Analyse d'opinion dans les interactions orales

Valentin Barrière

université PARIS-SACLAY

TELECOM ParisTech

NNT : 2019SACL016

Thèse de doctorat

Analyse d'opinion dans les interactions orales

Thèse de doctorat de l'Université Paris-Saclay
préparée à Télécom ParisTech

Ecole doctorale n°580 Sciences et Techniques de l'Information et des Communications (STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Paris, le 15 Avril 2019, par

VALENTIN BARRIERE

Composition du Jury :

Pierre ZWEIGENBAUM DR, Université Paris Sud (LIMSI)	Président
Mohamed CHETOUANI PR, Sorbonne Université (ISIR)	Rapporteur
Xavier TANNIER PR, Sorbonne Université (LIMICS)	Rapporteur
Fabien RINGEVAL MCF, Université Grenoble Alpes (LIG)	Examinateur
Julien VELCIN PR, Université Lyon 2 (ERIC)	Examinateur
Chloé CLAVEL MCF, Télécom ParisTech (LTCI)	Co-directeur de thèse
Slim ESSID PR, Télécom ParisTech (LTCI)	Directeur de thèse

Table des matières

Table des figures

TABLE DES FIGURES

Liste des tableaux

LISTE DES TABLEAUX

Glossaire

AA : Apprentissage Automaque
ANEW : Affective Norms for English Words
ANN : Réseau de neurones artificiel
BD : Base de Données
BoNG : Bag-of-N-Grams
EDA : Ensemble de Descripteurs Acoustiques
EDT : Ensemble de Descripteurs Textuels
GCI : Instant de fermeture glottal
GRU : Gated Recurrent Unit
HCRF : Hidden Conditional Random Fields **PA** : Paire Adjacente
RNN : Recurrent Neural Network
SST : Standford Sentiment Treebank
TFN : Tensor Fusion Network

1 Introduction générale

Résumé du chapitre

— La communication humaine est définie comme la transmission d'un signal multimodal soumis à l'interprétation (modélisé par les filtres de l'émetteur et du récepteur) des individus. L'interaction est une communication avec rétroaction de la part d'une des entités.
— Le signal multimodal transporte des informations contenues dans la voix, le texte, les expressions faciales, les gestes et les postures,...
— Les opinions sont des phénomènes qui n'impliquent pas nécessairement une reaction physiologique, contrairement aux émotions.
— Les études sur les émotions vont favoriser le signal vocal alors que les études sur l'opinion vont favoriser l'analyse du contenu textuel. La prise en compte du contexte interactionnel et l'intégration des caractéristiques propres à l'oral est cruciale.

L'analyse automatique des sentiments et opinions est un champ d'étude relativement récent qui connait une ascension fulgurante ces dernières années et il devient difficile de s'en passer dans de nombreux domaines. Les entreprises utilisent les informations nouvellement disponibles permettant de mieux connaître chaque client pour répondre plus efficacement à ses demandes et anticiper ses besoins. Les applications sont multiples : automatisation dans les centres d'appel via la détection d'une situation conflictuelle, amélioration de l'interaction avec un Agent Virtuel en lui apportant des informations émotionnelles sur l'utilisateur ou simplement fouille de données dans les réseaux sociaux afin de mieux connaître l'opinion générale de la population sur un sujet.

La finalité de ce travail de thèse est l'étude et la création d'un système permettant d'analyser les opinions dans des interactions dyadiques, ce qui correspond à une discussion entre deux entités. Une partie de ce travail de thèse a été consacrée à la création d'un modèle comprenant l'extraction et la modélisation des données textuelles et acoustiques et l'utilisation de ces données par un classifieur approprié.

Bien que discipline nouvelle et pleine de défis à relever, l'analyse de sentiments utilisant uniquement le texte est déjà bien ancrée dans l'état de l'art et nous souhaitons utiliser l'audio pour permettre de repérer des structures distribuées sur ces deux modalités. L'utilisation des différentes modalités permettrait d'obtenir des informations complémentaires et ainsi équilibrer des prédictions qui peuvent être faussées lorsque l'on a accès à moins d'informations (par exemple intonation de la voix en désaccord avec la valence du texte écrit).

Dans cette introduction, nous commencerons en section 1.1 par une présentation de ce qui nous permettra de définir le contexte des interactions orales. La sous-section 1.1.1 définit sur des bases semblables aux théories de Shannon et du Signal Social les principes de la communication humaine multimodale. On s'intéressera en particulier à la communication humaine orale et ses spécificités. À partir de là, nous définirons une interaction humaine orale comme une communication orale avec des rétroactions des deux côtés (sous-section 1.1.2)

Dans la section 1.2, nous aborderons la deuxième partie du sujet de thèse en définissant les opinions dans les interactions orales. Nous nous intéresserons tout d'abord à l'ensemble des phénomènes affectifs (sous-section 1.2.1) pour définir ce qu'est une opinion, puis ce qu'est la réalisation du phénomène d'opinion dans une interaction de parole (sous-section 1.2.2). Nous finirons par une mise en contexte de la détection automatique des opinions dans les interactions orales (sous-section 1.2.3)

Pour conclure cette introduction, nous présenterons notre méthodologie, nos questions de recherche et le plan de l'organisation de cette thèse ainsi que nos contributions.

1.1 COMMUNICATION HUMAINE ET INTERACTION ORALE

Nous allons commencer par situer le contexte dans lequel se situe le travail de cette thèse. L'analyse des opinions dans les interactions orales dyadiques implique plusieurs points :

— l'existence d'au moins deux entités, qu'elles soient humaines ou robotiques (Mordatch et Abbeel, 2017; Lewis et collab., 2017) ;
— l'existence d'un moyen de communication de la part des deux entités permettant un transfert d'information ;
— une rétroaction de la part d'une de ces entités afin que la communication ne soit pas unilatérale et qu'il y ait interaction (Abric, 2008).

En premier lieu (sous-section 1.1.1), nous définirons la communication humaine. La communication sera définie dans un contexte proche de celui de la théorie de Shannon (1948) et de la théorie du signal social de Vinciarelli et collab. (2009) (bien que cette dernière soit restreinte aux signaux non-verbaux). En effet la communication humaine, c'est-à-dire quand un humain ou un agent avec des capacités sociales est émetteur, fait appel à toutes sortes de signaux sociaux. Dans cette même partie, nous caractériserons les spécificités de la communication orale.

Une fois que nous aurons défini le concept de communication humaine, nous définirons ce qu'est une interaction (sous-section 1.1.2). Toujours sur les bases de la théorie du signal social, nous définirons une interaction comme un ensemble de conditions permettant une communication entre deux entités avec une rétroaction d'au moins un des interacteurs. On s'attardera en particulier sur les interactions orales humaines qui sont un des objets d'étude de cette thèse.

1.1.1 Communication humaine orale

D'après la définition du Larousse, la communication est l'*action de transmettre quelque chose*. La communication implique la présence d'un émetteur (actif) et d'un potentiel récepteur (qui peut être passif). Afin de poser toutes les bases, nous définissons dans ces travaux la communication humaine comme un acte de communication ayant pour émetteur un humain ou un agent socialement compétent. Après avoir introduit ce qu'est la communication humaine via une approche basée sur les théories de l'information de Shannon (1948) et celle du signal social Pentland (2007), nous nous restreindrons aux spécificités de la communication orale.

1.1.1.1 Signal social et transport d'information

L'intelligence émotionnelle est une capacité qui s'est développée chez les mammifères au fil de l'évolution comme un atout pour survivre signalant les changements

FIGURE 1.1: Schématisation de la propagation d'une onde magnétique entre une antenne émettrice et une antenne réceptrice.

(réels ou imaginaires) dans les relations entre un individu et son environnement afin de fournir une réponse adéquate (Mayer et collab., 2008). De la même manière, les humains font preuve d'une intelligence sociale lors de leurs interactions avec le reste du monde (Albrecht, 2006).

L'Homme communique avec ses pairs grâce à différents moyens : il peut utiliser des gestes, des postures, des intonations, des expressions faciales, des mots, etc... Lors d'un acte communicatif, l'être humain utilise ces moyens pour transmettre de l'information. Ces signaux ont pour but d'avoir une communication de meilleure qualité. Ils se manifestent à haut niveau sous la forme de phénomènes comme l'empathie ou la conscience sociale (Goleman, 2006). On peut les détecter à bas niveau à l'aide de descripteurs adaptés et d'une structure de modèle assez complexe, et c'est ce que l'on fera dans cette thèse.

En se basant sur la théorie de Shannon (1948), on peut modéliser un acte communicatif comme l'émission, par une source qui est le locuteur, d'un signal ou d'un ensemble de signaux (Ringeval, 2011). Le récepteur est à la fin du canal de propagation du signal et peut décoder ce dernier suite à sa réception (voir Figure 1.1). Ainsi, un signal est généré puis envoyé par l'émetteur afin de transmettre une information au récepteur, par exemple, un état d'esprit ou une attitude sociale.

Les filtres des émetteur et récepteur sont construits par rapport à la manière de formuler ou d'interpréter un signal d'information. Ces filtres sont liés aux modalités utilisées, au contexte social et à l'individu : ils sont sens, perception et interprétation. Pour un décodage efficace de l'information contenue dans le signal de l'émetteur, il faut que les filtres de l'émetteur et ceux du ou des récepteurs soient partagés entre les différentes entités de la communication (Vinciarelli et collab., 2008). Un mauvais alignement des filtres peut induire une mauvaise interprétation de l'intention de l'émetteur, par exemple ce qui arrive lors d'un *culture clash*.

Le canal de propagation correspond au média utilisé et peut faire subir des pertes au signal de différentes manières. Ce peut être une vidéo-conférence, une ligne téléphonique, ou bien un lieu sombre ou venteux. Par exemple, les postures de l'animateur ne sont pas reçues par la personne écoutant la chronique sur son poste radio. Il est intéressant de savoir que certaines informations qui paraissent dépendre d'une modalité sont en fait partagés entre celles-ci. Par exemple, même si le sourire est principalement visible via la vidéo, il est possible de le retrouver à l'oreille (Arias et collab., 2018).

Le signal est le concept qui est transmis de l'émetteur au récepteur. Le passage par les filtres des émetteur et récepteur et les pertes dues au média le déforment. Il existe une différence entre signaux communicatifs, qui sont envoyés par l'émetteur dans le but de communiquer avec le récepteur, et signaux informatifs qui permettent au récepteur de reconstruire une intention communicative de l'émetteur qu'il n'a pas souhaité coder dans son signal (Vinciarelli et collab., 2011). La prise en compte de cette différence n'est pas utile dans notre cas, car on souhaite créer un système à la fois général et indépendant de l'individu.

Un signal étant une "*variation d'une grandeur physique de nature quelconque, transportant de l'information*" (Larousse, 2018), on peut séparer les différentes grandeurs physiques qui permettent de transporter l'information par modalités. Lors d'un acte communicatif, l'expression d'un phénomène se faisant de manière asynchrone par rapport au contenu textuel, à l'acoustique, aux gestes, aux expressions faciales, nous irons plus loin en séparant ces différentes modalités en fonction des caractéristiques de réalisation d'un phénomène social dans une interaction humaine. Ainsi, **on parlera dans la suite de cette thèse de signal audio, signal textuel, signal gestuel**, etc... pour parler du signal de l'information transmis par un conteneur particulier, que ce soit d'une manière volontaire ou non, lors d'un acte communicatif.

Ces signaux sont captés par le récepteur grâce à ses différents capteurs biologiques personnels (sens) : plusieurs signaux multimodaux sont captés puis traités par le cerveau (voir Figure 1.2).

L'interprétation de ces signaux captés dépend des filtres introduits plus haut, du contexte et de la culture. La reconstruction du signal implique un apriori sur l'autre interacteur et sa culture afin d'approximer ses filtres. Les concepts haut-niveau peuvent être exprimés par une accumulation de descripteurs bas niveaux. Par exemple, on peut retrouver l'émotion avec les unités d'action liées aux expressions faciales, les intonations, le vocabulaire employé et l'agencement de la phrase.

Un exemple concret de communication est montré en Figure 1.3 où un émetteur parle face à une caméra, comme dans les Vlogs (blog vidéo) que l'on peut trouver sur

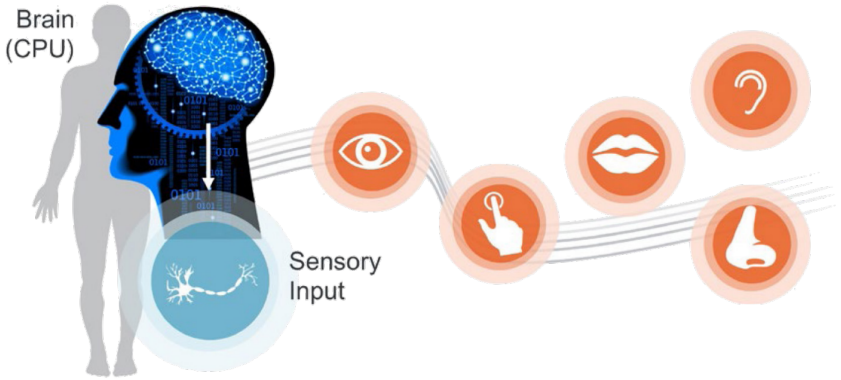


FIGURE 1.2: L'homme traite l'information provenant de capteurs multimodaux afin de prendre une décision (de Poria et collab. (2017a)).

le web. C'est un cas simple où il n'y a pas d'interaction entre l'émetteur et le récepteur. Le média utilisé est l'audiovisuel et les informations sont transmises dans du texte, l'acoustique de la voix, les expressions faciales et les gestes.

Ces signaux de communication humaine peuvent être utilisés avec des modèles multimodaux pour différentes tâches permettant de repérer les traits de personnalité ou comportements des locuteurs comme ses émotions (Sahay et collab., 2018), sa capacité de persuasion, son humour, sa passion (Liang et collab., 2018) ou même sa capacité à être professionnellement compétent pour un type de poste (Hemamou et collab., 2019).

Pour finir, dans la suite de cette thèse nous nous intéresserons à des communications et des interactions orales. Ainsi, nous nous restreindrons dans notre approche multimodale au **signal verbal** qui contient les informations provenant du contenu textuel et au **signal vocal** qui contient les informations acoustiques de la parole.

1.1.1.2 Spécificités du domaine oral dans la communication humaine

La communication humaine peut se faire de différentes manières selon le mode d'émission du signal. Dans cette thèse, les deux modalités que nous allons étudier sont le vocal et le verbal. C'est pour ceci que nous nous intéresserons en particulier aux différences entre une communication écrite et une communication orale.

La manière de s'exprimer à l'oral est très différente de celle de s'exprimer à l'écrit. Le locuteur formulera ses idées d'une manière plus naturelle que s'il les énonçait à l'écrit, ceci dû au **caractère spontané de la parole** qui fait que le locuteur n'a pas autant le temps de penser à l'oral qu'à l'écrit. Le domaine oral est, du fait de son caractère spontané, beaucoup plus difficile à traiter que le domaine écrit.

Les phrases ayant un début mais pas de fin, et les mots à moitié énoncés sont cou-

FIGURE 1.3: Exemple d'une communication humaine : émission de signaux de la part d'un humain.

rants dans les discours oraux spontanés. Le locuteur peut débuter sur une idée, puis changer d'avis sur la formulation de cette idée (Dubuisson Duplessis, 2014). Cette particularité de l'oral en fait un type de donnée difficile à utiliser par rapport à une expression textuelle où le discours suit une forme plus classique (Palau et Moens, 2009), que ce soit au niveau local par rapport à la syntaxe des phrases ou bien au niveau global par rapport à l'agencement des paragraphes d'un discours.

À l'oral on observe beaucoup de pauses du locuteur, qui peuvent être remplies ou non. Ces pauses sont des auto-interruptions du locuteur qui permettent de structurer le discours. L'utilisation de pauses est un indice important de la structure discursive du locuteur (Shriberg et collab., 2000). Une pause verbalisée permet de gérer le rythme du discours et la gestion des tours de parole. Elle est aussi un indice important de l'hésitation du locuteur, mais aussi du désir qu'il a de continuer son tour de parole s'il est en interaction, ou du moins sa phrase s'il est en communication sans interlocuteur (Ten Bosch et collab., 2005). Cependant, certaines pauses ont aussi simplement des rôles respiratoires et articulatoires et sont sans lien avec l'information verbale (Campione et Véronis, 2002).

De la même manière que les pauses peuvent aider à structurer le discours oral, la prosodie du locuteur permet d'apporter divers types d'informations pour contrôler le message envoyé lors d'un discours (Ringeval, 2011). Par exemple, elle contient des informations sur le contenu syntaxique des phrases (Warren, 1996). L'accent tonique sert

aussi à accentuer un mot par rapport à d'autres, et permet d'attirer l'attention du locuteur sur une partie de la phrase énoncée. Ainsi il est aussi lié au contenu linguistique et enrichit la quantité d'information contenue uniquement dans le texte sur la syntaxe de la phrase (Paul et collab., 2005).

Pour finir, le registre lexical et le vocabulaire changent entre l'écrit et l'oral. Si l'on souhaite utiliser des descripteurs textuels performants appris sur une grande quantité de texte, on aura généralement accès à du texte écrit et non à des transcriptions orales. Cependant, la sémantique des mots sera différente et une adaptation de domaine est nécessaire pour utiliser ces représentations. Un '*yeah*' est bien plus présent dans un discours oral que dans un discours écrit et son signifié est plus proche de celui du '*yes*' à l'écrit. Toutes ces transformations du langage obligent de considérer le discours écrit et le discours oral comme des domaines différents. Il est possible de passer d'un domaine à l'autre avec un modèle approprié, ce qu'on verra lors de certaines expériences menées dans cette thèse.

1.1.2 Interaction orale : communication orale avec rétroaction

Après avoir introduit le principe de communication humaine multimodale, nous pouvons commencer à parler d'interaction humaine multimodale. Ici contrairement au cas précédent, nous n'avons plus un émetteur et un récepteur mais chacune des entités peut jouer son propre rôle.

Une interaction n'est pas aussi simple qu'une communication comme on a pu le voir plus tôt. Le récepteur peut aussi envoyer des messages à l'émetteur lors du transfert d'information, influencer celui-ci en temps réel et lui faire changer le contenu initial de son message. Ceci est particulièrement vrai dans le cas de l'oral où l'énoncé du locuteur se construit de manière incrémentale.

En effet la théorie de Shannon suppose que le message est envoyé de manière complète, alors que lors d'une interaction des signaux sont échangés en temps continu et non pas de manière discrète. Le locuteur peut alors changer son message en temps réel pour s'adapter à la réponse ou au retour (*back-channel*) de son interactant (Dubuisson Duplessis, 2014). Ce processus de rétroaction permet de réguler les interactions (Abric, 2008).

Dans les interactions humaines et lors de l'analyse d'un locuteur, il est important de prendre en compte le contexte afin de pouvoir définir les phénomènes observés. Par exemple, Ghosh et collab. (2018) arrivent à détecter le sarcasme dans du texte de manière plus performante lorsqu'ils prennent en compte le contexte conversationnel.

Dans cette section, nous avons défini les concepts de communication humaine, de communication humaine orale et d'interaction qui est une communication avec rétro-

FIGURE 1.4: Exemple d'une interaction humaine : émission et réception de signaux de la part des deux interacteurs humains.

action. Dans la section suivante, on s'intéresse uniquement à un type de phénomène que l'on caractérise par des signaux sociaux, qui est l'opinion dans le contexte de la communication humaine à l'oral.

1.2 LES OPINIONS DANS LES INTERACTIONS ORALES

Les interactions orales et la communication humaine permettent de faire passer de nombreux messages à travers des signaux sociaux, dont les opinions font partie. Dans cette section nous pouvons passer à l'introduction et la définition de la deuxième grande partie de l'intitulé du sujet, à savoir les opinions, et leur manifestations dans les interactions orales. Les opinions sont des phénomènes subjectifs et leur définition ne fait pas consensus dans la littérature scientifique, chaque communauté ayant ses propres attributs pour définir les phénomènes affectifs. On définira les manifestations liées aux opinions dans la sous-section 1.2.1, avant de s'intéresser à la réalisation du phénomène d'opinion dans les interactions orales humaines (sous-section 1.2.2). Finalement, nous aborderons la détection automatique des opinions dans les interactions de paroles en sous-section 1.2.3, permettant de répondre à : pourquoi faire ce travail de recherche, qu'est-ce qui se fait déjà dans ce domaine, et comment se situer par rapport à l'état de l'art?

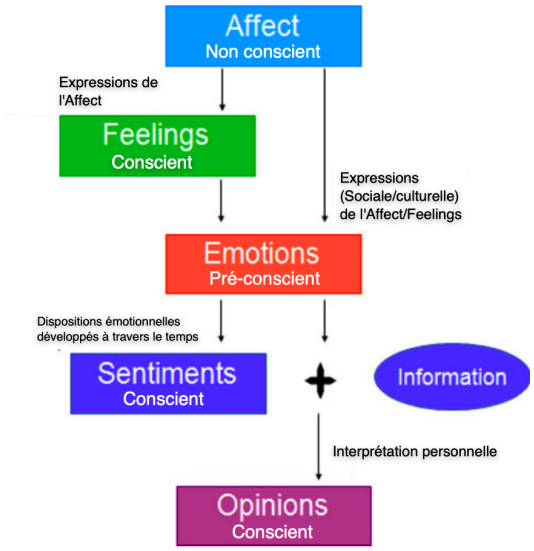


FIGURE 1.5: Représentation par Munezero et collab. (2014) des différents phénomènes affectifs.

1.2.1 Définitions générales des différents phénomènes liés aux opinions

Les termes affects, ressentis, [1] émotions, sentiments [2] et opinions sont utilisés de manière interchangeable dans la littérature et leur utilisation dépend des domaines de recherche (Clavel et Callejas, 2016). Comme notre travail se situe dans le cadre d'études sur la communication humaine à l'oral, nous nous baserons sur le schéma proposé par Munezero et collab. (2014) qui travaillent sur les phénomènes affectifs dans le texte, qui est un domaine proche. Ce schéma visible en figure 1.5 permet de distinguer les différents phénomènes affectifs humains traités dans la littérature.

Un affect se définit comme un état affectif élémentaire et il traduit une pulsion. Il est non-conscient et c'est lui qui est à l'origine de tous les phénomènes affectifs (Munezero et collab., 2014). D'après les mêmes auteurs, l'expression de cet affect s'exprimant de manière consciente et personnelle est le ressenti.

Contrairement aux émotions qui sont des expressions sociales pré-conscientes et influencées par la culture (Shouse, 2005), le ressenti est interne. Comme l'affect, il est moins ancré dans le contexte d'intelligence sociale qui nous intéresse pour l'analyse de la communication humaine. Nous nous concentrerons par la suite sur les émotions, sentiments et opinions qui découlent d'une conscience sociale, et délaisserons les no-

1. *feelings*
2. on utilisera le sens anglais de sentiment, comme il est utilisé dans la littérature, par souci de simplicité

tions d'affect et de ressenti qui sont peu liées à un échange d'information.

