wh-Question	qw	1707	1.58	1110	1.48	310	1.86	287	1.75
Hold Before Answer/Agreement	h	792	0.73	474	0.63	134	0.80	184	1.12
Or-Clause	qrr	392	0.36	244	0.33	75	0.45	73	0.44
Rhetorical Question	qh	352	0.33	260	0.35	56	0.34	36	0.22
Or Question	qr	207	0.19	131	0.17	37	0.22	39	0.24
Open-ended Question	qo	169	0.16	116	0.15	28	0.17	25	0.15

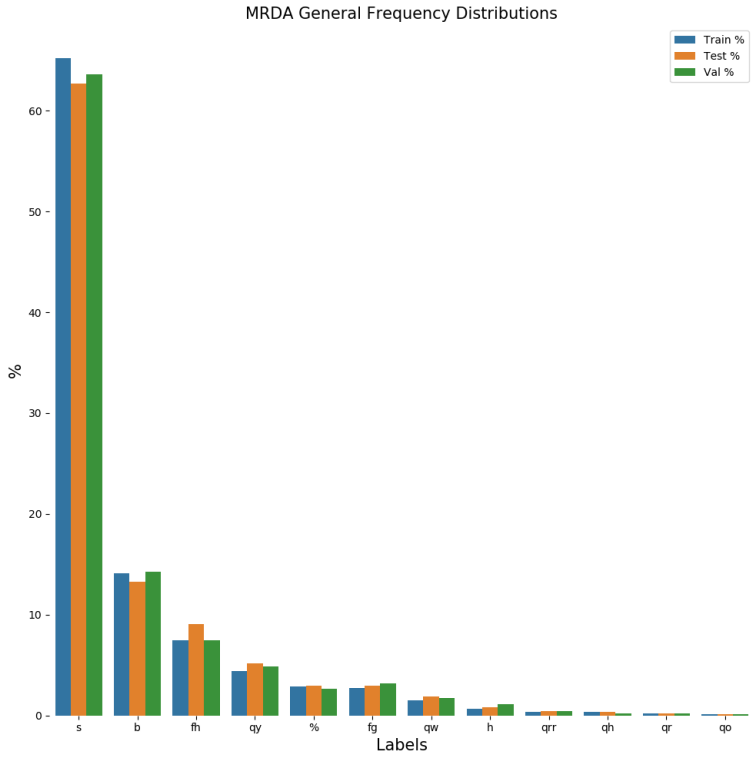


Figure B.2 – MRDA : Distribution des étiquettes générales

B.1.3 Étiquettes fines

Dialogue Act	Labels	Count	%	Train Count	Train %	Test Count	Test %	Val Count	Val %
Statement	s	33472	30.93	23238	30.96	4971	29.76	5263	32.03
Continuer	b	15013	13.87	10517	14.01	2175	13.02	2321	14.12
Floor Holder	fh	8362	7.73	5617	7.48	1520	9.10	1225	7.45
Acknowledge-answer	bk	7177	6.63	5117	6.82	1031	6.17	1029	6.26
Accept	aa	5898	5.45	4097	5.46	903	5.41	898	5.46
Defending/Explanation	df	3724	3.44	2790	3.72	531	3.18	403	2.45
Expansions of y/n Answers	e	3200	2.96	2360	3.14	540	3.23	300	1.83
Interrupted/Abandoned/Uninterpretable	%	3103	2.87	2171	2.89	492	2.95	440	2.68
Rising Tone	rt	3101	2.87	2015	2.68	516	3.09	570	3.47
Floor Grabber	fg	3092	2.86	2076	2.77	489	2.93	527	3.21
Offer	cs	2662	2.46	1878	2.50	402	2.41	382	2.32
Assessment/Appreciation	ba	2216	2.05	1605	2.14	354	2.12	257	1.56
Understanding Check	bu	2091	1.93	1405	1.87	371	2.22	315	1.92
Declarative-Question	d	1805	1.67	1153	1.54	350	2.10	302	1.84
Affirmative Non-yes Answers	na	1112	1.03	870	1.16	133	0.80	109	0.66
Wh-Question	qw	951	0.88	630	0.84	160	0.96	161	0.98
Reject	ar	908	0.84	594	0.79	152	0.91	162	0.99
Collaborative Completion	2	841	0.78	571	0.76	136	0.81	134	0.82
Other Answers	no	828	0.77	563	0.75	98	0.59	167	1.02
Hold Before Answer/Agreement	h	792	0.73	474	0.63	134	0.80	184	1.12
Action-directive	co	674	0.62	460	0.61	97	0.58	117	0.71
Yes-No-question	qy	669	0.62	476	0.63	90	0.54	103	0.63
Dispreferred Answers	nd	483	0.45	341	0.45	82	0.49	60	0.37
Humorous Material	j	463	0.43	326	0.43	67	0.40	70	0.43
Downplayer	bd	387	0.36	290	0.39	68	0.41	29	0.18
Commit	cc	371	0.34	258	0.34	51	0.31	62	0.38
Negative Non-no Answers	ng	351	0.32	236	0.31	56	0.34	59	0.36
Maybe	am	349	0.32	224	0.30	66	0.40	59	0.36
Or-Clause	qrr	345	0.32	216	0.29	66	0.40	63	0.38
Exclamation	fe	307	0.28	195	0.26	56	0.34	56	0.34
Mimic Other	m	293	0.27	200	0.27	48	0.29	45	0.27
Apology	fa	259	0.24	181	0.24	46	0.28	32	0.19
About-task	t	253	0.23	154	0.21	42	0.25	57	0.35
Signal-non-understanding	br	236	0.22	161	0.21	39	0.23	36	0.22
Accept-part	aap	219	0.20	158	0.21	27	0.16	34	0.21
Rhetorical-Question	qh	214	0.20	166	0.22	30	0.18	18	0.11
Topic Change	tc	212	0.20	127	0.17	35	0.21	50	0.30
Repeat	r	208	0.19	131	0.17	45	0.27	32	0.19
Self-talk	t1	198	0.18	120	0.16	38	0.23	40	0.24
3rd-party-talk	t3	165	0.15	105	0.14	36	0.22	24	0.15
Rhetorical-question Continue	bh	154	0.14	109	0.15	26	0.16	19	0.12
Reject-part	bsc	150	0.14	94	0.13	22	0.13	34	0.21
Misspeak Self-Correction	arp	150	0.14	89	0.12	18	0.11	43	0.26
Reformulate/Summarize	bs	141	0.13	89	0.12	17	0.10	35	0.21
"Follow Me"	f	128	0.12	98	0.13	12	0.07	18	0.11
Or-Question	qr	127	0.12	88	0.12	17	0.10	22	0.13
Thanking	ft	119	0.11	88	0.12	9	0.05	22	0.13
Tag-Question	g	87	0.08	58	0.08	9	0.05	20	0.12
Open-Question	qo	74	0.07	49	0.07	14	0.08	11	0.07
Correct-misspeaking	bc	51	0.05	29	0.04	13	0.08	9	0.05
Sympathy	by	11	0.01	5	0.01	2	0.01	4	0.02
Welcome	fw	6	0.01	5	0.01	0	0.00	1	0.01