Les émotions sont des manifestations sociales qui dépendent fortement de notre culture et des interactions que l'on a avec le monde (Wierzbicka, 1999), mais elles sont aussi définies par rapport au temps. Scherer (2005) définit les émotions comme un phénomène sur le court terme, contrairement à l'humeur[3] qui est un état diffus pouvant durer de quelques heures à plusieurs jours. De la même manière, la temporalité du phénomène peut être utilisée pour séparer les sentiments et les émotions dans des contextes où le manque d'information rend cette catégorisation difficile, comme lors de l'utilisation de données textuelles uniquement (Munezero et collab., 2014). Contrairement à l'émotion qui est un phénomène sur le court-terme, le sentiment s'établit sur le long terme, et il est défini comme une superposition de dispositions émotionnelles s'étant développées à travers le temps et perdurent.

Les sentiments, à l'inverse des émotions qui sont des manifestations sociales externes, ne se manifestent pas forcément par des comportements ou des expressions. Et ceci alors que les sentiments sont motivés par les émotions.

Les opinions sont des phénomènes provenant d'une réflexion personnelle "logique" (subjective ou non) et motivés non entièrement par des émotions, mais par des probabilités provenant d'information (subjective ou non) sur un sujet particulier. **Ce sont des interprétations personnelles non nécessairement contraintes par des conventions sociales, à l'inverse des sentiments qui sont construits socialement.** Ainsi, les opinions sont des jugements sujets à la controverse mais ceux-ci peuvent ou peuvent ne pas être émotionnellement chargés (Munezero et collab., 2014). Dans le cadre de la communication humaine, on pourrait dire que le sentiment provient d'une verbalisation des émotions (qui peut être interne), alors que l'opinion proviendrait d'une réflexion cognitive logique pas forcement chargé émotionnellement.

Pour résumer, les opinions sont basées sur des émotions qui sont des phénomènes sociaux utilisés par l'homme pour communiquer avec ses pairs, avec un ajout d'information couplé à un procédé cognitif. Ce sont des phénomènes réalisés à partir d'une réflexion comprenant de la logique. Par exemple, l'utilisateur faisant des critiques en ligne et donnant des notes à différents aspects d'une entité est capable d'expliquer la valeur des notes données.

Il existe de nombreuses théories permettant de modéliser des phénomènes affectifs et nous citerons aussi celle de l'*Appraisal* de Martin et White (2003) qui introduit la notion (nous utiliserons les mots anglais) d'*attitude* qui a été utilisée dans cette thèse. L'*attitude* est une notion de la théorie de l'*Appraisal* comprenant les *affects*, les *judgements* et les *appreciations* dans cette même théorie de l'*Appraisal*, et qui correspond

3. *mood*

à des évaluations de la part du locuteur. Nous avons utilisé dans une partie de cette thèse des annotations en attitude, qui est une phénomène proche de l'opinion définie plus haut.

Dans la partie suivante, nous nous concentrerons sur les opinions, qui sont des phénomènes complexes logiques, au sein d'une interaction et non plus dans le cadre d'un discours ou d'une critique orale.

1.2.2 Expression d'une opinion dans la parole et dans une interaction orale

On a vu dans la sous-section 1.2.1 qu'il existe de nombreux phénomènes liés aux opinions, et que ces derniers sont très différents les uns des autres. Les opinions, contrairement aux émotions, ne se réalisent pas de la même manière et ne se transmettent par les mêmes moyens.

En effet, on a pu voir dans la sous-section précédente que les opinions sont des phénomènes calculés à partir d'information extérieure (Munezero : "*opinions are prominently based on probabilities of information about a topic*") et moins liés à la physiologie que les émotions. D'après les travaux de Mehrabian (1971), il ressort que le contenu émotionnel est à 55% dans le signal visuel (expressions du visage et du langage corporel), à 38% dans le signal vocal (intonation et son de la voix) et à 7% dans le signal verbal (par la signification des mots et l'agencement de la phrase). Ce qui n'est pas le cas pour les opinions où une plus grande partie du message est contenue dans le signal verbal (Munezero et collab., 2014).

Si les travaux sur l'analyse automatique des émotions et des affects dans la parole orale sont nombreux, l'analyse des sentiments et des opinions dans des énonciations est un domaine relativement nouveau (Soleymani et collab., 2017). Les études qui se penchent sur la partie acoustique de la parole ont du mal à faire des délimitations entre sentiment et émotion. Cependant, des études comme celle de Mairesse et collab. (2012) ont trouvé des liens entre le sentiment du locuteur et des descripteurs purement acoustiques reliés à la fréquence fondamentale.

Dans une interaction, le phénomène d'opinion peut être collaboratif et textuellement réparti entre les différents locuteurs. Par exemple lors d'une paire de tours de parole de type question-réponse, il est parfois nécessaire d'observer l'ensemble de la paire pour connaitre l'opinion d'un seul des locuteurs. Langlet (2018) définit le type d'interactions que nous allons utiliser dans cette thèse comme des interactions collaboratives où l'agent et l'utilisateur jouent des rôles complémentaires car seul l'agent, qui s'exprime en premier, est capable de poser des questions. Il est nécessaire de construire un système prenant en compte l'interaction afin de détecter les indices d'un phénomène se développant sur une longueur caractéristique plus grande qu'un

simple tour de parole.

À l'oral, et dans le corpus de discussions ouvertes que nous allons utiliser pour l'analyse des interactions, les locuteurs ont plus tendance à parler peu et vite et, ce faisant, les tours de parole s'enchainent rapidement. La possibilité de couper un interlocuteur au milieu de son tour de parole permet, par exemple, à l'autre interlocuteur de finir la phrase du premier interlocuteur (Cafaro et collab., 2016) ou de s'aligner par rapport à un phénomène affectif avec lui (Varni et collab., 2017).

Pour les différentes raisons évoquées plus haut, il est important de prendre en compte ce contexte spécial pour analyser des opinions dans des données interactionelles, et encore plus dans le cadre d'une communication orale ayant une structure plus dynamique et inconsistante que l'écrit.

1.2.3 Détection automatique des opinions dans les interactions orales : histoire, enjeux et applications

La détection de phénomènes affectifs dans la parole a été longtemps dominée par la reconnaissance d'émotions. Comme on l'a vu dans les parties précédentes, ce phénomène est plus simple à détecter à l'aide de la voix car lié à la condition physiologique du locuteur.

Dans les données textuelles, l'analyse des sentiments a été étudiée depuis plusieurs années à l'aide de méthodes d'apprentissage automatique (Pang et Lee, 2004). Puis, l'analyse fine des opinions a suivi (Breck et collab., 2007). Cependant, peu de travaux existent utilisant l'oral. En effet, nous avons pu voir plus haut que la parole possède une structure différente du texte, rendant la détection des phénomènes d'opinions et de sentiments plus floue et plus difficile.

Mais, depuis quelques années, on peut observer de plus en plus d'écrits sur ces sujets. Les enjeux sont multiples : communication avec un agent virtuel, analyse de conversations audio dans un centre d'appel ou bien analyse de critiques multimodales sont autant de domaines où l'analyse d'opinions est un axe d'amélioration.

L'intelligence émotionnelle et sociale est nécessaire pour avoir une bonne interaction avec la machine, car on a vu que c'est une compétence primordiale dans le processus de communication humaine. Ainsi, de nombreux travaux s'attardent sur ce sujet dans un but d'amélioration de la qualité de l'interaction, dans des configurations humain-humain ou humain-agent. L'analyse d'opinion est un pas de plus dans cette direction.

Langlet et Clavel (2016) évoquent la théorie de la balance de Heider (1958) comme base pour une meilleure interaction. Partager les mêmes opinions entre les 2 interactants permet de les rapprocher et il peut donc être important pour l'agent virtuel de reconnaitre les opinions de l'utilisateur afin de se rapprocher de lui.

Ben Youssef et collab. (2015) travaillent dans le contexte d'entretiens d'embauche avec un agent virtuel jouant le rôle du recruteur. Les auteurs utilisent l'audio pour adapter le comportement d'un agent virtuel par rapport à celui de son interlocuteur et obtiennent de meilleurs résultats vis à vis de l'expérience utilisateur et de la crédibilité perçue de l'agent.

L'agent virtuel SimSensei de Devault et collab. (2014) a pour but d'améliorer l'évaluation diagnostique des praticiens pour des patients ayant des troubles psychologiques. Un système détecte et utilise l'audio ainsi que d'autres signaux non verbaux ayant un impact sur la communication et les interactions humaines pour améliorer l'engagement de l'utilisateur dans une conversation humain-agent (Rizzo et collab., 2014). Le système peut aussi interpréter des signaux sociaux dans le but d'aider à construire un diagnostique clinique

Finalement, on trouve de plus en plus de données multimodales liées à l'analyse des opinions sur Internet que l'on peut interpréter. Si la tâche de fouiller les opinions finement dans les critiques textuelles en ligne est connu depuis de longues années, celle sur les critiques Vlogs est plus récente (Zadeh et collab., 2018d). Néanmoins, il est important de noter que ces analyses de critiques Vlogs, si elles sont de plus en plus nombreuses, n'utilisent pas le contexte interactionnel pour obtenir un système plus performant. Le nombre de travaux prenant en compte un contexte interactionnel de manière multimodale s'intéressent uniquement aux émotions, sont très récents et s'élèvent à notre connaissance au nombre de trois : Hazarika et collab. (2018a); Majumder et collab. (2018); Hazarika et collab. (2018b)

Positionnement

Les émotions étant la cible principale des études sur la voix et les expressions faciales, l'étude des sentiments ou des opinions reste peu abordée. Quand c'est le cas, on s'intéresse uniquement au contenu du locuteur sans prendre en compte le contexte interactionnel qui s'avère crucial pour la compréhension des phénomènes d'opinions. Nous nous proposons dans cette thèse de modéliser à la fois l'interaction et les opinions contenues dans un tour de parole afin d'étudier ce phénomène

— dans une dynamique intra-locuteur avec une analyse de la communication humaine sous la forme d'une critique multimodale
— dans une dynamique inter-locuteurs avec une analyse d'une interaction humaine sous la forme d'une conversation entre un humain et un agent virtuel.

1.3 NOTRE MÉTHODOLOGIE

La méthodologie que nous avons utilisée dans cette thèse pour traiter le problème est basée sur une réponse à des interrogations qui se décomposent en 3 grandes questions de recherche (QR) :

QR 1 : Quels sont les descripteurs verbaux et vocaux pertinents que nous devons extraire pour modéliser une opinion ?

QR 2 : Quelle est la granularité de segmentation optimale pour découper le discours et l'interaction en unités efficaces pour une bonne représentation de la dynamique de l'opinion intra-locuteur et inter-locuteurs ?

QR 3 : Quel est le modèle d'apprentissage automatique à utiliser afin d'utiliser pleinement la dynamique et la séquentialité des opinions dans la parole ?

Ces trois QR sont résumées sur le schéma global de notre système présenté en figure 1.6. La QR1 est traitée par l'encadré de gauche portant sur les descripteurs à extraire, la QR2 par celui du milieu portant sur la segmentation à choisir et la QR3 par l'encadré de droite sur le modèle d'apprentissage.

Étant donné que nous utilisons des modèles d'apprentissage automatique supervisé, il est nécessaire d'utiliser un corpus d'interactions humaines orales annoté en opinion. Comme ce type de base de données n'existe pas à l'heure actuelle dans des dimensions assez grandes, nous avons fait annoter la base de données d'interactions humain-agent SEMAINE (McKeown et collab., 2012).

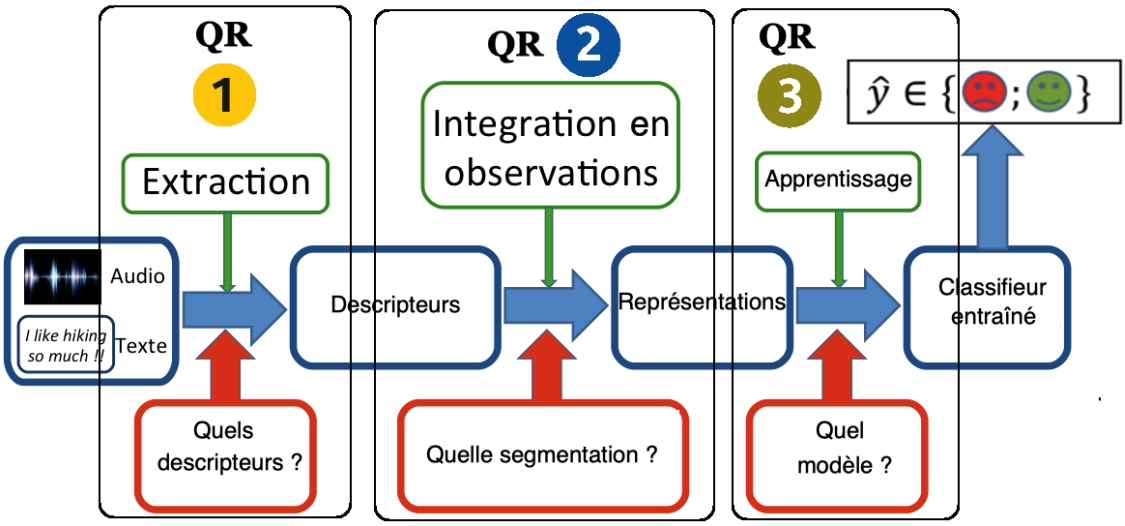


FIGURE 1.6: Diagramme en blocs de notre système avec les différents points à aborder

1.4 CONTRIBUTIONS ET ORGANISATION DE LA THÈSE

Contributions :

Les différents travaux de cette thèse sont dans le cadre de l'apprentissage automatique supervisé pour l'analyse des opinions. En répondant aux questions de recherches posées dans section 1.3, nous avons apporté différentes contributions : sur le processus d'annotation, sur les données annotées, sur l'étude d'une segmentation automatique et sur l'emploi de classifieurs de type HCRF pour l'analyse de phénomènes affectifs.
Les contributions de ce travail de thèse se résument ainsi :

— annotation en opinion d'un corpus d'interactions humaines, présentée dans le chapitre 5 ;
— une étude d'une segmentation avec les pauses du locuteur, présentée dans le chapitre 6 ;
— une adaptation du domaine écrit au domaine oral de mots-vecteurs appris extérieurement, présentée dans le chapitre 8 ;
— une modélisation de la dynamique intra-locuteur, présentée dans le chapitre 8 ;
— une modélisation de la dynamique inter-locuteurs, présentée dans le chapitre 9 ;
— une étude de l'utilisation jointe de l'audio et du texte pour de l'analyse d'opinion dans des interactions humaines orales, présentée dans le chapitre 10 ;

Cette thèse a donné lieu à plusieurs publications nationales et internationales listées ci-dessous, ainsi qu'à des communications comme au GT-ACAI [4].

- Barriere et collab. (2017) : **Valentin Barriere**, Chloe Clavel, and Slim Essid. Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields. In INTERSPEECH, 2017.
- Barriere (2017) : **Valentin Barriere**. Hybrid Models for Opinion Analysis in Speech Interactions. In ICMI, pages 647–651, 2017.
- Barriere et collab. (2018a) : **Valentin Barriere**, Chloe Clavel, and Slim Essid. Attitude Classification in Adjacency Pairs of a Human-Agent Interaction with Hidden Conditional Random Fields. In ICASSP, 2018.
- Barriere et collab. (2018b) : **Valentin Barriere**, Chloé Clavel, and Slim Essid. Classification d'attitude dans des paires adjacentes à l'aide de champs aléatoires conditionnels cachés. In WACAI, 2018.

Organisation de la thèse :

Le manuscrit s'organise en 3 grandes parties : la première partie est consacrée à l'état de l'art, la seconde à la méthodologie employée pour répondre aux questions de re-

4. https ://acai.limsi.fr/doku.php

cherche auxquelles nous avons répondu au cours de cette thèse, et la dernière aux différentes expériences que nous avons pu mener :

— La partie I comprend un état de l'art des différents blocs de la figure 1.6 pour un modèle d'Apprentissage automatique (AA)
  - Le chapitre 2 pour définir les descripteurs du signal de parole adaptés à la tâche et positionner dans le paysage scientifique les choix adoptés ;
  - Le chapitre 3 pour présenter les modèles d'apprentissage actuels, leurs défauts et qualités ;
  - Le chapitre 4 pour présenter les bases de données disponibles, leurs forces et faiblesses et pourquoi il a été nécessaire de faire une nouvelle annotation.
— La partie II est consacrée à notre méthodologie et présente en détail les choix réalisés dans notre travail de création d'une base de données et de modèles d'apprentissage
  - Le chapitre 5 pour présenter le travail effectué sur les base de données utilisées dans cette thèse ;
  - Le chapitre 6 pour présenter en détails les descripteurs que nous avons choisis pour représenter le signal de parole ;
  - Le chapitre 7 pour présenter les différents systèmes construits par rapport aux classifieurs, à la segmentation et à la représentation de l'interaction dyadique.
— La partie III contient les expériences que nous avons faites afin d'évaluer et de comparer nos modèles
  - Le chapitre 8 est consacré à l'analyse de l'opinion intra-locuteur ;
  - Le chapitre 9 est consacré à l'analyse de l'opinion inter-locuteurs ;
  - Le chapitre 10 est consacré à l'analyse multimodale de l'opinion.

Première partie

Apprentissage pour l'analyse d'opinions dans la parole : état de l'art

2

Descripteurs verbaux, vocaux et interactionnels

Résumé du chapitre

— Nous présentons les représentations textuelles existantes selon deux groupes : i) les descripteurs extraits de manière experte tels que des descripteurs issus de lexiques ou de structures linguistiques ; ii) les représentations distribuées apprises en amont sur de grandes quantités de données.

— Nous présentons les représentations audio existantes selon deux groupes : i) les descripteurs extraits de manière experte définis en sous-groupes de descripteurs cepstraux et spectraux, prosodiques et de qualité de voix ; ii) les représentations distribuées apprises en amont sur de grandes quantités de données.

— Il existe aussi des descripteurs et des modèles prenant en compte la dynamique interactionnelle.

— Les descripteurs utilisés dans cette thèse sont pour le texte des descripteurs linguistiques et issus de lexiques ainsi que des représentations distribuées et pour l'audio des descripteurs extraits de manière experte, principalement de qualité de la voix, et de représentations apprises.

L'analyse des phénomènes liés aux sentiments et aux opinions est un domaine qui a longtemps été, et est toujours, dominé par les modèles utilisant uniquement le texte. Les études centrées sur l'analyse de l'audio se concentrent en grande partie sur une tâche connexe, l'analyse d'émotions. Une autre différence est que les textes sont des compositions ordonnées de mots qui sont des valeurs discrètes ayant des relations sémantiques entre eux, alors que l'audio est un signal physique continu. Il apparaît donc important de séparer l'état de l'art des descripteurs utilisés pour l'analyse d'opinion sur ces 2 domaines.

Une autre différence entre les modalités vient du fait que les opinions se manifestent sous des formes différentes à l'écrit et à l'oral. Cela peut être une structure linguistique syntaxique dans le texte alors que la manifestation acoustique dans le même signal sera une augmentation de la fréquence fondamentale sur un mot particulier dans la voix. **Ces structures possèdent des temporalités différentes qui ne sont pas forcément synchronisées**. Pour cela, afin de modéliser l'opinion dans un contexte multimodal il est nécessaire d'effectuer un état de l'art et un travail de réflexion sur la fusion de données multimodales. Bien qu'important à mentionner, cet aspect a été jugé plus dépendant des modèles que des descripteurs et sera abordé dans la partie des modèles (chapitre 3).

Dernier point mais non des moindres, il paraissait important de s'attarder sur les différents travaux qui ont porté sur l'analyse des opinions dans un contexte interactionnel. En effet peu de systèmes se sont attelés à utiliser les interactions orales ou écrites entre les individus pour analyser ou modéliser les phénomènes liés aux opinions.

Pour résumer, ce chapitre présentera les différents descripteurs utilisés avec des méthodes d'apprentissage automatique pour l'analyse des phénomènes liés aux sentiments et aux opinions. Nous commencerons par les descripteurs représentant l'information contenue dans un texte (section 2.1), puis dans un signal acoustique (section 2.2). Enfin, une présentation des travaux existants étudiant les phénomènes liés aux opinions dans un contexte interactionnel sera effectué en section 2.3 avant un positionnement global sur tous ces points.

2.1 Représentations textuelles

Le domaine de l'analyse d'opinion s'est considérablement développé ces dernières années. La tâche peut être effectuée à des granularités différentes : au niveau d'un document entier (Pang et Lee, 2004), d'un texte court (Taboada et collab., 2011), d'une phrase (Wilson et collab., 2005) ou au niveau du mot (Warriner et collab., 2013). Les différentes approches regroupent des méthodes non-supervisées utilisant des lexiques de sentiments (Taboada et collab., 2011) qui ne dépendent pas du sujet traité dans le

texte : ces approches utilisent des lexiques créés en amont et donnent des information locales au niveau d'un mot (sous-section 2.1.1). On compte des approches supervisées qui utilisent des représentations basiques dépendant du corpus comme le Sac de N grammes (*Bag-of-N-Grams* : BoNG) (Schuller et collab., 2009a) qui donne des informations au niveau des N-grammes. On compte aussi des représentations distribuées plus complexes que l'on peut pré-entrainer séparément ou non (Mikolov et collab., 2013a) et qui peuvent représenter un mot, une expression multi-mots, une phrase ou un document. Finalement, on compte aussi les approches hybrides (Johansson et Alessandro, 2010; Yang et Cardie, 2013; Breck et collab., 2007), combinant la robustesse et la généralisation des algorithmes d'AA avec la stabilité des lexiques et des structures linguistiques : ces approches utilisent des descripteurs linguistiques (Sous-section 2.1.2) et d'autres représentations dans des méthodes d'apprentissage automatique.

Plus récemment, les approches utilisant les descripteurs développés de manière experte sont de plus en plus délaissées pour des modèles utilisant des descripteurs provenant d'un apprentissage, ou bien des approches de *bout-en-bout* plus générales nécessitant que de grandes quantités de données soient déjà disponibles (Sous-section 2.1.3).

2.1.1 Descripteurs issus de lexiques

Les descripteurs issus de lexiques permettent d'utiliser dans la représentation du signal de la connaissance humaine et d'injecter des informations sur une qualité, un aspect particulier de certains mots. Utiliser de la connaissance amassée en amont (par exemple via un pré-entrainement) enrichit la représentation du signal, permettant au modèle d'apprentissage d'observer des indices qui sont jugés importants pour la réalisation de la tâche.

Les lexiques sont spécialement efficaces pour faire une adaptation de domaine, repérer des mots particuliers que l'on sait intéressants pour une tâche spécifique et ainsi permettre au système de repérer les aspects les plus importants dans le discours du locuteur. Ces mots changent selon le domaine dans lequel on se place : la réalisation du phénomène d'opinion dans des discussions ouvertes n'aura pas les mêmes codes que dans un avis en ligne servant à noter un appareil photo.

Un exemple : en analyse de sentiments classique, savoir si le mot "*enjoyable*" est plutôt négatif ou positif peut avoir une grande importance dans la représentation du signal utilisée. Cependant, selon le type de domaine analysé, ces valeurs pourront changer selon les mots et même être opposés pour un même mot : "*long*" qui est un adjectif positif lorsqu'on parle de la batterie d'un appareil, devient négatif lorsqu'il est employé pour décrire le temps d'attente dans un restaurant.

Concernant les lexiques de subjectivité, il existe de nombreuses ressources qui ont

été créées au fil des années et qui diffèrent en terme d'annotation. Bradley et Lang (1999, 2010) introduisent le lexique ANEW (*Affective Norms for English Words*) qui contient 1000 mots annotés sous 3 axes classiques : la valence, la dominance et l'activation, à l'aide du SAM (*Self-Assessment Manikin*). Le SAM est une méthode classique utilisée pour l'annotation selon des axes différents tels que Valence, Activation, Dominance, Surprise (voir Bradley et Lang (1994) pour plus de détails). Trouvant que ce lexique est limité pour certains domaines, car ne contenant pas d'obscénités, Nielsen (2011) propose une version de ce lexique de 2477 mots, pour les données provenant de microblogs. Le lexique classique est amélioré par Warriner et collab. (2013) pour atteindre une taille d'environ 14000 mots.

Baccianella et collab. (2010) et Strapparava et Valitutti (2004) proposent des annotations complémentaires du graphe sémantique WordNet de Fellbaum (1998). Le premier lexique, qui contient des valeurs selon la valence, s'appelle SentiWordNet (SWN). Le second, contenant des catégories affectives s'appelle WordNet-Affect. SWN notamment a été utilisé par Morency et collab. (2011) pour une tâche d'analyse de sentiments multimodale dans des Vlogs (voir la section 4.1 pour plus de détails sur les Vlogs).

Wilson et collab. (2005) créent un lexique de subjectivité (lexique MPQA) qu'ils emploient dans un système à base de règles utilisant le contexte textuel pour trouver la polarité d'une phrase dans le corpus MPQA. Ce même lexique MPQA est aussi utilisé par Breck et collab. (2007) pour une tâche d'extraction d'opinion dans des textes du corpus MPQA.

Musto et collab. (2014) font une étude comparative de SenticNet, du lexique MPQA, de SWN et de WordNet-Affect pour une tâche d'analyse de sentiments dans des tweets. Les résultats montrent que les lexiques MPQA et SWN ont des résultats plus stables.

Mohammad (2018) propose un lexique de subjectivité, composé de 20 000 mots annotés en Valence, Activation et Dominance, comme ANEW. Les auteurs affirment que ces valeurs changent peu entre les différents langues et fournissent ce lexique pour plus d'une centaine de langues, grâce à un outil de traduction automatique fourni sur internet. [1]

Pour une tâche d'extraction d'opinions, de source et d'entités jointes, Choi et collab. (2006) utilisent comme descripteurs les hyperonymes qui sont extraits de WordNet. Les hyperonymes sont les mots de catégorie générique dont le sens inclut celui de plusieurs mots. Par exemple le mot canin est un hyperonyme de chien et de loup. Ce descripteur permet au système de généraliser.

1. Google Translate

2.1.2 Descripteurs issus de règles syntaxiques

Il existe des descripteurs issus de règles syntaxiques, certains sont très simples comme par exemple la nature grammaticale du mot, et d'autres plus complexes comme des structures linguistiques flexibles.

La nature grammaticale des mots est un indice qui a longtemps été jugé utile et utilisé pour l'analyse d'opinions. De plus, l'utilisation de la nature grammaticale permet de désambiguïser le sens du mot afin de gérer la polysémie dans certains cas. Breck et collab. (2007) utilisent la nature grammaticale des mots pour une tâche d'extraction d'opinions dans des textes du corpus MPQA, à l'aide de champs aléatoires conditionnels.

Pour une tâche d'analyse de sentiment ciblée, Poria et collab. (2016) utilisent comme descripteurs, entre autres, des structures linguistiques constituées d'un ensemble de règles basées sur la nature grammaticale des mots et sur des lexiques de subjectivité. Par exemple, si un mot est un complément d'un verbe attributif (comme *be*, *appear*, *seem*, *look*, ...), alors ce mot est marqué comme un aspect de l'entité de la critique.

Hutto et Gilbert (2014) introduisent un modèle d'analyse de sentiments simple à base de règles. Le modèle utilise un lexique de subjectivité validé par des humains et des règles générales relatives à la grammaire et à la syntaxe. Il permet de capturer des règles généralisables et des heuristiques associées aux indices grammaticaux et syntaxiques utilisés par les utilisateurs. Le modèle est testé sur des tweets.

Pour une tâche d'extraction d'opinion et de cible et source associés, Yang et Cardie (2013) construisent un modèle de champs aléatoires conditionnels utilisant de nombreux descripteurs issus de règles syntaxiques. L'arbre de dépendance est calculé pour obtenir la distance la plus courte entre une opinion et une cible potentielle ainsi que pour obtenir la catégorie syntaxique du composant le plus profond qui contient la cible candidate.

Langlet et Clavel (2015b) construisent des règles sur des structures linguistiques pour analyser des attitudes dans le corpus d'interaction humain-agent SEMAINE. Ces structures utilisent des descripteurs basés sur des lexiques de subjectivité, sur la nature grammaticale des mots et sur la syntaxe de la phrase à l'aide d'un arbre de dépendance.

2.1.3 Représentations apprises distribuées

Une autre possibilité très en vogue ces derniers temps pour représenter les données est d'apprendre de manière non supervisée des descripteurs que l'on utilisera par la suite. Ces représentations proviennent d'un apprentissage par un modèle de type réseaux de neurones et sont appelés représentations distribuées. Hinton et collab. (1986) insistent sur le pouvoir de ces représentations : elles permettent de partager

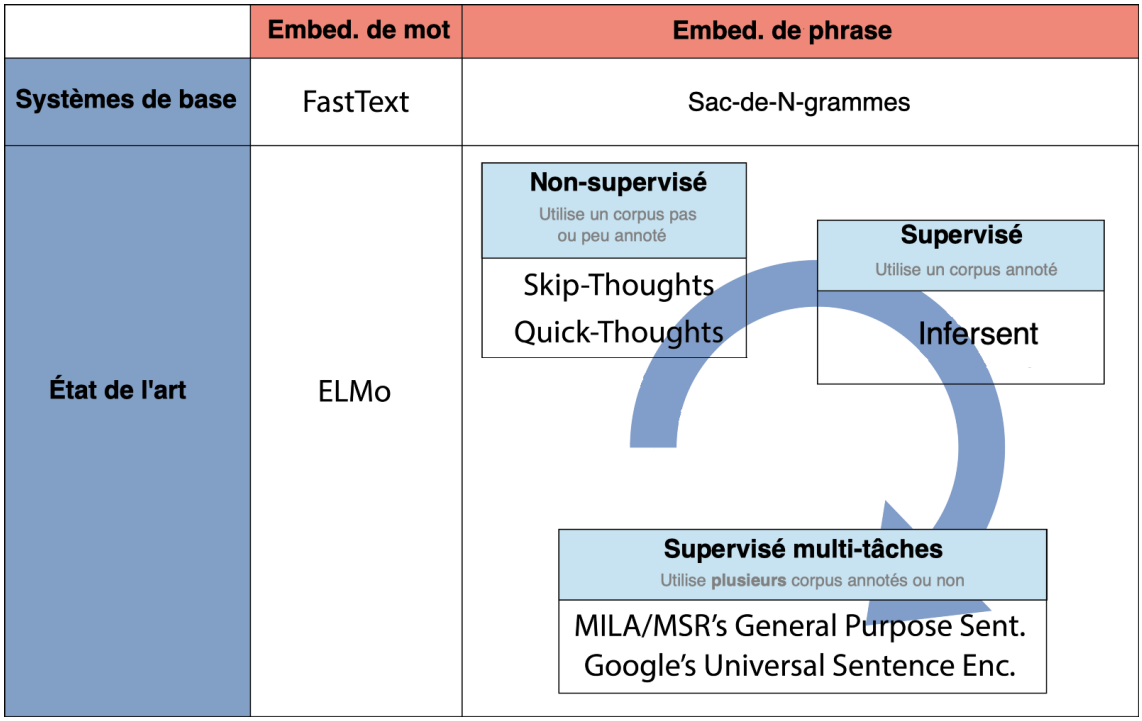


FIGURE 2.1: Récentes avancées en apprentissage de représentations sur du texte

des concepts entre les différentes données qu'elles représentent, via une représentation dans un espace commun de faible dimension. L'apprentissage de ces représentations sur les textes appelées plongements lexicaux (*embeddings*), demande une grande quantité de texte disponible pas forcément annotée. On peut, par exemple, utiliser les articles de Wikipédia. Cette apprentissage permet au modèle d'apprendre la structure sous-jacente du langage et les liens entre les différents mots et d'utiliser de la connaissance emmagasinée depuis une grande quantité de données. L'apprentissage peut être fait sur le corpus de la tâche ou bien sur un autre corpus de taille bien plus importante.

Depuis quelques années, la tendance a été de faire une sorte d'apprentissage par transfert (*transfer learning*) en apprenant des représentations du texte de manière non supervisée pour les utiliser ensuite dans une tâche supervisée. Cette technique permet de commencer l'entraînement de son système avec une représentation contenant déjà de l'information correspondant à la sémantique du mot, de la phrase ou bien du paragraphe Kiros et collab. (2015); Mikolov et collab. (2017); Peters et collab. (2018a). Depuis le début de l'année 2018, si l'apprentissage par transfert est toujours de rigueur, la manière d'effectuer le transfert de connaissance a changé. La tendance est maintenant à un apprentissage supervisé général, multitâche et plus profond que les précédentes méthodes Cer et collab. (2018); Howard et Ruder (2018); Conneau et collab. (2017) (voir Figure 2.1).

2.1.3.1 Plongement lexicaux sur des mots

Poria et collab. (2015) utilisent un plongement lexical de mots (*word embedding*) appelé word2vec (Mikolov et collab., 2013c) pour une tâche de classification binaire de valence sur de la parole pendant que Irsoy et Cardie (2014) utilisent la même représentation pour une tâche d'extraction d'expressions subjectives directives et expressives sur des données textuelles avec le corpus MPQA. Une autre représentation équivalente à word2vec, GloVe de Pennington et collab. (2014), est aussi utilisée de manière similaire, comme par Zadeh et collab. (2017) pour une tâche d'analyse de sentiments sur des Vlogs.

Il existe aussi des plongement lexicaux de mots spécifiquement créés pour l'analyse de sentiment. Tang et collab. (2014) développèrent une alternative au modèle de plongement lexical de mots C&W de Collobert et collab. (2011) appelé SSWE (*Sentiment Specific Word Embedding*). Spécialement créé pour l'analyse de sentiments, ce modèle utilise les annotations en sentiments directement dans son apprentissage en s'appuyant sur le fait qu'elles apportent de grandes quantités d'information. Khosla et collab. (2017) présentent Aff2vec qui utilise les valeurs de Valence, Activation et Domination des mots du lexique ANEW de Warriner et collab. (2013) pour la création de ses vecteurs. Cependant, ces plongements sont appris sur de petites quantités de texte et il est notoire que plus l'ensemble d'entrainement est grand plus les représentations sont de qualité élevée. Sahay et collab. (2018) ont utilisé les mots-vecteurs ELMo de Peters et collab. (2018a) pour une tâche d'analyse de sentiment sur le corpus multimodal de critiques Vlogs CMU-MOSEI de Zadeh et collab. (2018c). ELMo permet de créer un vecteur en fonction du contexte de la phrase dans lequel le mot est employé : un même mot aura donc plusieurs vecteurs s'il est employé dans différentes phrases. Ainsi, une première désambiguïsation est effectuée pour gérer la polysémie.

2.1.3.2 Plongement lexicaux sur des paragraphes

Les représentations peuvent être apprises sur différents niveaux lexicaux, ainsi il est possible d'utiliser des plongements lexicaux de paragraphes. Ce domaine est aussi en expansion depuis les dernières années, plus particulièrement à partir de l'arrivée de *doc2vec* (Le et Mikolov, 2014), des méthodes comme *Skip-Thought* Kiros et collab. (2015), *Infersent* Conneau et collab. (2017) et le *Google's Universal Sentence Encoder* Cer et collab. (2018) ont vu le jour.

La recherche de l'*embedding* général le plus performant est un domaine aussi en expansion depuis les dernières années. Cela consiste à trouver la méthode d'apprentissage permettant d'obtenir des représentations de paragraphes ou de documents de qualité pour toute sorte de tâches. Une des tâches est généralement l'analyse de sen-

timent. Ainsi toutes ces méthodes sont évaluées au moins sur une tâche d'analyse de sentiment, généralement sur le corpus de critiques de films IMDb.

L'avantage de ces représentations est que l'ensemble du document est contenu dans un unique vecteur le représentant et qui peut contenir énormément d'information. En utilisant ces représentations en entrée d'un classifieur, l'avantage est de ne plus avoir à modéliser la séquentialité de la parole, car cette connaissance complexe est déjà englobée dans le modèle de génération des vecteurs qui utilisent des RNN. Le désavantage est de ne plus pouvoir interpréter la séquentialité de la parole car elle est intrinsèque au vecteur dans un espace abstrait de grande dimension. Un autre inconvénient est qu'utiliser un vecteur représentant un morceau de texte conséquent nécessite d'utiliser une représentation de l'audio d'une taille aussi conséquente, ce qui ne rend pas forcement compte de la temporalité des manifestations de l'opinion dans le signal vocal.

2.2 Représentations acoustiques

Du côté de l'audio, différents choix sont possibles quant à la représentation du signal audio mais l'utilisation de descripteurs liés au sentiment et à l'émotion est plus répandue. Des études récentes sur des tâches liées à l'analyse de sentiments à l'oral s'attachent à la détermination des descripteurs physiques du signal audio complémentaires pour créer le meilleur ensemble de descripteurs adapté aux tâches liés à l'émotion et aux sentiments (Eyben et collab., 2015; Tahon et Devillers, 2016). D'autres systèmes se concentrent sur un apprentissage de bout-en-bout (*end-to-end*) utilisant des techniques d'apprentissage profond avec une grande quantité de paramètres (Tzirakis et collab., 2017). Enfin, il est aussi possible d'utiliser des méthodes d'apprentissage non supervisé sur une grande quantité de données, à la manière de ce qui se fait pour le texte, puis d'utiliser ces représentations en entrée de classifieurs Freitag et collab. (2017).

Dans les tâches liées à l'analyse de sentiments, on retrouve souvent les mêmes descripteurs extraits qui sont utilisés avec différents modèles d'apprentissage. Il est possible de séparer les représentations en 4 catégories différentes (Clavel, 2007) :

— *Les descripteurs cepstraux et spectraux* qui sont caractéristiques du contenu vocal et de la structuration du flux oral. Ils servent à décrire les types de phonèmes ainsi que la manière dont la source (dans une modélisation source-filtre de la parole Fant (1970)) émet les ondes mécaniques sonores avant qu'elles n'entrent dans le conduit vocal.

— *Les descripteurs de la prosodie* : ils sont caractéristiques de la manière de parler, comme le débit de parole, les intonations et les accents toniques utilisés. Par conséquent, ces descripteurs permettent de caractériser l'état émotionnel du

locuteur

— *Les descripteurs de qualité de voix* : ils sont en majorité caractéristiques de la configuration glottale et par conséquent des changement physiologiques du locuteur. Ces paramètres qui ont été utilisés pour de multiples études liées à l'analyse de sentiments sont principalement tirés du modèle du flux glottal de Linjencrant-Fant (Fant, 1995) et liés aux instants de fermeture glottale.
— *Les représentations générées* : ce sont des représentations provenant d'un apprentissage. Comme à la manière des représentations textuelles, on peut distinguer plusieurs types d'apprentissage. L'apprentissage peut être fait au préalable, permettant d'aboutir à la création d'un modèle capable de générer des représentations distribuées contenant une grande quantité d'information sémantique à partir d'un signal audio, que ce soit de manière supervisée (Abu-El-Haija et collab., 2016) ou bien non supervisée (Freitag et collab., 2017). L'apprentissage peut aussi être fait de manière jointe avec la tâche (Trigeorgis et collab., 2016). Les représentations en entrée du modèle d'apprentissage peuvent aussi bien être une représentation temps-fréquence (Amiriparian et collab., 2017) ou bien le signal brut (Trigeorgis et collab., 2016; Tzirakis et collab., 2017).

Nous allons présenter les différents types des descripteurs et représentations de chaque groupe en les replaçant dans l'état de l'art, en décrivant précisément les sens physiques des descripteurs acoustiques utilisés et leurs intérêts. On commencera en sous-section 2.2.1 par des descripteurs physiques classiques, puis on parlera des ensembles qui sont utilisés afin d'obtenir des représentations cohérentes (Sous-section 2.2.2), et enfin on abordera les représentations apprises (Sous-section 2.2.3).

2.2.1 Représentations provenant de descripteurs extraits de façon experte

La voix est un signal quasi-stationnaire (il change peu sur des périodes allant de 5 à 100ms) et les sons produits peuvent être très différents, avec de gros changements dans la structure du signal. Ce dernier peut être semi-périodique ou aléatoire, selon si la parole est voisée (c'est-à-dire que les cordes vocales vibrent) ou non-voisée.

Un signal de parole est composé de parties voisées et non-voisées. Les parties voisées sont quasi-périodiques, il est donc possible de prendre le spectre du signal pour l'analyser. Regarder les différentes relations entre les formants peut apporter des informations descriptives, par exemple sur les différents phonèmes.

Les descripteurs comme l'intensité, la hauteur et des descripteurs de qualité de la voix sont intéressants pour l'analyse des phénomènes liés à l'émotion dans la parole et la littérature en psychologie recommande leur utilisation. Il est aussi utile d'utiliser d'autres descripteurs qui peuvent contenir des informations latentes et non-linéaires,

non contenues dans les premiers descripteurs. Cependant, il ne faut pas utiliser un ensemble trop grand pouvant bruiter les données. Nous avons décidé de construire un ensemble avec des descripteurs adaptés et aussi de tester une représentation générée dans notre travail de thèse.

2.2.1.1 Descripteurs de la prosodie

Ils sont caractéristiques de la structure de la parole, et responsables de l'intelligibilité lorsqu'on modélise la parole. Ces descripteurs sont appropriés pour la modélisation de l'état émotionnel du locuteur. Finalement, on abordera l'intégration de ces descripteurs, extraits à fréquence temporelle élevée, pour obtenir une représentation de l'acoustique au niveau de l'unité utilisée pour la segmentation.

La prosodie permet d'apporter divers types d'information. Par exemple, elle contient des informations sur le contenu syntaxique des phrases (Warren, 1996) : un pitch descendant est généralement significatif d'une fin de phrase ou d'une affirmation alors qu'un contour de pitch montant sera généralement présent dans une phrase interrogative.

L'accent tonique sert aussi à accentuer un mot par rapport à d'autres, ainsi il est aussi lié au contenu linguistique car il permet d'attirer l'attention du locuteur sur une partie de la phrase énoncé, et enrichit la quantité d'information contenue uniquement dans la syntaxe (Paul et collab., 2005).

Pour finir, l'utilisation de pauses est un indice important de la structure discursive (Shriberg et collab., 2000). Une pause verbalisée (par exemple : "*heuu*") est aussi un indice important de l'hésitation du locuteur, mais aussi du désir qu'il a de continuer son tour de parole s'il est en interaction, ou du moins sa phrase s'il est en communication sans interlocuteur. Cependant, certaines pauses ont aussi simplement des rôles respiratoires et articulaires et sont sans lien avec l'information verbale (Campione et Véronis, 2002).

2.2.1.2 Descripteurs cepstraux et spectraux

Ils sont caractéristiques des fréquences et de l'enveloppe spectrale du signal de parole de manière générale. Ces descripteurs sont très généraux, et permettent d'obtenir de bons résultats sur de nombreuses tâches utilisant le signal de parole.

Nous nous intéresserons en particulier aux MFCC (*MFCC* : Mel Frequency Cepsral Coefficient), qui sont des coefficients cepstraux et contiennent beaucoup d'information. Utilisés à l'origine pour des tâches de reconnaissance de la parole, les coefficients MFCC sont devenus indispensables pour des tâches comme l'analyse des émotions ou des intentions du locuteur. Les coefficients cepstraux se sont montrés très utiles

pour modéliser les états affectifs des locuteurs comme les émotions (Schuller et collab., 2007), l'interêt (Schuller et Rigoll, 2009) ou détecter l'autisme (Marchi et collab., 2012).

Il existe de nombreuses manières de modéliser le signal de parole et l'une d'elle est le modèle source-filtre de Fant (1981). Ce modèle présente le signal comme la résultante d'un produit de convolution entre le signal émis par les cordes vocales qui sont considérées comme une source, et le conduit vocal qui est considéré comme un filtre.

Les MFCC permettent de séparer ces deux composantes par une simple fonction mathématique utilisant, entre autres, transformées de Fourier et logarithme. Ceci permet de filtrer les variabilités du signal provenant du conduit vocal afin de se concentrer sur la partie source. Il est important de normaliser ces coefficients qui ont été utilisés de nombreuses fois pour des tâches de reconnaissance du locuteur, afin de ne pas biaiser les systèmes de reconnaissance de phénomènes liées aux opinions par des variations qui sont dues au locuteur.

2.2.1.3 Descripteurs de qualité de voix et modèle Linjencrant-Fant (LF)

Ils sont en majorité caractéristiques de la configuration glottale et par conséquent des changements physiologiques du locuteur. Les émotions étant fortement liées aux changements physiologiques comme le rythme cardiaque ou la tension dans le conduit vocal (Clavel, 2007), il est important de compter ces paramètres dans notre ensemble de descripteurs acoustiques (EDA). Ces paramètres qui ont été utilisés pour de multiples études liées à l'analyse de sentiments sont principalement tirés du modèle du flux glottal de Linjencrant-Fant (Fant, 1995) (LF) et liés aux instants de fermeture glottale.

Modèle Linjencrant-Fant (LF) du flux glottal :
L'estimation du flux glottal est le procédé d'estimation du conduit vocal et des composantes du flux glottal depuis un signal de parole ; décrit dans Fant et collab. (1985); Fant (1995) comme le modèle Liljencrants-Fant (LF). Plus précisément, le modèle LF est un modèle à 5 paramètres du flux glottal différencié composé de 2 phases : les phases d'ouverture et de fermeture.