B.2 Métadonnées

— Total number of utterances : 108202

— Max utterance length : 85

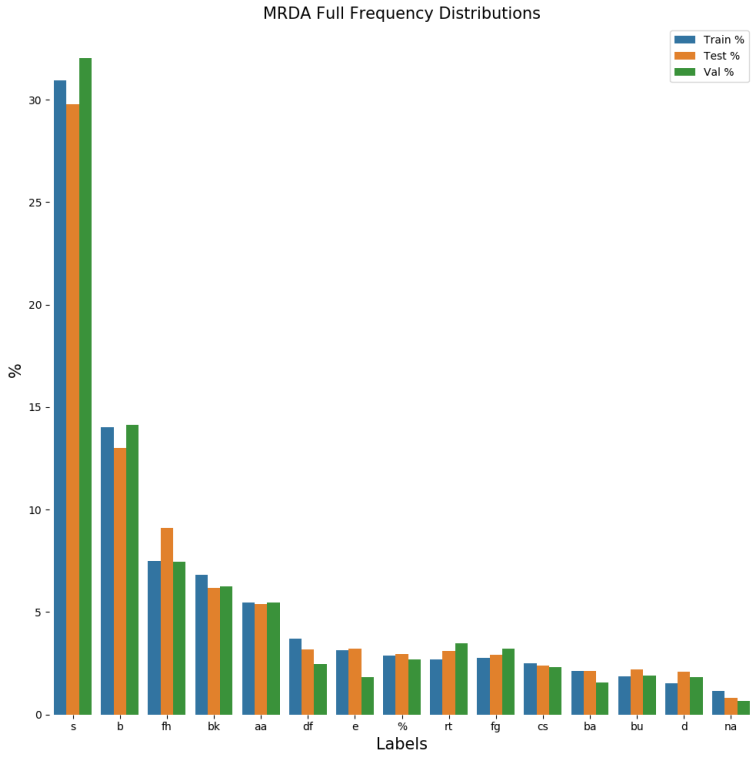


FIGURE B.3 – MRDA : Distribution des étiquettes fines

— Mean utterance length : 8.01
— Total Number of dialogues : 75
— Max dialogue length : 3391
— Mean dialogue length : 1442.69
— Vocabulary size : 10866
— Number of basic labels : 5
— Number of general labels : 12
— Number of full labels : 52
— Number of speakers : 52

B.3 RÉPARTITION DES DONNÉES

Données d'entraînement

— Number of dialogues : 51
— Max utterance length : 85
— Mean dialogue length : 1471.9