Le calcul du modèle LF est fait comme décrit dans Scherer et collab. (2013) à l'aide de l'algorithme présenté par Gobl et Ní Chasaide (2003) (voir en annexe D pour plus de détails).

Pour estimer le modèle LF, la première nécessité est de détecter chaque instant de fermeture glottal (*Glottal Closure Instants* : GCI) à l'aide de la méthode fournie par Drugman et Dutoit (2009). Une fois que ces instants sont estimés, il est possible d'appliquer un filtrage inverse adaptif itératif (*Iterative Adaptive Inverse Filtering* : IAIF) (Alku, 1992) sur des fenêtres temporelles au voisinage des GCI. Ce filtrage consiste à

appliquer des prédictions linéaires à haut et bas ordres et utiliser l'inverse des filtres pour estimer à la fois le flux glottal et le conduit vocal, comme il est classiquement modélisé dans le modèle source-filtre. Le modèle du flux glottal permet ensuite d'extraire les paramètres d'après leurs définitions (voir chapitre 6 où sont présentés les différents descripteurs utilisés).

2.2.1.4 Intégration

Les descripteurs sont généralement extraits à une fréquence élevée par rapport à la fréquence moyenne qui est utilisée pour la segmentation. Par conséquent, pour chaque descripteur on obtient une séquence de valeurs de taille variable sur chaque observation. Ainsi, une méthode commune permettant de représenter l'ensemble de ces séquences de tailles variables en un vecteur de la même taille est d'**utiliser des fonctionnelles décrivant l'évolution du descripteur dans l'intervalle temporel**. Un ensemble de fonctionnelles utilisé pour la création des représentations pour le défi Émotion 2009 (Schuller et collab., 2009b) est disponible en Tableau 2.1.

Fonctionnelles	Type
Moyenne arithmétique, Position relative	Temporel
Amplitude, Minimum, Maximum	Temporel
Écart-type, Asymétrie, Coefficient d'acuité	Moments
Pente et Erreur moyenne quadratique	Régression linéaire

TABLE 2.1: Ensemble de fonctionnelles appliquées pour la création de représentations dans le cadre du défi de reconnaissance d'émotion d'Interspeech 2009 (Schuller et collab., 2009b)

L'utilisation de fonctionnelles permet de créer un vecteur de valeurs représentant la dynamique d'un descripteur extrait sur un intervalle temporel. Comme toute représentation, les fonctionnelles peuvent être plus ou moins adaptées à une bonne représentation du signal.

2.2.2 Ensemble de descripteurs et outils d'extraction

Dans les différents travaux que l'on peut trouver dans la littérature, les auteurs utilisent généralement les descripteurs par **ensembles cohérents**. Beaucoup de travaux étudiant les phénomènes liés aux opinions à l'aide du contenu vocal du locuteur ont utilisé des ensemble de descripteurs provenant de deux sources.

Dans l'analyse de sentiments, d'émotions ou d'opinions c'est majoritairement des descripteurs automatiquement extraits grâce à des boites à outils. La première et plus

classique est OpenSMILE de Eyben et collab. (2013) qui a été utilisée pour de nombreux défis mis en place lors de différentes conférences (Sous-section 2.2.2.1). Les ensembles de descripteurs extraits avec OpenSMILE sont généraux, pour l'étude de la musique comme de la voix. La deuxième boite à outils est COVAREP de Degottex et collab. (2014), elle a aussi été utilisée lors de nombreux travaux sur l'analyse de l'opinion et permet d'extraire des descripteurs plus spécialisés basés sur la qualité de la voix (Sous-section 2.2.2.2).

2.2.2.1 OpenSMILE : Ensembles généraux

OpenSMILE de Eyben et collab. (2009); Eyben (2010); Eyben et collab. (2013) est un logiciel regroupant des implémentation de méthodes d'extraction de nombreux descripteurs. Des ensembles de ces descripteurs ont été utilisés pour de nombreuses tâches liées à la détection de phénomènes paralinguistiques, émotionnels, d'états ou de traits du locuteur. Ces ensembles sont ainsi devenus des références dans le domaine, afin de comparer les différents modèles entre eux. Il est possible de paramétrer l'extraction pour obtenir les ensembles utilisés lors de ces différentes tâches. Les descripteurs contenus dans les différents ensembles sont assez généraux et peuvent être utilisés aussi pour la musique Weninger et collab. (2013). Dernièrement, un ensemble de descripteurs particulier a été mis en valeur par Eyben et collab. (2015) pour être l'ensemble synthétique contenant les descripteurs nécessaires à l'analyse de paralinguistique comme les tâches présentées ci-dessous.

OpenSMILE a été utilisé pour extraire des descripteurs pour des défis portant sur la détection de phénomènes para-linguistiques dans la voix, tels que la détection :

— de l'émotion Schuller et collab. (2009b) ;
— du degré d'alcoolisation Schuller et collab. (2011a) ;
— d'autisme et de niveau de conflit Schuller et collab. (2013) ;
— de mensonge et d'accent Schuller et collab. (2016).

Ces ensembles de descripteurs peuvent être aussi utilisés pour des tâches de prédiction de valence et d'activation dans la musique Pellegrini et Barriere (2015), les liens entre émotions dans la voix et dans la musique étant forts Weninger et collab. (2013).

2.2.2.2 COVAREP : Ensembles centrés sur la qualité de voix

COVAREP est une autre API permettant d'extraire un ensemble de descripteurs audio, qui est basé principalement sur la qualité de la voix. Les descripteurs extraits ont été utilisés pour toutes sortes de tâches liées à l'expressivité des émotions, ou des opinions comme :

— En entrée d'un modèle d'apprentissage pour une tâche d'analyse de sentiments multimodale sur le corpus MOSI par Zadeh et collab. (2017)
— Pour l'étude de la dépression par Scherer et collab. (2014); Venek et collab. (2014)
— Pour des tâches d'analyse du pouvoir de persuasion d'un locuteur Brilman et Scherer (2015); Park et collab. (2014, 2015)
— En entrée d'un modèle d'apprentissage pour diverses tâches de compréhension de la communication humaine multimodale par Zadeh et collab. (2018c,a,b)

Les différentes études ayant été menées avec l'aide des paramètres extraits par COVAREP sont de bons indicateurs de la qualité de ces représentations pour notre tâche.

2.2.3 Représentations apprises

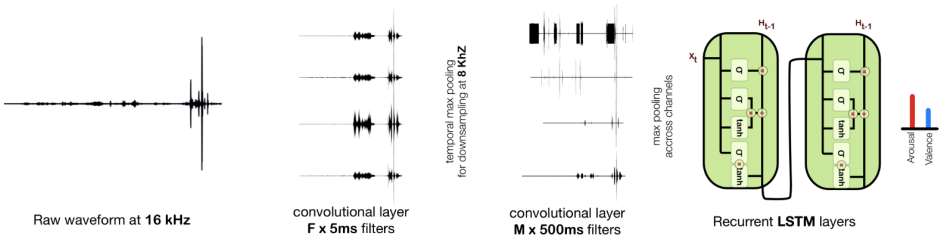
Les approches classiques et qui sont utilisées depuis des décennies fonctionnent sur le même schéma : les descripteurs extraits pour l'analyse des phénomènes paralinguistiques sont des descripteurs que des études en recherche auditive et psychologique ont trouvé reliés à certains phénomènes (Scherer, 2003). On extrait donc des descripteurs qui sont créés et choisis par des humains et qui sont reconnus comme utiles dans la caractérisation de certains phénomènes. Cependant, des phénomènes haut-niveau comme les opinions ne sont pas aussi aisément caractérisables dans la voix que dans le texte. **Les descripteurs du signal vocal classiquement utilisés ne représentent peut-être pas ce qui est important dans le signal pour les tâches visées, et il est possible qu'une approche utilisant le signal brut soit plus efficace pour intrinsèquement extraire des représentations spécifiquement utiles à une certaine tâche**.

Ghosh et collab. (2016a) étudient des représentations apprises et les compare à des descripteurs extraits complexes tels que ceux obtenus avec la boite à outils COVAREP (Degottex et collab., 2014) sur une tâche de reconnaissance d'émotion vocale dans le corpus IEMOCAP (Busso et collab., 2008). Dans cet article, les auteurs démontrent que l'apprentissage de représentation à l'aide de couches d'auto-encodeurs entrainés sans descripteurs extraits mais avec des représentation temps-fréquence, donne des représentations plus efficaces que des descripteurs complexes extraits comme des MFCC ou des descripteurs de qualité de voix.

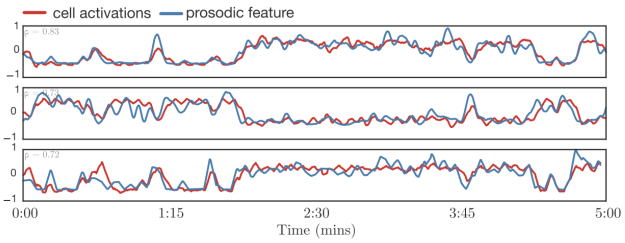
Freitag et collab. (2017) utilisent des représentations temps-fréquence afin d'apprendre une représentation dense distribuée à l'aide d'auto-encodeurs de type RNN-LSTM (*RNN-LSTM* : Recurrent Neural Network - Long Short Term Memory). En comparaison avec Ghosh et collab. (2016a) qui utilisaient des descripteurs extraits en entrée des auto-encodeurs, les auteurs de (Freitag et collab., 2017) n'utilisent aucun descripteur extrait et partent du signal pur. Cet approche permet d'entrainer un modèle à générer des représentations sur de larges bases de données non annotées, et d'utiliser

ces représentations pour des tâches atypiques avec peu de données.

La question de recherche précédente (en gras plus haut) quant à la pertinence des descripteurs extraits est aussi posée par Trigeorgis et collab. (2016) pour une tâche de reconnaissance d'émotion continue (régression dans l'espace Valence-Activation) dans le signal vocal. Dans cet article, les auteurs proposent un modèle d'apprentissage de bout-en-bout c'est à dire sans extraction de descripteurs, simplement à l'aide du signal brut. Le modèle d'apprentissage est un hybride de réseaux de neurones convolutif et récurrent CNN-LSTM (sous-figure 2.2a). La partie CNN permet de repérer des structures locales dans le signal de parole et d'obtenir des représentations moins denses qu'un signal audio brut. La partie LSTM permet de repérer des dépendances temporelles éloignés. Ce système arrive a obtenir d'excellent résultats, mais un des aspects les plus important pour les auteurs est la capacité du réseau à "retrouver" des descripteurs déjà existants. En analysant les activations des différentes cellules du RNN, il est possible d'observer la corrélation avec certains descripteurs extraits à la main comme vu en sous-figure 2.2b. Cela permet de "valider" avec une approche d'apprentissage la pertinence des descripteurs non appris, tout en obtenant des descripteurs proches mais différents, plus adaptés à une tâche avec un certain type de données.



(a) CNN-LSTM utilisé pour la prédiction de Valence et d'Activation.



(b) Comparaison des activations avec des descripteurs prosodiques extraits.

FIGURE 2.2: Approche de bout-en-bout par Trigeorgis et collab. (2016)

Ghosh et collab. (2016b) ont aussi fait de l'apprentissage de représentations avec des RNN pour de la reconnaissance d'émotions sur de courtes énonciations du corpus

IEMOCAP.

2.3 OPINIONS DANS LES INTERACTIONS

Dans cette partie nous allons aborder les études qui ont été faites en prenant en compte un contexte d'interaction Humain-Humain ou Humain-Agent dans la création des modèles ou dans la création des descripteurs.

*** :**Descripteurs

Si les applications et les études traitant de l'analyse de sentiments à l'oral sont très nombreuses, les études proposant des modèles effectuant de l'analyse de sentiments dans les interactions sont bien moins fréquentes. Clavel et Callejas (2016) font une intéressante revue sur les différents stratégies d'analyse de sentiments dans la généralité et dans les interactions Humain-Agent. On trouve des environnements de travail pour des interactions entre humain et agent mais avec des modules de détection de sentiments et d'émotions qui ne prennent pas en compte l'interaction.

Yildirim et collab. (2011) utilisent comme descripteurs les actes de dialogue précédents avec un contexte de 2 tours de parole pour prédire l'état émotionnel d'enfants dans une interaction Humain-Agent. Dans cet article, les actes de dialogue qui sont utilisés comme descripteurs sont annotés à la main, ce qui n'est pas pratique lorsqu'on veut faire un système automatique, car couteux en temps humain. Ces descripteurs obtiennent de bons résultats par rapport aux N-grammes. Henderson (2015) passent en revue les différentes techniques d'AA qui sont disponibles afin de faire de l'annotation automatique d'actes de dialogue. C'est un exercice difficile car les annotations sont généralement sur des données très précises et on obtient un système qui n'est pas adaptable sur différents corpus. Néanmoins, ce descripteur pourrait avoir une grande utilité dans les interactions, et des recherches sont en cours pour améliorer ce domaine (He et collab., 2018).

Somasundaran et collab. (2008) opèrent une tâche de détection d'opinions liés dans 2 phrases différentes. Pour cela ils utilisent des descripteurs qui peuvent être utilisés dans le cadre d'interactions tels que le nombre de mots qu'ont en commun les deux énoncés, ou bien le pourcentage du nombre de phrases nominales en commun entre les 2 phrases.

Lai et collab. (2013) utilisent des descripteurs propres à l'interaction dans les réunions du corpus AMI McCowan et collab. (2005) pour une tâche de modélisation de l'affect des participants. Ces descripteurs ne sont pas adaptés a des modèles séquentiels mais sont plutôt représentatifs d'une réunion entière, comme le calcul d'un score représentant l'équité de la participation des locuteurs.

*** :**Modèles

Hazarika et collab. (2018a) utilisent une méthode neuronale basée sur des RNN-

LSTM afin de modéliser l'émotion au sein d'une interaction dyadique multimodale du corpus IEMOCAP. Les auteurs utilisent un mécanisme d'attention sur les différents tours de parole de chaque locuteur, permettant au modèle de donner une importance plus élevée aux tours de paroles passées pertinents pour la prise de décision finale.

Dans un contexte d'interaction Humain-Agent de type magicien d'Oz acté, Langlet et Clavel (2015b) s'intéressent à l'extraction de cibles d'opinions de l'utilisateur. Pour cela les auteurs prennent en compte des paires de tours de parole, l'agent parlant toujours en premier, et extraient les opinions de l'utilisateur selon un système de règles basé sur différentes structures linguistiques à l'intérieur de l'interaction, dépendantes du locuteur.

Un autre exemple d'utilisation de l'interaction et de l'intégration du contexte dialogique dans un modèle est le travail de Ghosh et collab. (2017). Dans cette étude sur la détection de l'ironie dans des discussion en ligne, les auteurs ont utilisé un réseau de neurone récurrent hiérarchique avec auto-attention permettant la prise en compte du tour de parole d'un interactant pour détecter, dans le tour de parole adjacent de son interlocuteur, la présence d'ironie. Cette astuce permet de prendre en compte le contexte d'une paire de tours de paroles (paire adjacente : PA) dans un modèle afin d'améliorer ses performances.

2.4 POSITIONNEMENT AU NIVEAU DES DESCRIPTEURS

Notre système se place dans la catégorie des hybrides, afin de combiner la robustesse des méthodes d'apprentissage et la précision des structures linguistiques. L'utilisation de descripteurs linguistiques extraits de façon experte est de plus en plus remise en question dans la littérature. Les approches nouvelles se basent de plus en plus sur des méthodes statistiques permettant d'obtenir des représentations puissantes car performantes sur de nombreuses tâches et différents domaines. Pour cela, nous utilisons aussi en plus des descripteurs linguistiques classiques des représentations distribuées apprises de manière non supervisé sur une grande quantité de texte.

Nous avons décidé d'ajouter des informations acoustiques grâce à des descripteurs audio recommandés pour l'analyse de sentiments. L'approche choisie pour trouver un ensemble de descripteurs acoustiques (EDA) adapté est motivée par la volonté de se baser sur un EDA simple déjà éprouvé et validé par la communauté, puis de l'enrichir de descripteurs physiquement pertinents. Pour cela, nous avons premièrement utilisé des descripteurs acoustiques classiques provenant du GeMAPS qui est un EDA pour les tâches liées au sentiment conseillé par une équipe de spécialistes de ces différents domaines (Eyben et collab., 2015). Nous avons ensuite enrichi cet EDA avec des paramètres du flux glottal et des MFCC (*Mel-Frequency Cepstral Coefficients*).

Notre modèle graphique permet aussi d'incorporer des descripteurs caractéristiques

de l'interaction et des intégrations non symétriques modélisant un ensemble question-réponse. Nous avons choisi une fusion des données précoce afin de détecter des structures multimodales, puis un modèle multi-vues pour gérer l'asynchronie (voir sous-section 3.4.2).

3

Modèles pour l'analyse d'opinions

Résumé du chapitre

— Les modèles graphiques séquentiels allient faible besoin en données et interprétabilité. Les méthodes neuronales sont efficaces mais nécessitent beaucoup de données. Les méthodes linguistiques sont basées sur des règles expertes utilisant des représentations extraites de manière experte.

— Les méthodes hybrides utilisent des représentations expertes dans des algorithmes d'apprentissage automatique. Les modèles utilisées peuvent être des modèles graphiques ou neuronaux.

— Il existe des méthodes de fusion de données multimodales classiques telles que la fusion précoce ou la fusion tardive, et des méthodes plus perfectionnées comme la fusion hybride qui est effectuée à l'intérieur d'un modèle d'apprentissage, dit multi-vues.

— Les modèles d'opinions proposés dans cette thèse reposent sur des modèles graphiques hybrides avec une fusion des données précoce, et hybride

Dans le chapitre précédent, nous avons présenté l'état de l'art sur les descripteurs acoustiques et textuels utilisés pour des tâches connexes à l'analyse d'opinion comme l'analyse de sentiment ou d'émotion. Comme expliqué dans la figure 1.6, les systèmes classiques basés sur l'AA [1] sont composés d'une extraction de descripteurs (point 2), puis d'un modèle d'apprentissage (point 3). Certaines méthodes neuronales n'ont pas besoin d'effectuer une extraction des descripteurs en amont, au prix d'une grande quantité de données nécessaire à l'apprentissage. Pour ces méthodes dites de bout-en-bout (*ent-to-end*, l'extraction se fait directement à l'intérieur du modèle d'apprentissage. Quant aux méthodes basées sur des règles linguistiques, si elles utilisent les mots sous leurs formes brutes en les associant avec des règles, une extraction de la nature grammaticale et des valeurs de subjectivité des mots peut aussi se faire en amont. Dans ce chapitre, nous présenterons l'état de l'art des différents modèles d'apprentissage pour des tâches liées à l'analyse des opinions. Nous aborderons les modèles à états graphiques qui permettent d'analyser les données de façon séquentielle et interprétable (Section 3.1), puis nous aborderons les méthodes neuronales (Section 3.2) et enfin les modèles de linguistique computationnelle et les méthodes hybrides utilisant des descripteurs linguistiques (Section 3.3). Finalement, la question de la fusion multimodale sera abordée, par rapport à comment elle a été perçue et étudiée dans la recherche liée à l'analyse de la communication humaine (Section 3.4).

3

3.1 MODÈLES À ÉTATS ET MODÈLES GRAPHIQUES SÉQUENTIELS

Les modèles graphiques séquentiels , qu'ils soient à états cachés ou non ont longtemps été utilisés pour travailler sur des séquences du fait de leur capacité à modéliser explicitement les procédés dynamiques.

L'utilisation des Champs Aléatoires Conditionnels (*Conditional Random Fields* : CRF) pour l'analyse de phénomènes liées aux opinions remonte aux travaux de Wöllmer et collab. (2008), sur une tâche de classification d'émotion, via une partition de l'espace Valence-Activation. L'étude a été faite sur le corpus HUMAINE de Douglas-Cowie et collab. (2007) contenant des interactions multimodales dyadiques entre un agent virtuel et un utilisateur. Les descripteurs utilisés pour ce travail sont uniquement acoustiques, construits à partir de fonctionnelles appliquées à des descripteurs bas niveaux, tels que le Pitch, l'Intensité ou le HNR (*HNR* : Harmonics-to-Noise Ratio).

Dans la même veine, les auteurs de Ramirez et collab. (2011) utilisent un modèle de Champs Aléatoires Conditionnels Latents Dynamiques (*Latent-Dynamic Conditional Random Fields* : LDCRF) (Morency et collab., 2007) pour de la classification séquentielle d'émotions, et plus précisément sur les labels : Valence, Activation, Surprise et Dominance. Ce travail a été effectué sur le corpus SEMAINE de McKeown et col-

1. Apprentissage automatique

lab. (2012) contenant des interactions dyadiques multimodales entre un humain et un agent virtuel opéré par un humain et fourni par Schuller et collab. (2011b) dans le cadre du challenge AVEC2011. Le LDCRF permet un apprentissage explicite aussi bien de la sous-structure des signaux affectifs que de la dynamique intrinsèque entre les labels d'émotions.

Un des premiers travaux utilisant des modèles graphiques pour une tâche d'analyse de sentiment de manière multimodale a été Morency et collab. (2011) qui ont utilisé des Modèles de Markov Cachées (*Hidden Markov Model* : HMM) (Rabiner et Juang, 1986) sur un corpus de critiques de films sous forme de Vlog. L'intérêt de cette méthode est de pouvoir modéliser une dynamique caractéristique de la communication humaine. Avec leurs états cachés, les HMM peuvent symboliser par des états latents des variables non observées, comme l'état mental du locuteur. Cependant, l'utilisation d'un modèle génératif restreint le nombre de variables disponibles en entrée. En effet, les HMM étant des modèles génératifs et non discriminatifs (Mccallum et collab., 2000), sont sensibles au bruit apportée par l'utilisation de représentations non utiles à la discrimination d'une classe ou d'une autre, car ils modélisent aussi la distribution des variables observées. Les auteurs utilisent une analyse statistique préalable afin de sélectionner 5 descripteurs au total : 1 textuel, 2 acoustiques et 2 visuels.

Sur une tâche liée à l'analyse d'opinion, qui est la détection d'accord et de désaccord dans des débats télévisés, Bousmalis et collab. (2011) utilisent des champs aléatoires conditionnels cachés (*Hidden Conditional Random Fields* : HCRF) (Quattoni et collab., 2007). Les auteurs utilisent des descripteurs non-verbaux visuels et acoustiques. Cependant, les descripteurs visuels sont des descripteurs des gestes des locuteurs issus d'annotations précises et couteuses, effectuées en amont sur les 43 heures de vidéo du corpus.

Täckström et McDonald (2011) utilisent aussi des HCRF pour une tâche de classification d'opinions dans des critiques de films à l'écrit. Ils utilisent principalement des descripteurs textuels issus de lexiques de subjectivité, ne modélisant pas l'intégralité du contenu verbal comme il est possible avec une représentation en Sac-de-mot. Le type des données diffère de ce à quoi nous aurons affaire dans la suite de cette thèse, car ce sont des critiques écrites et non pas des transcriptions écrites de critiques orales.

Bousmalis et collab. (2013) ont continué le travail présenté plus haut de Bousmalis et collab. (2011) en améliorant le modèle graphique utilisé pour l'analyse du comportement humain. Ils introduisent dans ces articles les HCRF Infinis (*Infinite Hidden Conditional Random Fields* : IHCRF) qui sont une extension des HCRF à l'aide d'un modèle non-paramétrique permettant d'inférer le nombre d'états cachés.

Les HCRF sont aussi utilisés par Zhou et collab. (2013) pour une tâche de détection de troubles mentaux dans une interaction dyadique multimodale entre un humain et un agent virtuel contrôlé par une technique de Magicien d'Oz. L'utilisation de

HCRF combinée avec une segmentation en Paires Adjacentes (PA) permet de modéliser efficacement la dynamique de la conversation entière. La fusion multimodale des données est opérée en concaténant les vecteurs de représentation de chaque modalité ensemble.

Toujours pour une tâche d'analyse du comportement humain similaire à celle étudiée par Bousmalis et collab. (2011), Song et collab. (2012b) présentent un modèle graphique à états cachés multi-vues, permettant de modéliser des données multimodales. Le HCRF multi-vues (*Multi-view Hidden Conditional Random Fields* : MV-HCRF) est une amélioration des HCRF. Basé sur le principe que les différentes modalités n'interagissent pas entre elles à bas niveaux, ce modèle permet de cloisonner les différentes modalités et d'éviter une fusion précoce simple à base de concaténation des vecteurs de représentations des différentes modalités comme ce qui a été fait par Zhou et collab. (2013). Chaque modalité possède ses propres états cachés, fournissant ainsi une discrétisation haut niveau de chaque signal unimodal, qui interagissent ensuite (voir Figure 3.1). Un pré-traitement des données à base d'Analyse de Corrélation Canonique à Noyaux (*Kernel Canonical Correlation Analysis* : KCCA) permet, de la même manière que pour un SVM (*SVM* : Support Vector Machine) à noyaux, de gérer le caractère linéairement local des HCRF, et aussi diminuer la dimension de l'espace des descripteurs.

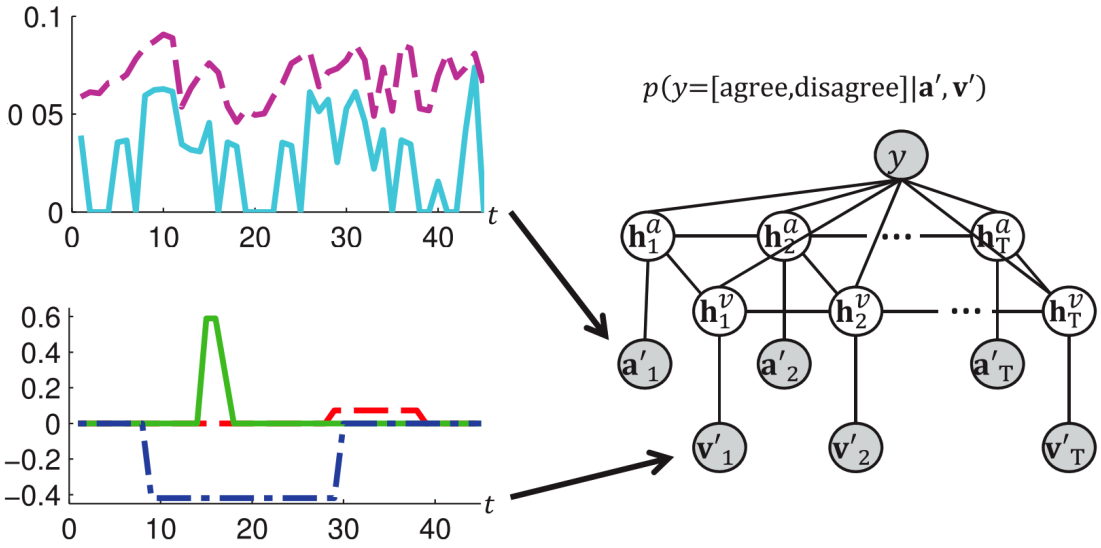


FIGURE 3.1: Modèle de MV-HCRF de Song et collab. (2012a) utilisé pour la détection de (dés)accord. y désigne les labels, t le temps, les $\mathbf{a}_i$ et $\mathbf{v}_i$ désignent les observations des 2 modalités et les $\mathbf{h}_i^v$ et $\mathbf{h}_i^a$ les états cachées associés.

Un hybride de modèle graphique et de réseaux de neurones a été utilisé par Rakicevic et collab. (2017) pour une tâche d'estimation du niveau d'accord dans les interac-

tions multimodales dyadiques de MAHNOB-MIMICRY (Bilakhia et collab., 2015). Les auteurs utilisent des données audio-visuelles : les Unités d'Activation faciales et des descripteurs acoustiques extraient à partir d'OpenSMILE sont fusionnés de manière non-linéaire à l'aide d'un Perceptron multi-couches à fusion hybride, avant d'être utilisés en entrée d'un Champ Aléatoire Conditionnel Ordonné (Kim et Pavlovic, 2010).

Zhang et collab. (2015) utilisent eux aussi un modèle hybride de Champ Aléatoire Neuronal (*Conditional Neural Field* : CNF) multi-tâche sur la base de données de critiques en ligne de Mitchell et collab. (2013). La première tâche est la détection de la valence de l'opinion contenue dans la phrase, la deuxième est l'extraction de l'entité nommée.

3.2 MÉTHODES NEURONALES

Depuis les dernières années, de plus en plus de BD (Bases de Données) pour l'analyse des phénomènes liées aux opinions sont disponibles. Simultanément la puissance computationnelle augmente. Les méthodes neuronales, gourmandes en données, sont celles qui se sont imposées depuis peu.

Pour les systèmes traitant uniquement le texte, (Sous-section 3.2.1) on peut distinguer 2 tendances. Soit une annotation fine des données, avec extraction de l'aspect sur lequel est porté l'opinion, ou alors une annotation au niveau du document avec des modèles complexes apprenant plusieurs tâches connexes à la fois. Les modèles à annotation fine nécessitent beaucoup de données annotées finement et permettent d'obtenir à la fois le sentiment du locuteur et la cible associée. Les seconds modèles utilisent les données disponibles sur le web afin de pré-entrainer des modèles généraux multi-tâches comme vu dans la section 2.1 pour les *word embeddings*. Ces modèles sont ensuite raffinées (*fine-tunés*, second apprentissage permettant un réglage plus fin des paramètres du modèle) sur la BD à utiliser. Ceci permet d'effectuer un apprentissage par transfert, utiliser des modèles plus complexes et obtenir de meilleurs résultats sur des BD de petites tailles.

Pour les systèmes traitant des données multimodales (Sous-section 3.2.2), on reste généralement sur des modèles mono-tâches qui s'appliquent à modéliser de manière intelligente la multimodalité afin de mieux faire interagir les différentes modalités entre elles. Ces systèmes utilisent le fait que les différentes vues contiennent des informations soit complémentaires, soit consensuelles ou soit erronées (voir Section 3.4). Le principe de ces systèmes dits multi-vues est de cloisonner les différentes modalités à bas niveau pour ensuite faire interagir des états ou représentations de plus haut niveau entre elles.

3.2.1 Les modèles neuronaux textuels

Le stockage des données de manière écrite a précédé de centaines de siècles le stockage des données de manière audio-visuelle. De même, il est plus simple à stocker sous forme numérique, le texte étant mois coûteux en mémoire de stockage. Ainsi, on a toujours trouvé plus de données disponibles sous la forme textuelle que sous la forme audio-visuelle et l'analyse de sentiment automatique s'est développée sur du texte dans sa genèse. Aussi de nos jours, les grandes quantités nécessaires aux modèles d'apprentissage neuronaux permettant une bonne généralisation se trouvent en plus grande partie sous forme textuelle. Dans cette sous-section, nous décrirons les modèles neuronaux utilisés sur du texte. On trouve des modèles nécessitant de grandes quantités d'annotations fines (Socher et collab., 2013), ou permettant une annotation fine grâce a une clusterisation intrinsèque (Angelidis et Lapata, 2017), certains basés sur un apprentissage non-supervisé de la langue (Howard et Ruder, 2018) et enfin d'autres jouant sur la connexité et la complémentarité de différentes tâches afin d'apprendre un modèle roi maîtrisant le langage (Mccann et collab., 2018).

Pour une tâche de classification de sentiment dans des phrases, Socher et collab. (2013) introduit le Réseau de Tenseur Neuronal Récursif (*Recursive Neural Tensor Network* : RNTN). À l'aide de l'arbre de dépendance de la phrase, le RNTN permet de trouver une représentation pour chacun des nœuds de l'arbre en combinant les représentations des noeuds-fils de l'arborescence. Le désavantage de cette approche est la nécessité d'annotation pour chacun des nœuds de l'arbre de dépendance. Un exemple illustrant la granularité fine et contraignante de ces annotations est disponible en figure 3.2. Ces annotations nécessaires à l'apprentissage sont disponibles dans la base de données SST (*Stanford Sentiment Treebank* : SST) introduite dans le même article.

Une autre application de NN (*NN* : Neural Network) utilisant l'arbre de dépendance de la phrase récursif est visible dans Irsoy et Cardie (2014) où les auteurs utilisent aussi la SST. La particularité de ce modèle est l'accumulation de couches, rendant le réseau profond par rapport à la structure de l'arbre et par rapport aux couches accumulées, et l'utilisation des mots-vecteurs word2vec pré-entrainés.

Irsoy et Cardie (2013) utilisent des Réseaux de Neurones Récursifs Bidirectionnels pour une tâche d'extraction d'Expression Subjective Expressive et d'Expression Subjective Directe, ainsi que les porteurs d'opinions et les cibles, sur le corpus MPQA Wiebe et collab. (2005). Cette tâche nécessite aussi beaucoup de données annotées à la main.

Angelidis et Lapata (2017) ont construit une amélioration du réseau de neurones hiérarchique avec attention de Yang et collab. (2016). Ce réseau à la capacité de labelliser au niveau des observations de la séquence, suite à un apprentissage supervisé au niveau d'un label global à la séquence entière. Les auteurs testent leur modèle sur des données *Yelp* et *IMDb*, obtenant un score F1 de 63% sur la partie IMDb, surpassant le

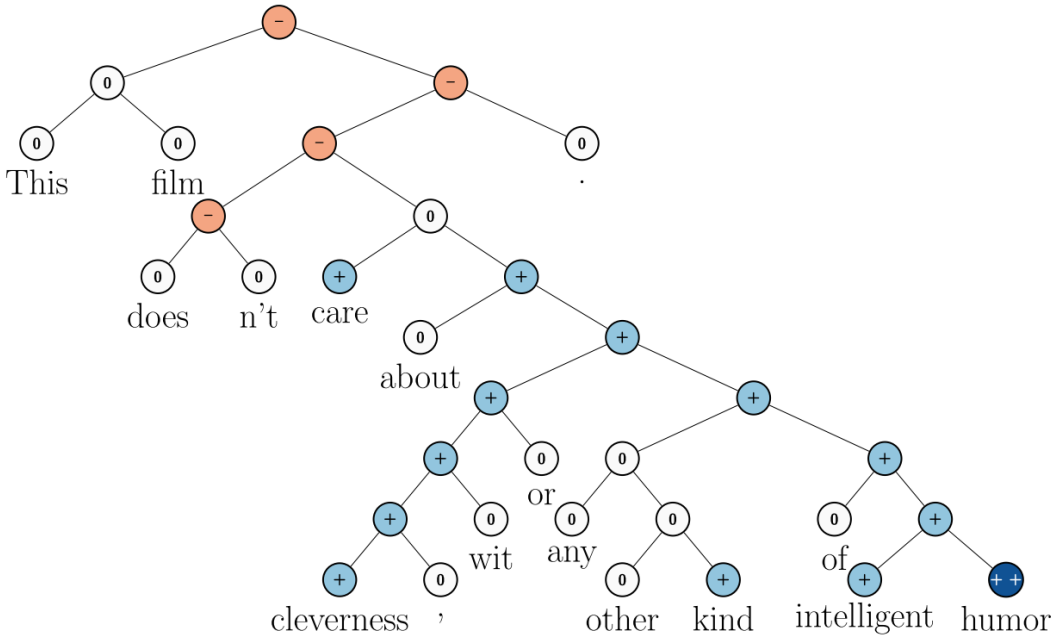


FIGURE 3.2: Exemples d'annotation du SST de Socher et collab. (2013)

SO-CAL de Taboada et collab. (2011).

Récemment, Howard et Ruder (2018) ont présenté un réseau de neurones permettant d'apprendre un modèle de langage de manière non supervisée sur un corpus classique comme wikipédia[2] avant d'apprendre une tâche particulière sur des données, disponibles puis de faire un réglage fin des paramètres (*fine-tuning*) sur le corpus cible pour la même tâche. Ce pré-entrainement permet d'obtenir de très bons résultats avec très peu d'exemples annotés sur le corpus cible, et d'obtenir de meilleurs résultats que l'état de l'art.

Radford et Salimans (2018) utilisent la même approche que Howard et Ruder (2018), basant leur modèle d'apprentissage sur le Transformer de Vaswani et collab. (2017). Ces modèles sont testés sur des bases de données telles que la SST et obtiennent des résultats à l'état de l'art.

Finalement, certains modèles misent sur des approches multi-tâches, se basant sur le fait que l'apprentissage joint de beaucoup de tâches complémentaires sont nécessaires à la création d'un modèle qui maitrise plusieurs aspects du langage, bien loin du modèle de langage basique prédisant le prochain mot. Cette idée a été mise en place il y a une dizaine d'année par Collobert et Weston (2008) qui utilisaient des mots-vecteurs avec un réseau de neurones et 6 types de tâches différentes. La même philosophie est reprise plus récemment par Mccann et collab. (2018) où les auteurs proposent un en-

2. https ://en.wikipedia.org

semble de 10 tâches à résoudre afin de créer un benchmark pour les modèles textuelles universels. Ils présentent le Décathlon du Langage Naturel (*Natural Language Decathlon* : decaNLP) et proposent le *Multitask Question Answering Network* (MQAN). Ce modèle résout les différentes tâches sous forme de *Question Answering*, et est basé sur des mécanismes de co-attention et d'auto-attention. Pour la tâche d'analyse de sentiment, la question est "*Is this sentence positive or negative?*" et le modèle répond en langage naturel "*positive*" ou "*negative*".

3.2.2 Les modèles neuronaux multimodaux

La majorité des modèles neuronaux pour l'analyse de sentiments sont des modèles très récents datant des deux dernières années. Ils nécessitent une plus grande quantité de données, même s'ils ont prouvé leur supériorité sur des bases de données de Vlogs de plus faibles tailles (Zadeh et collab., 2018a,b,c). La plupart de ces modèles sont présentés plus en amont dans ce manuscrit lorsque l'on aborde la fusion multimodale, et il nous paraissait important de les mentionner ici aussi. Le lecteur est renvoyé en partie 3.4 afin d'avoir plus de détails sur la majorité des modèles.

On ajoutera un modèle multimodal prenant en compte les interactions qui a été proposé récemment par Hazarika et collab. (2018a). Ce modèle permet de modéliser une interaction dyadique de manière séquentielle afin de déceler les émotions dans chaque tour de parole. Les auteurs utilisent des RNN (*RNN* : Recurrent Neural Network) de type GRU (*GRU* : Gated Recurrent Unit) Chung et collab. (2015) spécifiques à chaque interlocuteur et qui permettent de créer des vecteurs de mémoire sur l'historique de la conversation. Ces représentations de l'historique interagissent avec une représentation du tour de parole courant afin d'inférer la classe d'émotion. À notre connaissance, c'est l'unique modèle neuronal existant utilisant les interactions orales dans une dyade pour une tâche de reconnaissance d'un phénomène lié aux opinions.

3.3 LINGUISTIQUE COMPUTATIONNELLE ET MODÈLES HYBRIDES

La linguistique est un domaine de recherche qui a précédé historiquement l'informatique, les modèles de linguistique computationnelle ont aussi joué un rôle important dans les analyses des données textuelles. Ces modèles utilisent des indices linguistiques, lexicaux et syntaxiques dans des systèmes basés sur un ensemble de règles afin de détecter des phénomènes précis. Ces systèmes sont généralement construits suite à l'étude d'une base de données de petite taille, tout en restant des systèmes généraux qui ne sur-apprennent pas certaines règles trop spécifiques. Malheureusement, la création des règles doit être faite à la main, et ce sont des systèmes très peu robustes et sensibles au bruit.

Les systèmes hybrides quant à eux, sont des modèles d'apprentissage qui utilisent des représentations issues de la linguistique. Ces méthodes permettent d'allier la précision des lexiques à la robustesse et au pouvoir de généralisation des modèles statistiques.

3.3.1 Les modèles à base de règles

On sait que les mots peuvent avoir plusieurs significations selon le contexte d'utilisation, il en est de même pour le sentiment associé et il est intéressant d'étudier la polarité contextuelle dans une phrase. Wilson et collab. (2005) introduisent une nouvelle méthode pour détecter la polarité au niveau phrastique. Certains mots peuvent avoir plusieurs valeurs de sentiments (comme plusieurs significations), et en utilisant le contexte dans lequel chaque mot apparait, on peut modifier les valeurs en sentiment apportés par chaque mot pour obtenir une valeur globale contextuelle plus satisfaisante.

Kennedy et Inkpen (2006) est le premier à utiliser des inverseurs de valences de manière contextuelle pour une tâche d'analyse de sentiment dans des critiques de film. Dans Polanyi et Zaenen (2006), les auteurs vont plus loin et présentent une étude sur la modification de la valence en fonction du contexte dans lequel les mots sont employés. Ils affirment que la valence de l'attitude d'un objet lexical est modifiée par le contexte lexical et discursif. Les objets lexicaux interagissent les uns avec les autres. Ce ne sont pas forcement des adjectifs avec des valences fortes, inverseurs de valences ou modificateurs de degré pris de manière indépendante mais la composition de toutes les valences associées ensemble. Par exemple l'utilisation d'un adjectif négatif avec un nom positif ou l'inverse, comme dans "*a beautiful disaster*" ou "*an awful hero*".

Hutto et Gilbert (2014) présente un système de règles lexico-syntaxiques appelé VADER (*Valence Aware Dictionary for sEntiment Reasoning*) pour analyser les sentiments dans des tweets. Ce système consiste à utiliser un lexique de mots avec une valeur en sentiment associée, puis d'utiliser 5 règles générales incorporant des conventions syntaxiques et grammaticales pour l'expression, ou l'emphase de l'intensité, d'un sentiment. Les règles de VADER fonctionnent très bien sur des tweets car elles sont adaptées pour ce type de données en particulier (casse, répétition de lettres ou ponctuation,...).

Dans la même veine mais plus général, Taboada et collab. (2011) utilisent un ensemble de lexiques et de règles afin de calculer un score caractéristique de ce que les auteurs appellent l'orientation sémantique de document. Ce score est calculé localement à base de règles et de valeurs provenant de lexiques, de subjectivités, de négations et de modificateurs d'intensité. Les auteurs utilisent ce système sur le corpus de textes MPQA de Wiebe et collab. (2005).

Allant plus loin que le simple contexte phrastique et dans une interaction orale

humain-agent, Langlet et Clavel (2016) analysent les attitudes de l'utilisateur via un système composé de règles. Ce travail permet de prendre en compte un premier niveau de contexte dialogique en utilisant à la fois le tour de parole de l'utilisateur mais aussi le tour de parole adjacent précédent de l'agent (Figure 3.3).

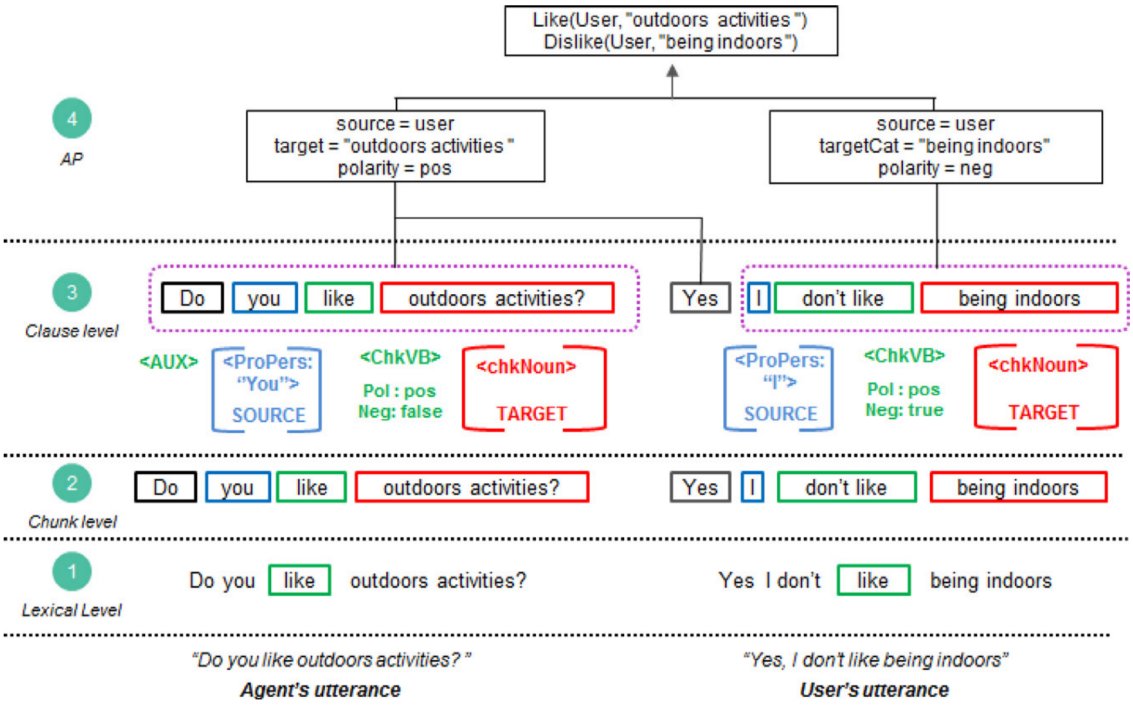


FIGURE 3.3: Exemple du modèle à règles de Langlet et Clavel (2016)

3.3.2 Les modèles hybrides

Les modèles hybrides ont été beaucoup utilisés pour des tâches précises et proches de la linguistique comme l'extraction d'opinion, de source et de cible proposés par le corpus MPQA (Choi et collab., 2005, 2006; Yang et Cardie, 2012, 2013). Sur ce genre de tâche difficile, les systèmes hybrides peuvent bénéficier des indices linguistiques utilisés comme descripteurs qui permettent d'injecter de la connaissance humaine dès l'initialisation du modèle d'apprentissage, pour atteindre le niveau de complexité demandé.