— Number of utterances : 75067

Données de test

— Number of dialogues : 12

— Max utterance length : 2028

— Mean dialogue length : 1391.83

— Number of utterances : 16702

Données de validation

— Number of dialogues : 12

— Max utterance length : 1969

— Mean dialogue length : 1369.42

— Number of utterances : 116433

C Description détaillée des actes de dialogues fins de MRDA

Group		Dialog acts	Labels	Description	Examples
STATEMENTS		Statement	s	Most widely used tag	
BACKCHANNELS / ACKNOWLEDGEMENT		Continuer (= Backchannel)	b	Utterances which function as backchannels are not made by the speaker who has the floor. Instead, backchannels are utterances made in the background that simply indicate that a listener is following along or at least is yielding the illusion that he is paying attention.	"uhhuh", "okay", "right", "oh", "oh yeah", "sure"...
FLOOR MECHANISMS		Floor Holder	fh	A floor holder occurs mid-speech by a speaker who has the floor. Spécifique aux échanges oraux.	but um - um... we can sort of...
BACKCHANNELS / ACKNOWLEDGEMENT		Acknowledge-answer	bk	Used to express a speaker's acknowledgment of a previous speaker's utterance or of a semantically significant portion of a previous speaker's utterance. Acknowledgments are neither positive nor negative, as they only serve to acknowledge, not to agree or disagree.	oh i see okay
ANSWERS	Positive	Accept	aa	The <aa> tag is used for utterances which exhibit agreement to or acceptance of a previous speaker's question, proposal, or statement. Utterances marked with the <aa> tag are quite short, as their lengthy counterparts are marked with the <na> tag.	"yes," "okay," "sure," "uhhuh," "right," "I agree," "exactly," "definitely," "that's true"...
SUPPORTIVE FUNCTIONS		Defending/Explanation	df	The <df> tag marks cases in which a speaker defends his own point or offers an explanation. Often, the word "because" signals an explanation.	"i was overreacting just because we've been talking about it"
SUPPORTIVE FUNCTIONS		Expansions of y/n Answers (= Elaboration)	e	This tag marks when a current speaker elaborates on a previous utterance of his by adding further details as opposed to simply continuing to speak on the same topic. When a speaker describes something using an example, the example is regarded as an elaboration.	
DISRUPTION FORMS		Interrupted/Abandoned/Uninterpretable	%		
FURTHER DESCRIPTION		Rising Tone	rt	The rising tone tag is used to mark utterances in which a speaker's tone rises at the end of his utterance.	
FLOOR MECHANISMS		Floor Grabber	fg	Floor grabbers usually mark instances in which a speaker has not been speaking and wants to gain the floor so that he may commence speaking. Spécifique aux échanges oraux.	"okay" "yeah but..."
ACTION MOTIVATORS		Offer (= suggestion)	cs	This tag marks proposals, offers, advice, and, most obviously, suggestions. Suggestions are often found in constructions such as "maybe we should..." Suggestions containing the word "maybe" are not to be confused with the maybe tag <am>. Additionally, if the phrase "excuse me" precedes something for which a speaker is negotiating permission, then it is marked as a suggestion rather than an apology <fa>.	
BACKCHANNELS / ACKNOWLEDGEMENT		Assessment/Appreciation	ba	These are acknowledgments directed at another speaker's utterances and function to express slightly more emotional involvement than what is seen in the utterances marked with the <ba> tag. Utterances marked with <ba> can be either positive or negative. When negative, utterances marked with the <ba> tag are often criticisms.	"that's good" "wonderful" "so this is slightly more complicated" "that's a whole lot of constructions"
CHECKS		Understanding Check	bu	This tag tag marks when a speaker checks to see if he understands what a previous speaker said or else to see if he understands some sort of information. With understanding checks, a speaker usually states what he is trying verify as correct and follows that with a tag question <g>.	"right?" "three thirty?"
FURTHER DESCRIPTION		Declarative-Question	d	The declarative question tag marks questions which have the syntactic appearance of a statement. In declarative questions, the subject precedes the verb and subject-auxiliary inversion and wh-movement do not occur. It is not uncommon for a rising tone <rt> to be found on a declarative question.	"nothing else?" "same idea?" "you don't know?"
ANSWERS	Positive	Affirmative Non-yes Answers	na	The <na> tag marks an utterances that act as narrative affirmative responses to questions, proposals, and statements. The difference between this tag and tag <aa> is that, as the <aa> tag is used for shorter utterances, the <na> tag is used for lengthy utterances.	
QUESTIONS		Wh-Question	qw	Questions that require a specific answer. What, which, where, when, who, why, or how...? However, not all questions containing a "wh" word are considered wh-questions. In determining whether an utterance is a declarative wh-question that does not contain a "wh" word, the surrounding context, in particular the response the question generates, is crucial to note. Most often, declarative wh-questions that do not contain "wh" words are requests for repetition.	"Why didn't you get the same results?" "What time do we have to leave?" "For who?"