Breck et collab. (2007) utilisent des CRF pour retrouver des Expressions Subjectives Directes/Expressives (*Direct/Expressive Subjective Expression* : DSE/ESE) dans le corpus de textes MPQA. Pour la représentation du signal textuel, on note l'utilisation de plusieurs descripteurs linguistiques comme les hyperonymes de WordNet Miller et col-

lab. (1990), le groupe dans lequel se situe les verbes et noms dans l'ontologie Framenet (Baker et collab., 1998) et des valeurs provenant d'un lexique de subjectivité (Wilson et collab., 2005). Aujourd'hui, l'utilisation de FrameNet ou WordNet peut être remplacée par des mots-vecteurs. En effet, les relations de proximité sémantique entre les mots comme une relation d'hyperonymie ou d'homonymie est implicitement incluse dans le positionnement spatial des mots-vecteurs.

Pour la même tâche, Yang et Cardie (2013) et Choi et collab. (2006) proposent des méthodes d'extraction de l'opinion, des sources et cibles associées dans le corpus MPQA de Wilson et collab. (2005). Pour le système de Yang et Cardie (2013), après avoir identifié l'expression d'opinion avec un CRF, un autre modèle est utilisé pour trouver les possibles arguments de l'opinion (source ou cible). Des descripteurs linguistiques comme des lexiques de subjectivités, le groupe des verbes dans FrameNet, la distance en terme de nombre de mots et en terme de chemin syntaxique entre l'argument et l'expression sont utilisées.

Mitchell et collab. (2013) introduisent un modèle d'analyse de sentiment ciblée. Le modèle s'attaque à une tâche de détection, dans une phrase, du sentiment et de la cible associée à ce sentiment, sur un corpus de critiques sous forme de texte. Les auteurs utilisent des CRF et de nombreux descripteurs linguistiques pour apprendre de manière jointe le sentiment et sa cible associée sur des corpus de tweets en anglais et en espagnol.

Les modèles hybrides peuvent aussi être des modèles neuronaux. Shin et collab. (2016) utilisent aussi des lexiques de subjectivité comme l'*Opinion Lexicon* de Hu et Liu (2004) et le *NRC Sentiment140 Lexicon* de Kiritchenko et collab. (2014) dans des réseaux de neurones convolutifs. Les auteurs testent leur système sur des tâches d'analyse de sentiments dans des critiques de films écrites et dans des tweets. Les lexiques ne couvrent pas plus de 10% des mots du corpus de critique de films : un modèle neuronal de mots-vecteurs est aussi utilisé pour représenter les mots en plus des lexiques qui perdent beaucoup d'information. Pour finir, l'utilisation de CNN implique une quantité de données d'apprentissage conséquente.

Présenté plus haut (sous-section 3.2.2), le RTN de Sahay et collab. (2018) utilise des descripteurs, comme Shin et collab. (2016), avec des valeurs issus de lexiques et aussi des scores provenant des règles lexico-syntaxiques de Hutto et Gilbert (2014), sur des critiques de film Vlogs. Ces descripteurs sont ensuite placés en entrée d'un réseau de neurones permettant la fusion des modalités.

Wang et collab. (2016) s'attaquent au même type de tâche sur un corpus de critiques Yelp de restaurants et d'ordinateurs sous forme de texte. Les auteurs utilisent un modèle de CRF neuronaux récursifs qui sont constitués d'un réseau de neurones récursif basé sur l'arbre de dépendances syntaxiques de la phrase (*Dependency-Tree Recursive Neural Network* : DT-RNN)), puis un modèle de CRF. Le DT-RNN permet d'ob-

tenir des représentations haut-niveau de chaque mot qui sont utilisés dans comme entrée du CRF qui permet d'extraire l'opinion et l'aspect associé. Le DT-RNN a l'avantage d'être un modèle de bout-en-bout mais de pouvoir aussi utiliser des descripteurs linguistiques en concaténation de la représentation du DT-RNN pour améliorer ses performances (comme la nature grammaticale du mot et des lexiques de subjectivités).

L'ontologie SenticNet de Cambria et collab. (2014) permet de retrouver "*l'information conceptuelle et affective associé à une expression du langage naturelle composée de plusieurs mots*". SenticNet a aussi été beaucoup utilisé pour récupérer des structures et les utiliser dans un modèle hybride. Poria et collab. (2016) utilise les concepts SenticNet de Cambria et Hussain (2015) (exemple en Figure 3.4) pour une tâche d'analyse de sentiments multimodale sur le corpus de critique Vlogs de Morency et collab. (2011).

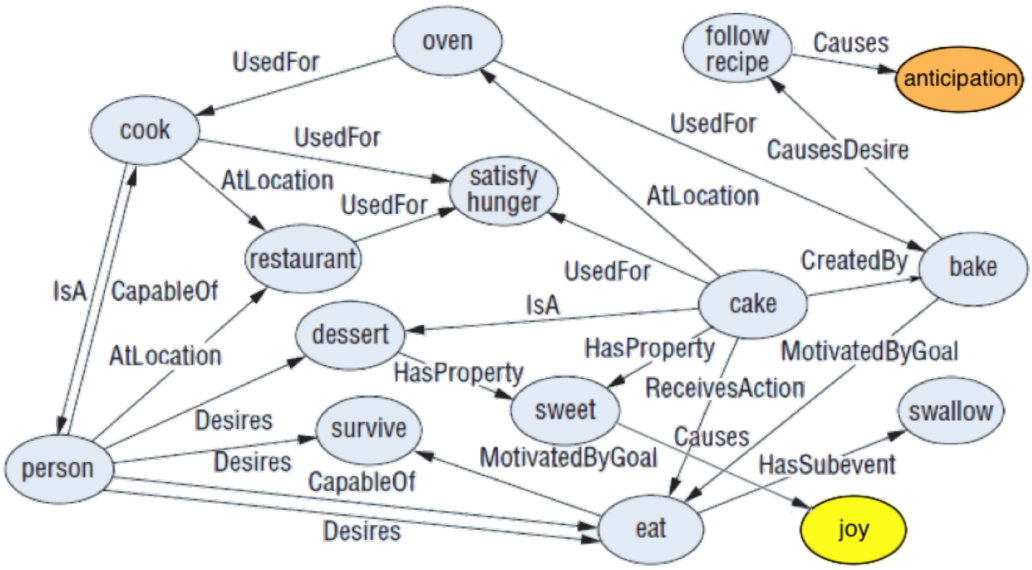


FIGURE 3.4: Modèle de Cambria et collab. (2014) pour le concept `cake`

Il existe finalement des approches utilisant les graphes de connaissances intégrés dans des réseaux de neurones grâce à des méthodes comme les travaux de Kumar et collab. (2018); Ma et collab. (2018). Kumar et collab. (2018) intègre le réseau WordNet dans un réseau de neurones en utilisant une méthode d'*embedding* de graphe de connaissance pour une tâche d'analyse de sentiment sur des micro-posts. Cette technique permet de représenter des relations entre différentes entités comme par exemple le synset de WordNet "(*bronze age, part of, prehistory*)".

Ma et collab. (2018) proposent le Sentic LSTM qui est un LSTM permettant d'intégrer la base de connaissances générale SenticNet dans un LSTM. Basé sur l'idée que

l'information au niveau du concept est complémentaire à l'information au niveau du mot. Les auteurs ajoutent une nouvelle porte dans l'architecture du réseau LSTM. L'utilisation des concepts contenus dans SenticNet pour chaque mot améliore les résultats sur une tâche d'analyse de sentiment ciblés dans des critiques en ligne. Un travail similaire est aussi présenté par Cambria et collab., avec une validation via une tâche d'analyse de sentiment binaire sur des phrases provenant de critiques textuelles.

3.4 FUSION DES DONNÉES MULTIMODALES ET SYSTÈMES MULTI-VUES

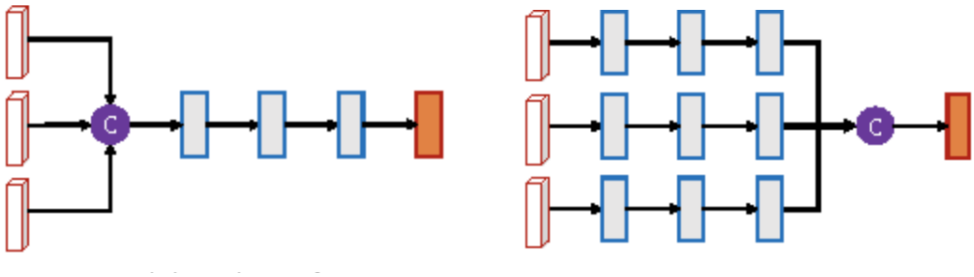
La fusion des données est une partie cruciale dans le traitement des données multimodales. Les manifestations de l'opinion dans les modalités ayant des structures de longueurs différentes et centrées sur des instants différents, il est parfois difficile pour le modèle d'apprentissage d'arriver à assembler des données venant de capteurs différents pour en sortir une information pertinente. Jusqu'à peu de temps auparavant, la majorité des systèmes pratiquaient des fusions classiques en concaténant les représentations des modalités avant l'entrée du classifieur ou alors fusionnaient les résultats de classifieurs unimodaux de manière plus ou moins complexe (Sous-section 3.4.1). De plus en plus de systèmes dit multi-vues ont été utilisés (Sous-section 3.4.2) afin de cloisonner les modalités au niveau des descripteurs bruts puis de les faire interagir à plus haut niveau.

3.4.1 Fusion classique

Cette partie abordera la manière dont on fusionne l'information : fusion tardive ou précoce, les différentes granularités des représentations et leurs modèles associés qui permettent d'avoir des résultats pour le domaine auquel on s'intéresse. La manière de fusionner l'information étant étroitement liée à la segmentation et finalement la manière de prendre en compte la dynamique, on abordera rapidement les études essayant de modéliser une dynamique d'opinion.

La fusion multimodale est une étape importante dans le processus de création d'un algorithme. Cette étape doit être liée avec le modèle, les données et la tâche. Un des obstacles les plus importants est de trouver une segmentation des données adéquate qui soit adaptée aussi bien aux descripteurs qu'au modèle. Le texte et l'audio sont généralement fortement corrélés et il est pourtant difficile de les aligner lorsqu'on se penche sur des descripteurs acceptant les mêmes intégrations temporelles. Cela peut mener à différents types de fusion : la fusion tardive et la fusion précoce. [3] Un exemple

3. *decision-level fusion* et *feature-level fusion*



(a) Fusion Précoce

(b) Fusion tardive

FIGURE 3.5: Exemples de fusions multimodales. Les boites rouges et blanches représentent les vecteurs des différentes modalités, les C représentent une concaténation de ces vecteurs, les boites bleues sont des représentations vectorielles (comme une couche cachée dans un NN) et les boites rouges et orange les vecteurs de sortie.

est disponible en Figure 3.5. On notera qu'il existe aussi la fusion qui permet de fusionner différentes représentations à différents niveaux Atrey et collab. (2010). Cette approche est efficace avec des réseaux de neurones qui permettent différents niveaux d'abstraction selon la profondeur.

La fusion de descripteurs (précoce) combine les caractéristiques extraites de chaque modalité en un unique vecteur, généralement la concaténation de chaque vecteur unimodal. Cet unique vecteur est ensuite utilisé comme entrée d'un classifieur durant les phases d'apprentissage et de tests. L'idée principale est de faire des liens entre les différentes modalités durant l'entraînement afin de détecter des structures multimodales intéressantes qu'il aurait été impossible de repérer sans cela. L'objectif final étant d'avoir un système commettant moins d'erreurs.

La fusion tardive s'attaque au problème en entraînant un classifieur par modalité et fusionne les décisions indépendamment obtenues par chacun des classifieurs. Avec cette méthode, l'extraction d'informations et de structures jointes temporellement entre les modalités est bien plus ardue. Les résultats unimodaux sont combinés à l'aide de métriques adaptées telles que des règles expertes faites à la main, des paramètres obtenus à la suite d'un apprentissage ou bien de simples décisions comme le vote majoritaire ou la moyenne arithmétique.

Le sens commun voudrait que l'on effectue une fusion précoce : animé par l'hypothèse que le modèle est adapté au problème, les structures multimodales émergeront statistiquement de l'apprentissage, comme il a été fait par Poria et collab. (2015). Néanmoins, pour une tâche de détection de persuasion sur un corpus multimodal, Nojavanasghari et collab. (2016) montrent qu'une approche par fusion tardive naïve surpasse une approche par fusion précoce. La fusion précoce augmente la dimensionnalité du vecteur d'entrée avec des descripteurs qui ne sont plus nécessairement im-

portants pour la prédiction et diminue les performances, provoquant un bruit indésirable à l'apprentissage. Paleari et Huet (2008) ont mené des expériences sur le corpus eNTERFACE pour une tâche de reconnaissance d'émotions avec fusions tardive et précoce sur des données audiovisuelles. Les résultats obtenus sont meilleurs avec la fusion tardive, simplement en moyennant les intervalles de confiance. **La configuration optimale pour une fusion d'information efficace est encore un problème ouvert** et un modèle discriminatif est utile, afin d'utiliser une fusion précoce et analyser les résultats.

Une difficulté principale de la fusion précoce est la synchronisation des différentes modalités en une représentation jointe cohérente. Par exemple, utiliser une représentation en Sac de mots pour représenter un long document en un vecteur donne de bons résultats pour une tâche d'analyse de sentiments (Maas et collab., 2011). Cependant, un vecteur représentant tout l'audio pour un document de plusieurs minutes n'est pas le meilleur des choix : le Pitch ayant bien moins de sens physique intégré sur un intervalle temporel aussi grand car on perd toute l'information sur la dynamique. Il est plus efficace d'exploiter l'aspect séquentiel du signal audio pour améliorer le modèle.

Dans certaines de nos expériences, nous avons choisi d'expérimenter différentes segmentations afin d'avoir des unités d'intégration qui peuvent contenir des structures importantes pour les différentes modalités. Utiliser les pauses orales pour segmenter est une technique déjà reconnue et utilisée par la communauté pour l'audio (Wöllmer et collab., 2013b) alors qu'une segmentation par pauses du texte est moins commune. Perez-Rosas et collab. (2013b) utilisent les pauses plus longues que 0,5 secondes pour segmenter les vidéos de 30 secondes du corpus MOOD en unités inter-pausales (UIP) et utilisent un Sac de mots pour une classification en sentiment au niveau des énonciations. De son côté, Poria et collab. (2015) préfèrent apprendre des représentations à l'aide d'un plongement de mots en entrée d'un CNN et obtenir des nouvelles représentations de dimension faible pour chacune des énonciations sur le même corpus.

3.4.2 Système multi-vues

Les systèmes multi-vues sont les systèmes qui permettent de faire ce qui est appelé une fusion hybride (Nojavanasghari et collab., 2016). La fusion est faite au sein du modèle d'apprentissage et permet de faire interagir les modalités à différents niveaux. La fusion tardive interdit toute détection de structures multimodales, car elle ne fait qu'utiliser les scores de probabilités des classifieurs. La fusion précoce concatène les modalités avant l'entrée du classifieur, ce qui fait interagir les modalités à bas niveau. Le problème de cette fusion est l'absence de prise en compte du **caractère asynchrone et la différence de longueur caractéristique des structures caractéristiques des différentes modalités**. En effet, si le locuteur énonce la phrase "*I love that movie!*", la

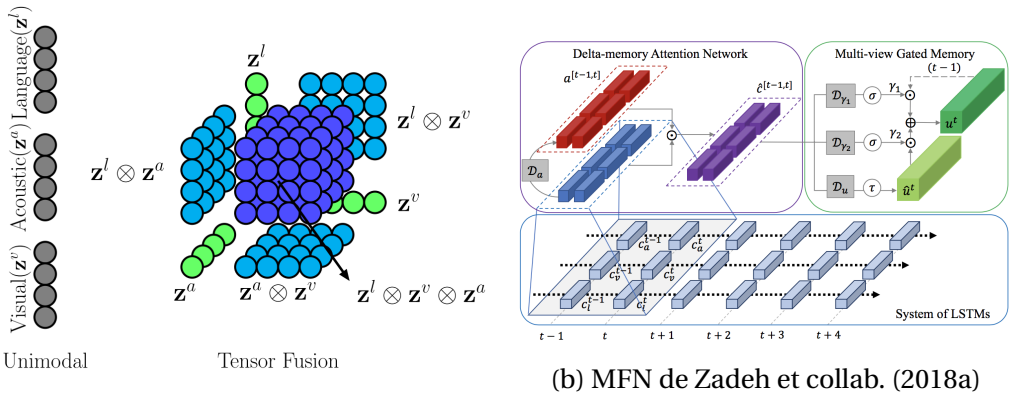
structure linguistique spécifique au phénomène d'opinion sera la phrase entière, alors que la structure vocale (l'intonation) ne sera peut être qu'une augmentation du pitch sur le "o" (en phonétique : [u :]) de "*movie*". La fusion multimodale hybride présente l'avantage de pouvoir cloisonner les modalités à bas niveaux afin de les faire interagir à plus haut niveau. Ainsi, une modalité qui est dans une condition particulière à un instant donné peut avoir un impact sur les autres modalités sur une échelle de temps différente, et finalement changer la décision finale.

Les modèles des travaux suivants sont utilisés pour l'étude de la communication humaine sur des tâches d'analyse de sentiment. Ils méritent d'être mentionnés, car s'ils ne sont pas spécifiques à l'interaction, ils s'intéressent de manière profonde à la fusion multimodale à l'aide de réseaux de neurones profonds :

— Rajagopalan et collab. (2016) étendent les modèles neuronaux séquentiels à l'aide d'artifices permettant de gérer les différentes modalités d'un signal. Ils proposent un LSTM multi-vues (*Multi-View Long-Short-Term-Memory* : MV-LSTM) qui permet de séparer les modalités au sein même de la couche cachée en faisant interagir leurs représentations respectives avec des degrés différents. Ce modèle est utilisé pour une tâche de détection d'engagement dans des interactions multimodales dyadiques chez l'enfant.
— Le *Tensor Fusion Network* de Zadeh et collab. (2017) fusionne les représentations des 3 modalités, obtenues après un traitement neuronal, via une suite de produits dyadiques puis tensoriels avant d'appliquer un CNN sur la matrice cube (voir Figure 3.6a)
— Une amélioration directe du TFN présenté plus haut est le *Relational Tensor Network* (RTN) de Sahay et collab. (2018). Motivé par le travail de Poria et collab. (2017b) qui fait passer les représentations de chaque observation (correspondant à une énonciation du locuteur) dans un Bi-LSTM afin des les imprégner du contexte de la vidéo entière, les auteurs utilisent le contexte de la vidéo entière pour tagger chaque énonciation. Le RTN fait passer une séquence d'observations unimodales dans un Bi-LSTM afin d'obtenir une séquence de représentations unimodales qui sont fusionnées via un produit tensoriel, avant de passer dans un CNN.
— Chen et collab. (2017) : introduisent le concept de la *Gated Multimodal Embedding* qui permet l'occlusion d'une modalité, lorsqu'elle est bruitée, via une couche d'attention. Par exemple quand il y a du bruit visuel ou auditif alors le système apprend a ne pas choisir cette modalité.
— Le *Mult-Attention Recurrent Network* de Zadeh et collab. (2018c) permet de séparer les modalités en attribuant des LSTM Hybrides spécifiques à chaque modalité qui permettent au système de stocker à la fois les dynamiques unimo-

dales et la dynamique inter-modale (possiblement asynchrone) importante pour chaque modalité. La représentation multimodale est calculée à l'aide d'un bloc multi-attention.

— Le *Memory Fusion Network* de Zadeh et collab. (2018a) utilise toutes les modalités pour calculer l'attention sur chaque neurone de toutes les modalités. Ceci permet aussi d'utiliser la dynamique inter-modale dans la création de l'attention. (voir Figure 3.6b)



(a) TFN de Zadeh et collab. (2017)

(b) MFN de Zadeh et collab. (2018a)

FIGURE 3.6: Exemples de fusions multimodales

3.5 POSITIONNEMENT

Les structures linguistiques sont coûteuses à créer et elles sont d'une grande rigidité. À l'heure actuelle, les modèles linguistiques à règles deviennent extrêmement minoritaires. Avec l'arrivée d'une nouvelle puissance computationnelle et des algorithmes permettant l'avènement de modèles complexes ayant des milliards de paramètres, la majorité des travaux se tournent vers des modèles statistiques. Cependant lorsque la tâche est atypique comme la nôtre à l'oral, et que la quantité de données disponible n'est pas suffisante, il est toujours possible d'utiliser des descripteurs linguistiques du signal textuel en entrée d'un modèle d'apprentissage statistique. Les approches neuronales sont très bonnes pour modéliser les interactions entre les modalités et les fusionner en une représentation efficace, mais elles nécessitent beaucoup de données.

Les modèles de champs aléatoires conditionnels cachés (HCRF; plus de détails sous-section 7.1.1) contenant des états cachés sont bien adaptés à un certain nombre de problèmes. Celui de la modélisation de signaux multimodaux pour la reconnaissance d'opinion peut être envisagé du fait du caractère séquentiel de la parole et de la compositionnalité des opinions. Les CRF sont bons pour structurer parmi les labels,

mais ils ont besoin de beaucoup de types de labels différents, ou de labels précis quand c'est une analyse ciblée. Avec l'utilisation du HCRF, on arrive à trouver ces "labels latents" grâce aux états cachés du HCRF. Quattoni et collab. (2007) l'ont fait avec succès en apprenant de manière discriminatoire une distribution d'états parmi les différentes classes de gestes, leur permettant non seulement de découvrir les configurations distinctives qui identifient de manière unique chaque classe, mais aussi d'apprendre une structure commune partagée entre les classes. Les approches hybrides utilisant des ontologies comme FrameNet ou SenticNet permettent de mettre des concepts sur des groupes de mots. Ceci n'est plus jugé si intéressant qu'auparavant grâce à l'utilisation de représentation distribuées de mots-vecteurs appris statistiquement. Avec ces représentations, les concepts sont intrinsèquement inclus dans la position spatiale du vecteur.

L'analyse des opinions se faisant surtout dans le texte, nous travaillons sur une tâche peu commune avec un phénomène complexe (voir chapitre 1) où la quantité de données disponibles est faible, car nécessitant des interactions orales annotées en opinion. C'est pour cela que nous nous concentrons sur des méthodes hybrides permettant apprentissage et utilisation de descripteurs linguistiques. De plus, nous cherchons à modéliser des signaux multimodaux séquentiels, tout en ayant un système interprétable ce qui est possible avec l'analyse des états latents des HCRF. Sachant aussi qu'il est souhaitable pour le système d'être efficace sans avoir besoin d'un nombre trop important d'exemples, en tant que modèle discriminatif, les HCRF pourraient être plus adaptés à nos petits ensembles de données qu'un modèle génératif. C'est pour ces raisons que nous utilisons des HCRF comme classifieurs à l'aide d'une fusion au niveau des descripteurs, puis la version multi-vues pour séparer les modalités.