ANSWERS	Negative	Reject	ar	The <ar> tag marks negative words such as "no" and other semantic equivalents that offer negative responses to questions, proposals, and statements.	"no," "nope," "no way," "nah," "not really," and "I don't think so"...
SUPPORTIVE FUNCTIONS		Collaborative Completion	2	This tag marks utterances in which a speaker attempts to complete a portion of another speaker's utterance.	
ANSWERS		Other Answers (= No knowledge)	no		"yeah i don't understand it" "i don't know" "didn't we already get that?"
FLOOR MECHANISMS		Hold Before Answer/Agreement	h	Used when a speaker who is given the floor and is expected to speak "holds off" prior to making an utterance. Spécifique aux échanges oraux.	let's see... well...
ACTION MOTIVATORS		Action-directive (= Command)	co	A command may arise in the form of a question (e.g., "Do you want to go ahead?") or as a statement (e.g., "Give me the microphone."). Commands are often confused with suggestions <cs>. If a rejection is considered impolite, the utterance is considered a command, otherwise it is considered a suggestion.	"just start up a new form" "let's get this clearer" "wait" "proceed"
QUESTIONS		Yes-No-question	qy	Essentially, an utterance is considered a yes-no question if it sounds as if it elicits a yes or no answer.	did i say that? right? the insertion number is quite high?
ANSWERS		Dispreferred Answers	nd	The <nd> tag marks statements which act explicit narrative forms of negative answers to previous speakers' questions, proposals, and statements. As with the <na> tag, the <nd> tag marks lengthier utterances than those marked with the <ar> tag which exhibit rejection. Often confused with <ng>. The <nd> tag marks utterances that offer explicit rejections and the <ng> tag marks utterances that offer implicit rejections through the use of hedging.	"it was more than that" "I'd prefer not to" "that's a different thing" "but there's no reason to do that"
FURTHER DESCRIPTION		Humorous Material	j	The <j> tag marks utterances of humorous or sarcastic nature.	
POLITENESS MECHANISMS		Downplayer	bd	The downplayer tag <bd> marks cases in which a speaker downplays or deemphasizes another utterance (made by the same speaker or a different speaker).	Short examples: "it's okay", "that's all right", "I'm kidding", "it's just a thought", "never mind". Other examples: "i don't know if this is at all useful", "Icould be wrong"
ACTION MOTIVATORS		Commitment	cc	This tag <cc> is used to mark utterances in which a speaker explicitly commits himself to some future cours With commitments, a speaker mentions what he will do in the future, not what he might do.	"i'll work on that" "i'll try to get to that" "i'll wait"
ANSWERS		Negative Non-no Answers	ng	As opposed to a dispreferred answer <nd> which explicitly offers a negative response to a previous speaker's question, proposal, or statement, a negative answer <ng> implicitly offers a negative response with the use of hedging.	"well you know, i do think eating while you're doing a meeting is going to be increasing the noise" "it just seems like that's a very different thing than what we're doing"
ANSWERS	Uncertain	Maybe	am	The <am> tag is one which the speaker asserts that his utterance is probable or possible, yet not definite. The <am> tag is often confused with suggestions <cs> which have the form of "maybe we should..."	"so i guess it's sort of averaging over all those three possibilities" "probably western, yeah" "maybe that's an interface issue that might be addressable"
QUESTIONS		Or-Clause (after Y/N question)	qrr	This tag marks when a speaker adds an "or" clause to a yes/no question.	
FURTHER DESCRIPTION		Exclamation	fe	The <fe> tag marks utterances in which a speaker expresses excitement, surprise, or enthusiasm. Utterances marked with the <fe> tag, excluding quotes, are punctuated with an exclamation mark < ! > within the transcript.	"wow!", "oops!"...
RESTATED INFORMATION	Repetition	Mimic Other	m	The mimic tag marks when a speaker repeats another speaker's utterance, or portion of another speaker's utterance. It does not have to be repeated verbatim in order to be considered a mimic. If a speaker's utterance is marked as a mimic, it may contain more speech in addition to what is mimicked.	
POLITENESS MECHANISMS		Apology	fa	An utterance is marked as an apology <fa> when a speaker apologizes for something he did.	"I'm sorry" "sorry to interrupt"
FURTHER DESCRIPTION		About-task	t	The about-task tag marks utterances that are in reference to meeting agendas or else address the direction of meeting conversations with regard to meeting agendas. The about-task tag is not to be confused with the topic change tag <tc>.	"i do have an agenda suggestion" "let's discuss agenda items"
CHECKS		Signal-non-understanding (= Repetition Request)	br	An utterance marked as a repetition request indicates that a speaker wishes for another speaker to repeat all or part of his previous utterance.	"what?", "sorry?", "huh?", "pardon?", "excuse me?", "say that again"...

ANSWERS	Positive	Accept-part	aap	The <aap> tag marks when a speaker explicitly accepts part of a previous speaker's utterance. Partial accepts are often conditional responses that accept or agree to another speaker's utterance. Partial accepts are often confused with partial rejections <arp>.	
QUESTIONS		Rhetorical-Question	qh	Questions to which no answer is really expected	"why not?" "I mean, who cares?"
FURTHER DESCRIPTION		Topic Change	tc	The <tc> tag marks utterances which either begin or end a topic.	"okay enough on forms" "i think we're sort of done" "what else we got?"
RESTATED INFORMATION	Repetition	Repeat	r	The repeat tag <r> is used when a speaker repeats himself. This often occurs in response to repetition requests or else to place emphasis on a certain point.	
FURTHER DESCRIPTION		Self-talk	t1	The <t1> tag is used when a speaker talks to himself.	
FURTHER DESCRIPTION		3rd-party-talk	t3	The third party talk tag marks utterances of side conversations. Side conversations are conversations which are not directed toward the main conversation and may only consist of a handful of utterances or may be quite lengthy.	
BACKCHANNELS / ACKNOWLEDGEMENT		Rhetorical-question Continue (= Rhetorical Question Backchannel)	bh	Rhetorical question backchannels lack semantic content and are syntactically similar to rhetorical questions, however they function as backchannels and acknowledgments	"yeah?" "really?"
ANSWERS	Negative	Reject-part	arp	The <arp> tag marks when a speaker explicitly rejects part of a previous speaker's utterance. Partial rejections are often responses posing exceptions when rejecting another speaker's utterance. Partial rejections are often confused with partial accepts <aap>.	"not when we were doing this." "not that much though." "except for it doesn't do well on short things remember."
RESTATED INFORMATION	Correction	Misspeak Self-Correction	bsc	The <bsc> tag marks when a speaker corrects his own error, with regard to either pronunciation or word choice.	
RESTATED INFORMATION	Repetition	Reformulate/Summarize	bs	The <bs> tag marks when a speaker summarizes a previous utterance or discussion, regardless of whose speech he is summarizing. Summaries are not to be confused with understanding checks <bs>.	
CHECKS		Follow me	f	The <f> tag marks utterances made by a speaker who wants to verify that what he is saying is being understood.	"do you know what i'm saying?" "right?"
QUESTIONS		Or-Question	qr	"Or" questions offer the listener at least two answers or options from which to choose.	"are you assuming that or not?" "do we just need to do that?" "is this the same as the e mail or different?" "is this uh just raw counts or is it?" "what if there was a door slam or something?"
POLITENESS MECHANISMS		Thanking	ft	The <ft> tag marks utterances in which a speaker thanks another speaker.	
FURTHER DESCRIPTION		Tag-Question	g	A tag question follows a statement and is a short question seeking confirmation of that statement.	
QUESTIONS		Open-Question	qo	An open-ended question places few syntactic or semantic constraints on the form of the answer it elicits. En gros, tout ce que ne correspond pas aux autres types de questions retenus.	what do you think about that? what about your trip yesterday?
RESTATED INFORMATION	Correction	Correct-misspeaking	bc	The <bc> tag is used when a speaker corrects another speaker's utterance.	
POLITENESS MECHANISMS		Sympathy	by	The <by> tag marks utterances in which a speaker exhibits sympathy.	"oh, i'm sorry" "it's just totally understandable"
POLITENESS MECHANISMS		Welcome	fw	The <fw> tag marks utterances which function as responses to utterances marked with the thanks tag <ft>. Phrases such as "you're welcome" and "my pleasure" are marked with the welcome tag <fw>.	

Bibliography

Ahmed Aboutaleb, Ahmed Fayed, Dina Ismail, Nada A. GabAllah, Ahmed Rafea, and Nourhan Sakr. 2021. Bert bilstm-attention similarity model. In 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pages 366–371.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames : a corpus for adding memory to goal-oriented dialogue systems. ArXiv, abs/1704.00057.

John Langshaw Austin. 1975. How to do things with words. Oxford university press.

Noa Avigdor-Elgrabli, Roei Gelbhart, Irena Grabovitch-Zuyev, and Ariel Raviv. 2018. More than threads: Identifying related email messages. CIKM '18, page 1711–1714, New York, NY, USA. Association for Computing Machinery.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein. 2020. Crawling and preprocessing mailing lists at scale for dialog analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1151–1158, Online. Association for Computational Linguistics.

Leonard Bloomfield. 1933. Language. Holt, New York.

Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. 2015. Efficient agglomerative hierarchical clustering. Expert Syst. Appl., 42(5) :2785–2797.

Kristy Elizabeth Boyer, Robert Phillips, Eun Young Ha, Michael Wallis, Mladen Vouk, and James Lester. 2009. Modeling dialogue structure with adjacency pair analysis and hidden Markov models. In Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers, pages 49–52, Boulder, Colorado. Association for Computational Linguistics.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4573–4577, Marseille, France. European Language Resources Association.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Harry Bunt. 1989. Information dialogues as communicative action in relation to user modelling and information processing. In The Structure of Multimodal Dialogue, Vol. 1, pages 47–74.