4

Bases de données pour l'apprentissage

Résumé du chapitre

— Il existe de nombreuses bases de données pour l'étude de la communication multimodale : des bases de données de type Vlog et des bases de données en interaction. Les annotations sont principalement en émotions pour les bases de données en interaction.

— Les deux bases de données utilisées dans cette thèse sont : le corpus VLog de critiques de films ICT-MMMO annoté en sentiment au niveau de la vidéo et le corpus d'interaction humain-agent SEMAINE annoté en attitude au niveau de la paire de tours de parole.

— Nous avons choisi de faire annoter la base de données entière SEMAINE en opinion

Le travail de cette thèse se base sur des algorithmes d'apprentissage automatique (AA) qui nécessitent des bases de données afin de pouvoir repérer des structures dans les données. Les structures à détecter sont des expressions d'opinion dans une interaction humaine orale. Cette section de l'état de l'art évoquera les différents corpus utilisables dans notre domaine de recherche qui contiennent des interactions. Plus particulièrement nous nous attarderons, dans le cadre de cette thèse, sur les bases de données répondant à des critères spécifiques. Nous souhaitons travailler avec des bases de données : en anglais (pour être cohérent entre nos différents travaux), avec des enregistrements audio, des transcriptions du texte, les timecodes (codes temporels) associés à chaque mot, contenant des interactions dyadiques, les plus spontanées possibles.

Après une présentation et un état de l'art des différentes bases de données audio disponibles à ce jour qui favorisent les émissions d'opinions (Section 4.1), nous présenterons plus en détails les bases de données que nous avons utilisées au cours de ce travail de thèse (Section 4.2). Nous finirons ensuite par une réflexion sur les limitations des bases de données proposées par la communauté pour une tâche d'analyse d'opinions en interaction suivie d'un positionnement sur le sujet (Section 4.3).

4.1 LES BASES DE DONNÉES AUDIO POUR L'APPRENTISSAGE DE MODÈLES D'ANALYSE D'OPINION

On compte de plus en plus de bases de données multimodales pour l'apprentissage automatique et l'analyse des phénomènes liées aux opinions. Il faut cependant noter que jusqu'à il y a peu de temps, une grande partie des études qui étaient effectuées sur ces bases de données étaient restreintes à une analyse des émotions des locuteurs plutôt que de l'analyse de leurs sentiments ou de leurs opinions. Ces derniers aspects étant plutôt étudiés sur des données textuelles à la syntaxe plus claire du fait de la complexité de l'annotation du phénomène et de la volatilité de la structure linguistique dans le langage oral.

Dans ses débuts, l'analyse de sentiments fut étudiée majoritairement sur des données textuelles. La prédiction de notes de critiques de films, qui sont des données contenant de nombreuses opinions, a longtemps été un exercice classique de l'analyse de sentiment. Nous commencerons notre état de l'art des bases de données multimodales de critiques en ligne (Sous-section 4.1.1), car c'est une BD de ce type que nous avons utilisé lors de la première expérience de cette thèse.

Cependant ces bases de données de critiques ont l'inconvénient majeur de ne pas contenir d'interactions. Les données sont des monologues d'un locuteur en face de la caméra. Dans ce contexte, la manière d'exprimer l'opinion est encore différente. Finalement, lorsque l'on souhaite avoir des interactions qui soient spontanées, il est difficile que celles-ci soient riches en opinions (Busso et collab., 2017). On parlera des

différents corpus multimodaux contenant des interactions qui constituent les bases d'apprentissages disponibles de notre champ d'étude (Sous-section 4.1.2).

4.1.1 Les bases de données d'opinions de critiques Vlogs

L'expansion de l'analyse de sentiment s'est réalisé au début des années 2000 avec l'analyse de critiques de films ou de produits que l'on pouvait désormais trouver à grande échelle sur le net. Des sites comme IMDb ont même uniquement la vocation à n'être qu'une grande base de données faite de manière collaborative. Les travaux de Turney (2002); Pang et Lee (2004) s'attardent sur l'analyse de critiques de film en construisant des systèmes permettant de retrouver la note que le critique a donné sur une échelle de Likert (1932). L'analyse de sentiments sur les critiques est un large domaine qui profite grandement aux systèmes de recommandations Chelliah et Sarkar (2017). Les critiques sont des bases de données qui ont une densité élevée en opinions, pour cela, elles sont des candidates idéales pour l'utilisation de modèles d'apprentissage pour l'analyse d'opinions. La sous-section suivante portera sur les bases de données contenant des interactions et favorisant les expressions d'opinions.

Depuis ***Youtube Movie Reviews*** de Morency et collab. (2011) et la création du premier corpus de Vlogs pour l'analyse de sentiments, de plus en plus de chercheurs ont compris l'intérêt d'utiliser ce genre de données. Ces bases de données ont l'avantage d'être exploitées directement depuis le web en téléchargeant les vidéos. Cette technique permet d'obtenir de grandes quantités de données provenant d'utilisateurs variés et d'avoir des qualités d'enregistrements très différentes, cette diversité représente ainsi d'une manière plus réelle les données que l'on peut être amené à trouver dans la réalité (dénommées "*in-the-wild*") : grand nombre de locuteurs avec des manières différentes de s'exprimer.

FIGURE 4.1: Exemples de videos provenant de la base de données de Morency et collab. (2011)

Depuis le *Youtube Movie Reviews* qui faisait moins de 50 vidéos, le nombre de vi-

déos des corpus de Vlog qui ont été produits augmente de plus en plus. Wöllmer et collab. (2013b) ont collectés le corpus **ICT-MMMO** (*'Institute of Creative Technologies - Multi-Modal Movie Opinion* : ICT-MMMO) de 370 vidéos de critiques de film annotées avec la note du critique, allant de 1 à 5.

D'autres auteurs ne se sont pas arrêtés à utiliser la simple note du locuteur mais sont allés plus loin en faisant annoter des énonciations de la vidéo de manière indépendante. Ainsi, les auteurs du corpus **CMU-MOSI** (*Carnegie Mellon University - Multimodal Opinion-level Sentiment Intensity* : CMU-MOSI) (Zadeh et collab., 2016) ont segmenté chaque vidéo et ont fait annoter indépendamment des énonciations plus courtes. Une segmentation fine de la sorte permet d'éviter des segments longs à annoter contenant plusieurs opinions et permet ainsi d'obtenir un accord inter-annotateur plus élevé. Le corpus CMU-MOSI comprend 2199 segments de vidéos annotés en sentiment de -3 à 3. Les mêmes auteurs ont depuis publié une version améliorée, le **CMU-MOSEI** (*Carnegie Mellon University - Multimodal Opinion Sentiment and Emotion Intensity* : CMU-MOSEI) Zadeh et collab. (2018b), qui contient 23453 segments provenant de 3228 videos, chacun annoté de la même manière que son prédécesseur. Ce corpus a été utilisé lors d'un Grand Challenge du 1er *Workshop Human Multimodal Language* à ACL (*ACL* : Association for Computational Linguistics) en 2018 (Zadeh et collab., 2018d). Une version espagnole de base de données de critiques Vlog est aussi disponible avec le **MOUD** (*Multimodal Opinion Utterances Dataset* : MOUD) de Perez-Rosas et collab. (2013a), composé de 56 vidéos segmentées en 498 énonciations de manière manuelle et annotées à l'aide de 3 étiquettes : positive, négative ou neutre.

Les avantages inhérents de ces bases de données ont fait que leur utilisation n'en a pas été restreinte uniquement à des tâches d'analyse de sentiments ou d'opinions. Le ***Youtube Personality dataset*** de Biel et Gatica-Perez (2013) est une collection de 404 Vlogs où les locuteurs s'expriment sur de nombreux sujets tels que la politique, la littérature ou le cinéma On compte des annotations de descripteurs du comportements, de transcriptions et d'annotation de personnalités (à l'aide du *Big Five*). Le **POM** (*Persuasive Opinion Multimedia* : POM) de Park et collab. (2014) est un corpus de 1000 vidéos récoltés sur ExpoTV.com qui est un site populaire hébergeant des vidéos de critiques de produits, et qui a été utilisé pour étudier le pouvoir de persuasion d'un point de vue multimodal. Ce corpus a été annoté par Garcia (2017) pour une étude d'analyse de sentiments ciblée au niveau des différents aspects notés dans la critique en plus de la note globale.

4.1.2 Les bases de données d'opinions en interactions

On compte de nombreux corpus qui ont été créés pour l'étude des interactions Humain-Humain et Humain-Agents, la majorité de ces corpus ne provenant pas du

web sont multimodaux afin de pouvoir étudier en profondeur les signaux sociaux. Les interactions entre différentes personnes sont soumises à un ensemble de paramètres variés qu'il faut prendre en compte lorsque l'on crée une base de données pour les étudier. La relation entre les différents interlocuteurs (degré de familiarité, rôle social dans l'interaction), le nombre d'interlocuteurs (deux ou plus), le type de tâche à effectuer ou encore la présence ou non d'un agent conversationnel. Le degré de familiarité entre les différents interlocuteurs (et la relation qui les lie en général) est très important. Par exemple lorsque 2 personnes se connaissent les informations échangées peuvent être plus courtes et restent comprises de chaque côté car les individus partagent un socle commun d'informations contextuelles. Dans le corpus **NoXi** de Cafaro et collab. (2017) (voir Figure 4.2), on introduit un autre biais en donnant un rôle d'expert ou de novice à chacun des membres de la dyade. Ceci donne un rôle social à chacun d'eux qui oriente et biaise l'interaction (l'expert parlant plus que le novice par exemple).

Un point important est le type d'interaction, qu'elle soit dyadique (deux personnes) ou polyadique (plus de deux). La manière de modéliser l'interaction est très différente selon le type de l'interaction étudiée. Le corpus AMI de McCowan et collab. (2005) est une base de données composée d'enregistrements de groupes de personnes ayant des réunions et qui bougent dans la pièce, alors que dans le corpus **MAHNOB-mimicry** (Bilakhia et collab., 2015) se trouvent uniquement des interactions dyadiques avec les sujets assis en face-à-face, ce qui est préférable pour notre sujet d'étude. Cependant peu d'opinions sont inclues dans MAHNOB-mimicry, de plus les tours de paroles sont généralement tres longs (plusieurs minutes), rendant l'interaction quelque peu artificielle car non dynamique.

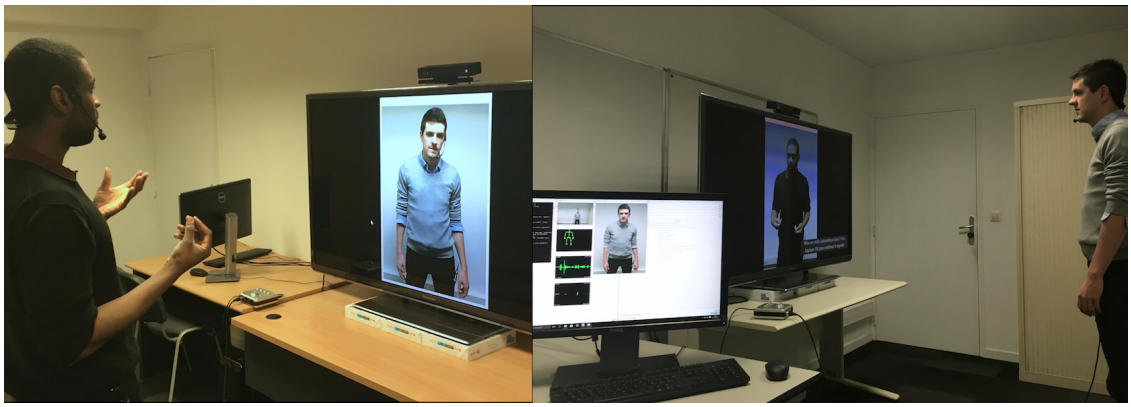


FIGURE 4.2: Photographie d'une interaction lors de la collecte des données de la base de données NoXi de Cafaro et collab. (2017)

Il est parfois difficile de trouver tout le panel des sentiments auquel l'humain est confronté dans la vraie vie, au sein d'un même corpus créé en laboratoire et particulièrement lorsqu'il s'agit des émotions. Afin de créer ces configurations particulières,

les interactions des corpus de laboratoires peuvent être spontanées ou actées. Dans certains corpus pour l'analyse visuelle d'émotion les auteurs demandaient aux participants de dire une phrase avec différentes intonations, ou bien utilisait des acteurs pour jouer des situations fortes en sentiments. Le corpus **IEMOCAP** de Busso et collab. (2008) contient une moitié de sessions actées et une autre moitié de sessions spontanées, toutes jouées par des acteurs professionnels. De même, **MSP-Improv** de Busso et collab. (2017) est un corpus d'interaction dyadiques multimodales. Les auteurs de ce corpus on fait le choix d'utiliser des données actées au détriment de données spontanées afin d'obtenir un corpus riche en opinion. D'après les auteurs de la base de données, beaucoup de questions de recherche intéressantes nécessitent des conditions contrôlées particulières. Ainsi, les interactions de MSP-Improv sont actées pour reproduire des situations qui ne se produisent que très peu dans la vie courante.

Le corpus **RECOLA** de Ringeval et collab. (2013) est un corpus d'interactions dyadiques multimodales qui correspond tout à fait à nos demandes, les participants devant effectuer la *Winter Task Survival* de Hall et Watson (1970), cependant ce corpus est en français et non en anglais.

Un peu à part, nous citerons aussi **MSP-Podcast** de Lotfian et Busso (2017) qui présente un corpus audio d'enregistrements de radio annotés en émotions, contenant 18 238 phrases annotées tirées de 920 enregistrements pour un total de 27h. Cette base de données est d'après ses auteurs, la plus grande base de données avec des interactions émotionnelles, avec des émotions naturelles, un grand nombre de locuteurs différents reflétant la diversité de la réalité et des données balancées. Néanmoins, cette base de données ne contient pas d'interaction dyadiques, n'a pas de transcriptions du texte qu'il serait difficile d'obtenir de manière automatique.

Enfin, il est intéressant pour notre problématique de bien discerner les cas Humain-Agent et Humain-Humain. Les corpus Humains-Agents sont plus rares que les corpus Humains-Humains et lorsqu'ils sont créés, des artifices sont utilisés la plupart du temps. Le corpus **SEMAINE** (McKeown et collab., 2010) est un corpus d'interactions dyadiques entre un humain-utilisateur et un humain-agent jouant un rôle prédéfini selon la technique du Magicien d'Oz. Un acteur joue le rôle de l'agent et ne peut dire que quelques phrases prédéfinis. Les agents joués par les acteurs sont au nombre de 4, et chacun avec une forte personnalité destinée à éveiller des réactions affectives chez l'utilisateur (plus de détails Sous-section 4.2.2). Le corpus d'interaction humain-agent **ChIMP** de Narayanan et Potamianos (2002) suit la même technique du Magicien d'Oz que précédemment pour une interaction avec des enfants afin de mieux comprendre la manière dont ces derniers interagissent avec des machines, cependant c'est un corpus textuel uniquement.

4.1.3 Résumé des BDD existantes

Le lecteur trouvera un résumé des différentes bases de données existantes dans le tableau 4.1. **Au vu des différents avantages qu'il a, le corpus SEMAINE est donc adapté à nos besoin**. C'est un corpus d'interactions dyadiques multimodales. Il n'est pas totalement acté, car les utilisateurs sont tout à fait spontanés et si les opérateurs jouent un personnage, ils ont de la marge dans leur créativité, permettant à des émotions artificielles mais proches d'émotions naturelles d'émerger Busso et collab. (2017). De plus, cette particularité du Magicien D'Oz permet au corpus d'être dense en phénomènes liées à l'opinion et aux émotions, car l'agent permet de déclencher ces phénomènes chez l'utilisateur qui réagit de manière totalement spontanée.

4.2 LES BASES DE DONNÉES UTILISÉES DANS NOS ÉTUDES

4.2.1 *ICT-MMMO* : Corpus de critiques Vlog

Le corpus ICT-MMMO sur lequel nous avons travaillé est une base de données vidéo de critiques de films provenant des sites *Youtube.com* et *ExpoTV.com* (Wöllmer et collab., 2013b). Le corpus contient 365 vidéos, variant de 1 à 3 minutes, pour un total de plus de 14h45 d'enregistrement. Les vidéos ont été faites par des critiques non professionnels et la qualité audio des enregistrements varie de manière significative. Des captures d'écran de locuteurs ayant des opinions diverses sont disponibles en Figure 4.3.

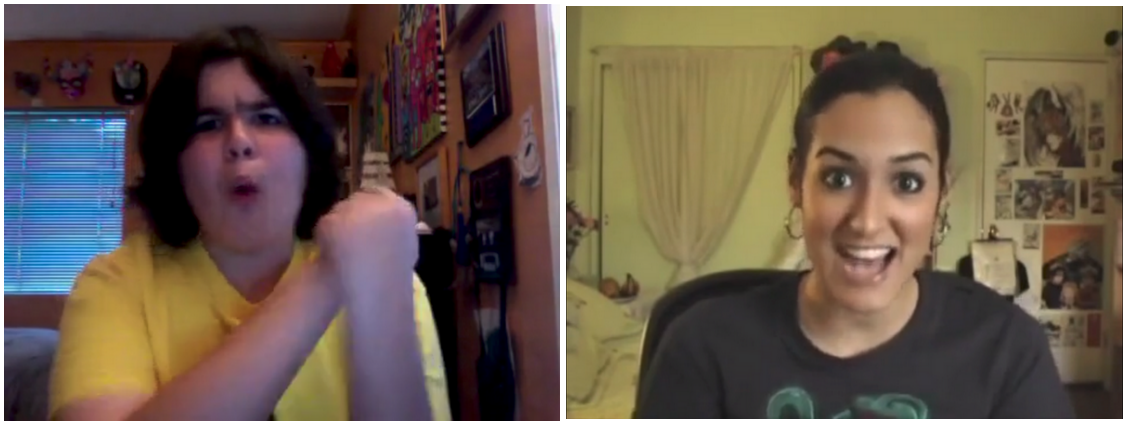
4.2.1.1 Transcriptions manuelles et annotations

Toutes les vidéos du corpus ont été annotées par un ou deux annotateurs en terme de valence globale. Les scores de valence sont sur une échelle allant de 1, qui signifie une opinion très négative de la part du locuteur, à 5, pour une opinion très positive, en passant par 3 pour une opinion neutre. L'étiquette "Je ne sais pas" est aussi une annotation valable et a été utilisée par un des deux annotateurs. Le consensus inter-annotateur est obtenu en prenant la moyenne arithmétique des scores disponibles. Le corpus contient 120 vidéos négatives, 38 neutres et 207 positives, ce qui donne un total de 365 vidéos.

Chaque vidéo est transcrite à la main à l'aide du logiciel Transcriber de (Barras et collab., 2001). Chaque pause de plus de 150ms est aussi transcrite, avec les temps de début et fin associés : ceci permet d'avoir la transcription de la vidéo en ayant en plus les codes temporels de chaque Unité Inter-Pausale (UIP). De plus, des annotations para-linguistiques sont présentes et formalisées selon un schéma précis (voir Tableau 4.2).

TABLE 4.1: Corpus d'interactions multimodaux en anglais

Corpus	Temps de parole	Nombre d'observations	Points +	Points -	Remarques
ATAROS Freeman et collab. (2014)	36.4h (**34** interactions of ~64'), + 5h d'entretiens et d'autres conversations	11489 énoncés labellisés : 2993+ / 4519 ~ / 517 - / 3445 NA	Interactions dyadiques dense en prise de positions (débats collaboratifs) ; Transcriptions et alignement des mots au niveau des pauses > 0.5s	Labels NA : trous dans l'annotation de la conversation.	
MAHNOB-MIMICRY Bilakhia et collab. (2015)	54 sessions, 11h40 (entre 5 et 20mn, ~ 12mn)		Interactions dyadiques naturelles favorisant l'opinion ; Participants engagés soit dans une discussion socio-politique (34) soit négociant une offre de location (20)	Pas de transcription. Tours de paroles de plusieurs minutes	Annotations : Actes de dialogue, prise de parole, affects, gestes de tête et de mains, mouvements du corps et expressions faciales, (des)accord
SEMAINE McKeown et collab. (2010)	24 sessions de 4 discussions dont 1 inutilisable : 95 discussions dont 79 transcrites.	Environ 6000 tours de paroles	Configuration favorise l'expression d'opinion. 79 Transcriptions	Agents actés ; Utilisateurs spontanés	95 discussions annotées en Valence et Activation continument
IEMOCAP Busso et collab. (2008)	10h de parole	10 039 tours (5255 scriptés ; 4784 spontanés). En moyenne 4.5 sec pour 11.4 mots	Interactions dyadiques ; Configuration favorise l'émotions, transcription au niveau des mots/syllabes/phonèmes	Acté ; une émotion par session. Ce n'est pas clair lorsque l'interaction est spontanée ou non (Bilakhia et collab., 2015)	Annotation de Valence, Activation, Dominance pour chaque tour de parole
MSP-Improv Busso et collab. (2017)	Plus de 9 heures	8,438 tours de paroles ; 80 interactions	Interactions dyadiques actées et spontanées	Pas de transcriptions ; Acté	Annotations d'Activation, Valence et Dominance pour chaque tour de parole
MSP-Podcast Lotfian et Busso (2017)	Podcasts radio, 27 heures de données émotionnelles	18 282 phrases	Plus grand corpus avec des interactions orales qui existe	Pas de transcriptions	
Noxi Cafaro et collab. (2017)	11h13mn de discussions	40 sessions	Interactions dyadiques naturelles	Pas de transcriptions ; Beaucoup de non natifs	Disposition Novice-Expert
Negociations (Gratch) Park et collab. (2012)	Humain-Humain : 42 sessions d'environ 12 minutes ; 8h24	10319 phrases (7840 uniques) avec 122.8 TP par dialogues ; 79396 mots (2516 uniques) avec 7.7) mots par TP	Interactions dyadiques	Pas de transcriptions	Peu d'information dans les articles



(a) Un instantané d'un utilisateur emettant une critique négative sur le film

(b) Un instantané d'un utilisateur emettant une critique positive sur le film

FIGURE 4.3: Instantanés de 2 vidéos de la BD

TABLE 4.2: Quelques annotations para-linguistiques du corpus ICT-MMMO

Para-linguistique	**Exemple**
Élongation	*I li : :ked it*
Mot incomplet	*jona-*
Mot accentués	*ex_a_c_tly*
Intonation montante	question
Intonation descendante	fin utterance

Des explications complémentaires sur la base de données ICT-MMMO que nous avons utilisée sont disponibles dans la section 5.1.