Harry Bunt. 1995. Dialogue control functions and interaction design. In Dialogue in Instruction, pages 197–214.

Harry Bunt. 2006. Dimensions in dialogue act annotation. In LREC, pages 919–924.

Harry Bunt. 2007. Multifunctionality and multidimensional dialogue act annotation. Communication-Action-Meaning. Gothenburg, pages 237–259.

Harry Bunt. 2009. The dit++ taxonomy for functional dialogue markup. In AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts, pages 13–24.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Harry Bunt, Dirk K. J. Heylen, Catherine Pelachaud, Roberta Catizone, and David R. Traum. 2009. The dit++ taxanomy for functional dialogue markup.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 549–558, Marseille, France. European Language Resources Association.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Fang. 2016. The DialogBank. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3151–3158, Portorož, Slovenia. European Language Resources Association (ELRA).

Vitor R. Carvalho and William W. Cohen. 2004. Learning to extract signature and reply lines from email. In IN PROCEEDINGS OF THE CONFERENCE ON EMAIL AND ANTI-SPAM.

Vitor R. Carvalho and William W. Cohen. 2005. On the collective classification of email "speech acts". In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, page 345–352, New York, NY, USA. Association for Computing Machinery.

Ann Cavoukian. 2010. Privacy by design: The definitive workshop. a foreword by ann cavoukian, ph.d. Identity in the Information Society, 3(2) :247–251.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. CoRR, abs/1807.05013.

Thierry Chanier, Céline Poudat, Benoît Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. 2014. The CoMeRe

corpus for French: structuring and annotating heterogeneous CMC genres. Journal for language technology and computational linguistics, 29(2) :1–30. Final version to Special Issue of JLCL (Journal of Language Technology and Computational Linguistics (JLCL, http ://jlcl.org/) : BUILDING AND ANNOTATING CORPORA OF COMPUTER-MEDIATED DISCOURSE : Issues and Challenges at the Interface of Corpus and Computational Linguistics (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel).

Wei Chen, Yeyun Gong, Can Xu, Huang Hu, Bolun Yao, Zhongyu Wei, Zhihao Fan, Xiaowu Hu, Bartuer Zhou, Biao Cheng, et al. 2021. Contextual fine-to-coarse distillation for coarse-grained response selection in open-domain conversations. arXiv preprint arXiv :2109.13087.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In The 41st International ACM SIGIR Conference on Research Development in Information Retrieval, SIGIR '18, page 225–234, New York, NY, USA. Association for Computing Machinery.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 76–86, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Andrew D Cohen. 1996a. Speech acts. Sociolinguistics and language teaching, 383 :420.

William W. Cohen. 1996b. Learning rules that classify e-mail. In AAAI spring symposium on machine learning in information access.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 309–316, Barcelona, Spain. Association for Computational Linguistics.

Marco Colombetti et al. 2000. A commitment-based approach to agent speech acts and conversations. In Proceedings of the Workshop on Agent Languages and Conversational Policies, pages 21–29.

V.J. Cook. 2014. The English Writing System. The English Language Series. Taylor & Francis.

Mark G Core and James Allen. 1997a. Coding dialogs with the damsl annotation scheme. In AAAI fall symposium on communicative action in humans and machines, volume 56, pages 28–35. Boston, MA.

Mark G. Core and James F. Allen. 1997b. Coding dialogs with the damsl annotation scheme. In Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines, pages 28–35, Cambridge, MA.

Ferdinand de Saussure. [1916] 1983. Course in General Linguistics. Duckworth, London. (trans. Roy Harris).

Mostafa Dehghani, Masoud Asadpour, and Azadeh Shakery. 2012. An Evolutionary-Based Method for Reconstructing Conversation Threads in Email Corpora. ASONAM '12, page 1132–1137, USA. IEEE Computer Society.

Mostafa Dehghani, Azadeh Shakery, Masoud Asadpour, and Arash Koushkestani. 2013. A learning approach for email conversation thread reconstruction. Journal of Information Science, 39 :846 – 863.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Giacomo Domeniconi, Konstantinos Semertzidis, Vanessa Lopez, Elizabeth M. Daly, Spyros Kotoulas, and Gianluca Moro. 2016. 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. SCITEPRESS - Science and Technology Publications, Lda.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. Computational Linguistics, 36(3) :389–409.

Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, pages 1179–1189, Portland, Oregon, USA. Association for Computational Linguistics.

Kristen Engel, Yiqing Hua, Taixiang Zeng, and Mor Naaman. 2022. Characterizing reddit participation of users who engage in the qanon conspiracy theories. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1) :1–22.

Shai Erera and David Carmel. 2008. Conversation detection in email systems. In Advances in Information Retrieval, pages 498–505, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. 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.

Dominique Estival, Tanja Gaustad, Ben Hutchinson, Son Bao Pham, and Will Radford. 2007. Author profiling for english emails. In In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 263–272.

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. Doc2dial : A goal-oriented document-grounded dialogue dataset. ArXiv, abs/2011.06623.

Michael Finke, Maria Lapata, Alon Lavie, Lori S. Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alexander H. Waibel, Klaus Zechner, and finkem. 2002. Clarity : Inferring discourse structure from speech. In Applying Machine Learning to Discourse Processing.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

J. Goldstein and R.E. Sabin. 2006. Using speech acts to categorize email and identify email genres. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), volume 3, pages 50b–50b.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 137–147, Seattle, USA. Association for Computational Linguistics.

M. A. K. Halliday and Deakin University. 1985. Spoken and written language / M.A.K. Halliday. Deakin University : distributed by Deakin University Press Waurn Ponds, Vic.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker turn modeling for dialogue act classification. In Findings of the Association for Computational Linguistics : EMNLP 2021, pages 2150–2157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter A. Heeman and James F. Allen. 1995. The trains 93 dialogues. Technical report, USA.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. arXiv preprint arXiv :1906.01543.

Lydia Mai Ho-Dac. 2021. Wikidisc. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

JR HOBBS. 1985. On the coherence and structure of discourse. Technical Report, 37.

G. Hripcsak and A. S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc, 12(3) :296–298.

Emily Jamison and Iryna Gurevych. 2013. Headerless, quoteless, but not hopeless? using pairwise email classification to disentangle email threads. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pages 327–335, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1250–1259, Singapore. Association for Computational Linguistics.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. 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. Association for Computational Linguistics.

Thorsten Joachims. 2001. A statistical learning learning model of text classification for support vector machines. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, page 128–136, New York, NY, USA. Association for Computing Machinery.

Sachindra Joshi, Danish Contractor, Kenney Ng, Prasad Deshpande, and Thomas Hampp. 2011. Auto-grouping emails for faster e-discovery. Proceedings of the VLDB Endowment, 4 :1284 – 1294.

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond T. Ng. 2010. Exploiting conversation structure in unsupervised topic segmentation for emails. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 388–398, Cambridge, MA. Association for Computational Linguistics.

Shafiq Joty and Tasnim Mohiuddin. 2018a. Modeling speech acts in asynchronous conversations: A neural-CRF approach. Computational Linguistics, 44(4) :859–894.

Shafiq Joty and Tasnim Mohiuddin. 2018b. Speech Act Modeling of Written Asynchronous Conversations: A Neural CRF Approach. Computational Linguistics (Special Issue on Language in Social Media, Exploiting discourse and other contextual information), pages 859–894.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive Summarization of Reddit Posts with Multi-level Memory Networks. In NAACL-HLT.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Bryan Klimt and Yiming Yang. 2004a. The enron corpus : A new dataset for email classification research. In Machine Learning : ECML 2004, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bryan Klimt and Yiming Yang. 2004b. The enron corpus : A new dataset for email classification research. In Machine Learning : ECML 2004, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

Manfred Krifka. 2022. Adjacency pairs in common ground update : Assertions, questions, greetings, offers, commands. In Proceedings of the 26th Workshop on the Semanics and Pragmatics of Dialogue, pages 94–105.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2017. Dialogue act sequence labeling using hierarchical encoder with CRF. CoRR, abs/1709.04250.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Enzo Laurenti, Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, and Camille Courgeon. 2022. Give me your intentions, I'll predict our actions: A two-level classification of speech acts for crisis management in social media. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4333–4343, Marseille, France. European Language Resources Association.

Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues. In Proceedings of the corpus linguistics 2003 conference, volume 16, pages 441–446. Lancaster : Lancaster University.

David D. Lewis. 1992. Representation and learning in information retrieval. Ph.D. thesis, University of Massachusetts. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Dernière mise à jour - 2023-02-24.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019a. A dual-attention hierarchical recurrent neural network for dialogue act classification. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 383–392, Hong Kong, China. Association for Computational Linguistics.

Wenting Li, Shangbing Gao, Hong Zhou, Zihe Huang, Kewen Zhang, and Wei Li. 2019b. The automatic text classification method based on bert and feature union. In 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), pages 774–777.

Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Hai-Tao Zheng, and Shuming Shi. 2020. The world is not binary : Learning to rank with grayscale data for dialogue response selection. arXiv preprint arXiv :2004.02421.

Di Liu, Zhen Zhao, and Li-Dong Gan. 2019. Intention detection based on bert-bilstm in tasko-riented dialogue system. In 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, pages 187–191.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. CoRR, abs/1711.05101.

WILLIAM C. MANN and SANDRA A. THOMPSON. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text - Interdisciplinary Journal for the Study of Discourse, 8(3) :243–281.

M. McCarthy. 1991. Discourse Analysis for Language Teachers. Cambridge Language Teaching Library. Cambridge University Press.

M. L. McHugh. 2012. Interrater reliability : the kappa statistic. Biochem Med (Zagreb), 22(3) :276–282.

Tahir Mehmood, Alfonso Gerevini, Alberto Lavelli, and Ivan Serina. 2019. Leveraging multi-task learning for biomedical named entity recognition. In AI*IA 2019 – Advances in Artificial Intelligence, pages 431–444, Cham. Springer International Publishing.

Stefano Mezza, Wayne Wobcke, and Alan D. Blair. 2022. A multi-dimensional, cross-domain and hierarchy-aware neural architecture for iso-standard dialogue act tagging. In International Conference on Computational Linguistics.

T. Daniel Midgley, Shelly Harrison, and Cara MacNish. 2006. Empirical verification of adjacency pairs using dialogue segmentation. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, pages 104–108, Sydney, Australia. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Einat Minkov, Richard C. Wang, and William W. Cohen. 2005. Extracting personal names from email: Applying named entity recognition to informal text. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 443–450, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Corrado Monti, Jacopo D'Ignazi, Michele Starnini, and Gianmarco De Francisci Morales. 2023. Evidence of demographic rather than ideological segregation in news discussion on reddit. ArXiv, abs/2302.07598.

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. 2020. Establishing a new state-of-the-art for French named entity recognition. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4631–4638, Marseille, France. European Language Resources Association.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. 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.

Pinelopi Papalampidi, Elias Iosif, and Alexandros Potamianos. 2017. Dialogue act semantic representation and classification using recurrent neural networks.

Kateřina Pardubová. 2006. Jazyk emailů: Formy psaného a mluveného jazyka [online]. SUPERVISOR : doc. Mgr. Jan Chovanec, Ph.D.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 86–95, Portland, Oregon. Association for Computational Linguistics.

Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access, 9 :112478–112489.

Massimo Poesio and David R. Traum. 1997. Conversational actions and discourse situations. Computational Intelligence, 13(3) :309–347.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. 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), pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Sujith Ravi and Zornitsa Kozareva. 2018. Self-governing neural networks for on-device short text classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 887–893, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. CoRR, abs/1908.10084.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2020. A survey of deep active learning.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr., 3(4) :333–389.

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

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality, pages 58–65, Vancouver, BC. Association for Computational Linguistics.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Robert Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3) :339–374.

Lionel Tadonfouet Tadjou, Fabrice Bourge, Tiphaine Marie, Laurent Romary, and Éric de la Clergerie. 2021. Building a corporate corpus for threads constitution. In Proceedings of the Student Research Workshop Associated with RANLP 2021, pages 193–202, Online. INCOMA Ltd.

Motoki Taniguchi, Yoshihiro Ueda, Tomoki Taniguchi, and Tomoko Ohkuma. 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, pages 4969–4980, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David R. Traum and Elizabeth A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. Computational Intelligence, 8.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In aaai email-2008 workshop.

Daniel Vanderveken. 1992. Meaning and speech acts. volume i : Principles of language use. volume ii : Formal semantics of success and satisfaction. Tijdschrift Voor Filosofie, 54(2) :340–340.

A. Viera and J. Garrett. 2005a. Understanding interobserver agreement : the kappa statistic. Family medicine, 37 5 :360–3.

Anthony J Viera and Joanne M Garrett. 2005b. Understanding interobserver agreement: the kappa statistic. Family medicine, 37(5) :360—363.

Wei Wang, Saghar Hosseini, Ahmed Hassan Awadallah, Paul N. Bennett, and Chris Quirk. 2019. 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. Association for Computing Machinery.

Xia Wang, Ming Xu, Ning Zheng, and Mo Chen. 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.

Yejun Wu and Douglas W. Oard. 2005. Indexing emails and email threads for retrieval. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 167–177, Singapore and Online. Association for Computational Linguistics.

Hongsheng Xu, Ganglong Fan, Guofang Kuang, and Chuqiao Wang. 2023. Exploring the potential of bert-bilstm-crf and the attention mechanism in building a tourism knowledge graph. Electronics, 12(4).

Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, and Zhoujun Li. 2018. Response selection from unstructured documents for human-computer conversation systems. Knowledge-Based Systems, 142 :149–159.

Jen-Yuan Yeh. 2006. Email thread reassembly using similarity matching. In International Conference on Email and Anti-Spam.

Dian Yu and Zhou Yu. 2019. Midas : A dialog act annotation scheme for open domain human machine spoken conversations. arXiv preprint arXiv :1908.10023.

Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. 2016. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC Medical Research Methodology, 16(1).

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In 11th AAAI International Conference on Web and Social Media (ICWSM).

Shanshan Zhang, Lihong He, Slobodan Vucetic, and Eduard Dragut. 2018. Regular expression guided entity mention mining from noisy web data. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1991–2000, Brussels, Belgium. Association for Computational Linguistics.

Table des figures

Liste des tableaux