4.2.2 SEMAINE

SEMAINE est la seconde base de données que nous avons utilisée dans ce travail de thèse. Le corpus de discussions ouvertes entre un opérateur jouant le rôle d'un agent virtuel et un utilisateur. Ce sont des interactions dyadiques multimodales, avec des transcriptions fournies. Nous allons présenter la base de données dans cette sous-section ainsi que les annotations d'attitudes collectées par Langlet et Clavel (2015b) et que nous avons utilisée pour un de nos travaux.

4.2.2.1 Base de données

La base de données **SEMAINE** a été réalisée par McKeown et collab. (2012, 2010) à la *Queen's University of Belfast*. SEMAINE est une grande base de données audio-visuelles créée dans le cadre d'une approche itérative visant à l'amélioration d'agents

conversationnels animés affectifs (*Sensible Artificial Listener* : SAL). Ces agents sont capables d'engager une personne dans une conversation soutenue et colorée sur le plan émotionnel. Les données utilisées pour construire les agents proviennent de la configuration SAL-solide, constituées d'interactions entre des utilisateurs et un "opérateur" simulant un agent SAL, via une pseudo-technique de Magicien d'Oz car jouant le rôle. D'autres configuration sont disponibles, comme les *semi-SAL* et *SAL-automatique*, mais dans cette thèse nous avons uniquement utilisé le sous-ensemble de SEMAINE appelé *SAL-solide* où un opérateur humain joue le rôle d'un agent virtuel.

Les différents personnages agents ont des personnalités bien définies, qui ont été choisies spécialement pour déclencher des réactions liées aux émotions chez les utilisateurs. On compte **Poppy**, joyeuse et extravertie, **Prudence**, pragmatique et dans la mesure, **Spike**, colérique et conflictuel et **Obadiah**, dépressif et morose. Les phrases de l'agent sont contraintes par un script (cependant, certaines déviations au script se produisent dans la base de données) visant à mettre l'utilisateur dans le même état que celui du personnage joué par acteur.

Une session classique dure environ vingt minutes avec une durée d'interaction approximative de cinq minutes par personnage. La durée réelle des interactions varie en fonction des individus. Les participants sont priés de demander un changement de personnage quand ils s'ennuient, s'énervent ou pensent qu'ils n'ont plus rien à dire à l'agent. L'opérateur peut également demander de changer de personnage si suffisamment de temps s'est écoulé avec une personnalité ou si une conversation arrive a une conclusion naturelle. L'opérateur est tenu de jouer chacun des quatre personnalités à son tour, mais sans ordre particulier (généralement l'utilisateur décide).

Les enregistrements de SEMAINE sont effectués dans une configuration particulière. Les participants ont été recrutés parmi les étudiants de la *Queen's University of Belfast* et l'équipe de recherche des auteurs. L'interaction se déroule à distance, l'opérateur et l'utilisateur étant dans différentes pièces. Le média utilisé pour la communication entre les participants est un écran avec des haut-parleurs associés (montré en Figure 4.4)

Comme dit plus haut, SEMAINE convient à nos besoins : c'est un corpus d'interactions dyadiques multimodales. Le cadre particulier de la collecte des données de SEMAINE en fait un corpus semi-acté, même si la mise en place de la situation est artificielle. En effet, les utilisateurs sont spontanés et même si les opérateurs jouent un personnage, ils ont de la marge dans leur créativité, permettant à des émotions artificielles mais proches d'émotions naturelles d'émerger (Busso et collab., 2017). Finalement, l'utilisation de pseudo-agent ayant un rôle particulier permet au corpus d'être dense en phénomènes liées à l'opinion et aux émotions, car l'agent permet de déclencher ces phénomènes chez l'utilisateur qui réagit de manière totalement spontanée.

FIGURE 4.4: Conditions d'enregistrement de SEMAINE de McKeown et collab. (2012). La salle de l'utilisateur est à gauche et la salle de l'opérateur est à droite

4.2.2.2 Annotations disponibles

En ce qui concerne les annotations, SEMAINE est annoté de manière continue sur 4 axes : Valence, Activation, Attente et Puissance (Wöllmer et collab., 2013a). Des exemples de valeurs fortes et faibles sur chacun de ces axes sont illustrés en figure 4.5. Des annotations sur des parties de la BD et collectées de diverses manières ont été fournies par (Langlet et Clavel, 2015b). Cependant, des annotations d'opinions ne sont pas encore disponibles sur cette base de données entière. C'est pour cela que nous allons introduire dans la sous-section suivante les annotations en attitude par paire de tour de parole que nous avons utilisées pour créer le premier corpus *SEMAINE-Léger* qui a servi de terrain d'étude dans un des travaux menés au cours de cette thèse.

Nous savons que la base de données SEMAINE contient le type de données que nous souhaiterions utiliser ; des interactions dyadiques multimodales semi-spontanées, riches en opinions. Il n'existe pas d'annotations disponibles pour la tâche qui nous intéresse sur le corpus entier. Pour pallier ce manque, nous avons choisi d'utiliser des annotations disponibles de Langlet et Clavel (2014) et de les fusionner en un ensemble cohérent.

Dans un premier temps, afin d'étudier les opinions au sein des interactions, nous avons utilisé des annotations qui ont été faites lors de précédents travaux de membres de notre équipe de recherche (plus de précisions sont disponibles dans le chapitre 5). Nous avons utilisé les annotations faites par les auteurs de Langlet et Clavel (2014) et de Langlet et Clavel (2015b). Ceci constitue deux sous-ensembles d'annotations différents. Le schéma d'annotation de Langlet et Clavel (2014) est donné en Figure 4.6.

Nous expliciterons en détails les manières d'obtentions et les différences de ces deux sous-ensembles, ainsi que la manière de les fusionner dans le chapitre 5.

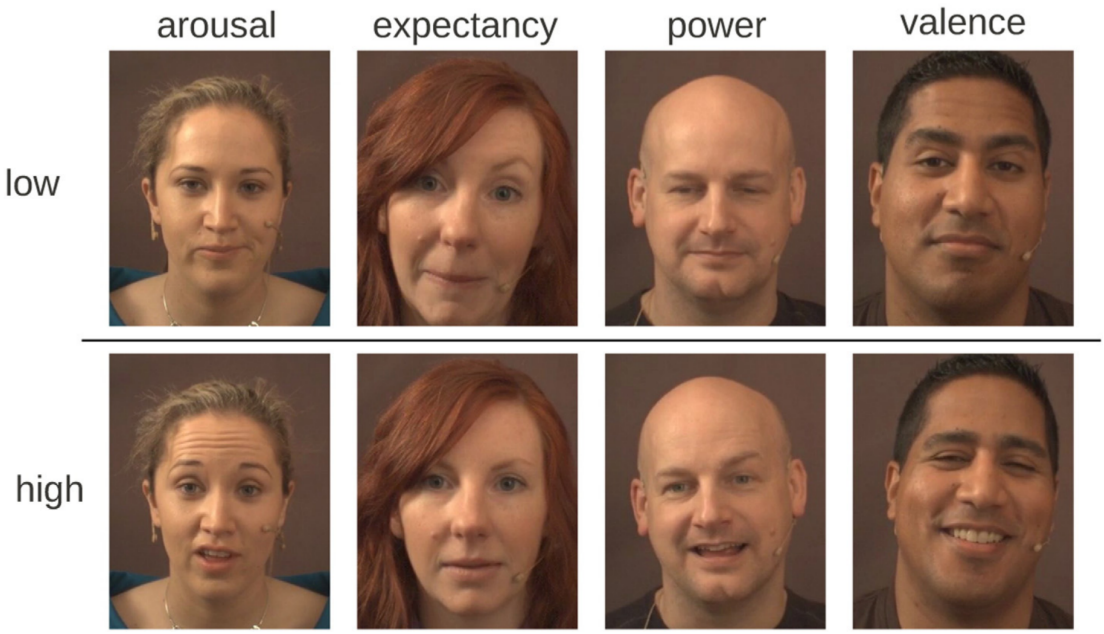


FIGURE 4.5: Exemples de valeurs faibles et fortes d'Activation (*Arousal*), Surprise (*Expectancy*), Dominance (*Power*) et Valence de Wöllmer et collab. (2013a)

4.3 LIMITATIONS ET POSITIONNEMENT PAR RAPPORT À L'ÉTAT DE L'ART

Les annotations existantes de SEMAINE nous ont permis de créer un ensemble de Paires Adjacentes annotées par rapport aux attitudes de l'utilisateur. Cependant, ces deux bases de données annotées sont limitées par plusieurs facteurs. Pour cela nous proposons comme alternative de faire annoter nous-même le corpus SEMAINE.

4.3.1 Limitations quant à la tâche proposée

Malgré les avantages énoncés ci-dessus de ICT-MMMO et de *SEMAINE-Attitude*, ces corpus sont limités, chacun par des points différents, par rapport à la tâche que nous souhaitons effectuer.

La base de données ICT-MMMO :

— **Ne contient pas d'interaction**. Ce corpus est constitué uniquement de monologues, bien qu'il soit utile pour une première étude, nous sommes restreint car le travail de cette thèse s'intéresse à une analyse des opinions dans les interactions orales, qui est un contexte bien différent.

— **Ne contient pas d'alignement de l'audio en mot par mot**. Seul les codes temporels des pauses sont disponibles. Il est donc difficile d'étudier une autre segmen-

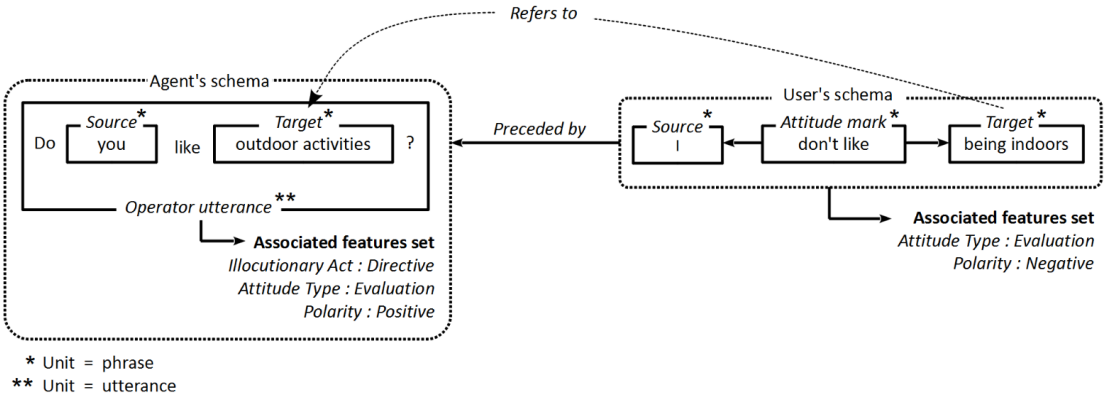


FIGURE 4.6: Exemple d'annotation à l'aide du schéma utilisé par Langlet (2018). *Judgments* et *appreciations* sont regroupés en *evaluation*

tation qui pourrait permettre d'utiliser des structures linguistiques plus fines, la segmentation rendant notre système "à la merci" des pauses du locuteurs.

— **Contient des annotations à gros grains**. Chaque vidéo contient beaucoup d'opinions mais seule l'annotation au niveau de la vidéo est disponible.

— **Taille limitée**. Le manque d'exemples ne permet pas d'obtenir une significativité dans les résultats lorsque l'on compare notre système à un autre.

La base de données SEMAINE-Léger :

— Contient des **PA annotées de manière indépendante**. Il n'y a pas d'annotation faites dans le contexte entier de la conversation, rendant impossible l'utilisation de la discussion entière avec un modèle d'étiquetage séquentiel.

— **Taille limitée**. Le manque d'exemples ne permet pas d'obtenir une significativité des résultats lorsque l'on compare notre système à un autre.

— **Des annotations sont sur des données fictives**. Suite à une création des PA via des méta-données produites par un ASR, une partie des PA créées et annotées sur SEMAINE-T ne sont pas des tours de paroles réels du corpus SEMAINE. [1]

— **L'annotation n'est pas homogène**. SEMAINE-D et SEMAINE-T n'ont pas été annotés selon la même méthodologie, étant des corpus de développement et de test d'un système linguistique. Par conséquent, le nombre d'annotateurs et les plateformes diffèrent, rendant la création d'une vérité terrain plus complexe.

1. Ceci n'est pas handicapant pour Langlet et Clavel (2016) qui n'utilisent que le texte et ont fait annoter ces PA factices, mais rédhibitoire avec une utilisation de l'audio.

4.3.2 Positionnement

L'analyse d'opinion est encore de nos jours une tâche surtout restreinte à des monologues ou des articles et non à des interactions (Poria et collab., 2017a), ce qui restreint beaucoup la disponibilité des bases de données. Au vu de l'état de l'art des bases de données utilisables pour l'analyse des opinions, il apparait impossible d'utiliser un corpus déjà existant et annoté pour étudier les opinions dans des interactions orales humain-humain ou humain-agent. Ainsi, nous avons décidé de nous positionner en faisant annoter une base de données déjà existante.

La base de données que nous avons choisi est SEMAINE car le corpus contient des interactions orales dyadiques dans une configuration qui favorise les expressions d'opinions. La distinction principale que nous souhaitons faire avec les base de données existantes est un corpus d'interaction qui soit annoté en opinions par tour de parole et ce de manière dépendante. Cela signifie que contrairement à SEMAINE-Attitude où l'annotation des PA est faite de manière indépendante, les annotateurs annoteront une conversation entière. **Cette caractéristique permet de modéliser de manière plus efficace la dynamique de l'opinion inter-locuteurs dans une discussion complète** par :

I. Une utilisation des aspects séquentiel et collaboratif de l'interaction orale ;
II. Une analyse des opinions plus contextuelle qui profite pleinement de l'historique conversationnel

Deuxième partie

Méthodologie

5 Mise en place des bases de données SEMAINE-Léger et SEMAINE-Opinions

Résumé du chapitre

— Nous avons effectué une normalisation et un alignement audio des transcriptions du corpus SEMAINE.

— Un schéma d'annotation et une plateforme web mis en place lors de cette thèse ont été utilisés pour la collecte d'annotations en opinion par tour de parole sur 79 discussions du corpus SEMAINE.

— Nous avons obtenus un corpus de 6h20 riche en opinion (52% des tour de parole) avec un accord inter-annotateur (α de Krippendorff) de 66.3

Dans cette partie nous présentons les bases de données sur lesquelles nous avons travaillé afin de récupérer des annotations, que ce soit à partir d'annotations existantes ou bien en faisant nous-mêmes annoter la base de données. La base de données impliquée dans la création de ces deux sous-corpus est SEMAINE de McKeown et collab. (2012). La première section (Section 5.1) est consacrée à la description du travail effectué pour créer une annotation d'attitude sur des paires de tours de parole du corpus SEMAINE, appelé SEMAINE-Léger. Ce corpus est celui qui a été utilisé dans les chapitres 9 et 10. La deuxième section (Section 5.2) est consacrée au nettoyage et à l'annotation en opinion par tour de parole de 79 discussions du corpus SEMAINE, afin d'obtenir un plus grand ensemble d'entrainement appelé SEMAINE-Opinions. Ce corpus plus grand, avec des annotations homogènes ayant été obtenue avec l'audio et le texte n'a pas pu être utilisé dans le cadre de cette thèse faute de temps.

5.1 PRÉSENTATION ET TRAVAIL EFFECTUÉ PAR RAPPORT À LA BASE DE DONNÉES ICT-MMMO

5

Le consensus inter-annotateur est atteint en faisant un vote majoritaire des annotateurs. Les fichiers annotés "Je ne sais pas" par un annotateur sont écartéscar il n'y a pas de consensus entre les annotateurs. Les fichiers dont l'annotation obtenue est neutre sont aussi écartés car ils contenaient tous des annotations avec des polarités différentes par les deux annotateurs (un annotateur notait 2 et l'autre 4). On obtient au final 321 critiques (116 négatives et 205 positives) pour un total de 13h12 d'audio, composé de **12625** UIPs et de **137 157** mots.

Dans les tableaux récapitulatifs 5.1 et 5.2, il est possible de voir le nombre moyen d'observations par vidéo et par UIP. Chaque vidéo est une séquence d'environ 40 UIP qui contiennent elles-mêmes en moyenne 11 mots.

TABLE 5.1: Tableau récapitulatif des valeurs caractéristiques de *ICT-MMMO* par vidéo

Unité	**UIP par vidéo**						**Pause par vidéo**					
	μ	σ^2	**min**	**max**	**med**	Σ	μ	σ^2	**min**	**max**	**med**	Σ
Mot	10,86	11,23	1	451	8	12 625	∅	∅	∅	∅	∅	∅
Temps (sec)	3,12	3,00	0,09	129,17	2,4	39 401	0,58	0,41	0,16	8,53	0,47	7 272

Le corpus ICT-MMMO n'est pas équilibré par rapport aux différentes classes : on trouve plus de vidéos positives que de vidéos négatives. C'est un biais classique que l'on retrouve dans de nombreux corpus (Zadeh et collab., 2018b), mais qui reflète la réalité des données.

Pour conclure, le corpus ICT-MMMO est une base de données de monologues riches en opinions. Elle dispose de transcriptions faites à la main comprennent aussi les temps

TABLE 5.2: Tableau récapitulatif des valeurs caractéristiques de *ICT-MMMO* des UIP et pauses par vidéo

Unité	Par vidéo					
	μ	σ^2	min	max	med	Σ
IUP	39,33	16,20	3	93	40	12 625
Mot	427,28	144,77	46	741	448	137 157
Temps (sec)	145,40	42,89	22,64	180,80	171,93	46 674

de début et fin de chaque pause. Toutes ces particularités le rendent utile pour la première étude de cette thèse portant sur l'analyse des opinions intra-locuteur, où nous étudions une segmentation automatique adaptée à la modélisation de la dynamique de l'opinion dans un long monologue grossièrement annoté.

5.2 MISE EN PLACE D'UNE BASE DE DONNÉES À L'AIDE D'ANNOTATIONS EXISTANTES : SEMAINE-LÉGER

Durant nos travaux, nous avons créé une base de données à partir d'annotations qui ont été faites au préalable sur le corpus SEMAINE de McKeown et collab. (2012). Ces annotations provenant de sources différentes, nous avons eu besoin de les retravailler afin de donner une unité à la base de données finale.

La première sous-section portera sur la fusion de ces annotations brutes en une vérité terrain utilisable par des modèles d'AA (Sous-section 5.2.1), puis on présentera quelques statistiques sur le corpus (Sous-section 5.2.2) avant de s'attaquer aux limitations de cette méthode (Sous-section 5.2.3) qui ont motivé l'annotation que nous avons fait faire et qui est présentée dans la section 5.3 de ce chapitre.

5.2.1 Fusion des différentes annotations

Nous avons utilisé les annotations faites par Langlet et Clavel (2014, 2015b). Ceci constitue deux sous-ensembles d'annotations différents. Nous allons expliciter en détails les manières d'obtentions et les différences de ces deux sous-ensembles :

La première partie est composée de 527 Paires Adjacentes (PA) ayant étés annotées en *attitude*, notion provenant de la théorie de l'*Appraisal* de Martin et White (2003), selon le schéma de Langlet et Clavel (2014). Cette partie du corpus a été utilisé pour le développement du système de Langlet et Clavel (2014), nous l'appellerons **Semaine-D**.

La seconde partie est composée de 600 PA ayant étés annotées en attitude par Langlet et Clavel (2015b) à l'aide de l'*Amazon Mechanical Turk*. Cette partie a

été utilisée pour le test du système de Langlet et Clavel (2015b), nous l'appellerons **Semaine-T**.

Semaine-D :

L'annotation de **Semaine-D** a été faite par les auteurs et utilisé comme corpus de développement pour leur système linguistique. Le schéma d'annotation utilisé est un schéma linguistique permettant la détection des attitudes de l'utilisateur (appelés *Likes* et *Dislikes* dans leur travail) dans le contexte dialogique d'une paire de tours de parole (ie Paire Adjacente). L'*attitude* est une notion de la théorie de l'*Appraisal* comprenant les *affects*, les *jugements* et les *appreciations*. L'attitude correspond aux évaluations émotionnelles (*affect*), esthétiques (*appreciations*) et éthiques (*judgments*) d'après Martin et White (2003). Le schéma d'annotation utilisé est décrit plus en détails dans le chapitre 6.1.1 de Langlet (2018) et un exemple est disponible en Figure 4.6. Les mots d'attitudes sont obligatoirement présents pour chaque type de structure du schéma d'annotation comme indice de la valence de l'attitude.

Semaine-T :

L'annotation de **Semaine-T** a été faite lors de l'évaluation du système de Langlet et Clavel (2015a). Les auteurs ont demandé à des annotateurs selon un protocole à 4 questions, détaillé dans (Langlet et Clavel, 2015a). La première question porte sur la présence ou non de *(dis)like* (équivalent attitude), la deuxième sur le nombre de ces expressions, la troisième sur les cibles de chacune de ces opinions et la dernière question sur les valences. Le corpus atteint la somme de 600 PA à annoter, avec pour chacune entre 3 et 6 annotateurs [1].

5.2.1.1 Annotations initiales

L'annotation de **SEMAINE-D** a été faite par les auteurs de Langlet et Clavel (2015a) et utilisée comme corpus de développement du système linguistique présenté dans le même travail. Cet ensemble est composé de 527 PA annotées en attitude de l'utilisateur avec les sources, cibles associées. Cette annotation est faite sur les mots déclencheurs d'attitude ou les réponses à des questions posées par l'agent déclenchant un attitude de l'utilisateur. L'annotation suit le schéma de Langlet et Clavel (2015a) basé sur la théorie de l'*Appraisal* de Martin et collab. (2006).

Dans le corpus **Semaine-T**, on compte entre 3 et 6 annotateurs pour chacune des 600 PA. Les annotations portent la valence des évaluations de l'utilisateur humain, et sur la cible associée si elle est présente. Après élimination d'erreurs sur la base de données, nous obtenons 594 PA. Nous avons remarqué qu'un peu moins de 25% des PA du

1. du à des bugs non reportés

corpus utilisées pour le test du système de Langlet et Clavel (2016) (Semaine-T) étaient aussi utilisées pour développer ce système (Semaine-D). Nous avons choisi de garder les PA de Semaine-D, ce qui a réduit la taille de Semaine-T a 457 PA.

5.2.1.2 Création des étiquettes de vérité terrain

Afin de créer une vérité terrain pour nos modèles d'apprentissage, nous avons dû choisir une manière d'agréger les étiquettes des différentes annotations et des différents annotateurs en une seule étiquette par PA. Le procédé utilisé est décrit en annexe B.1.

Des exemples de PA constituant des exemples de PA positives et négatives sont visibles dans la figure 5.1.

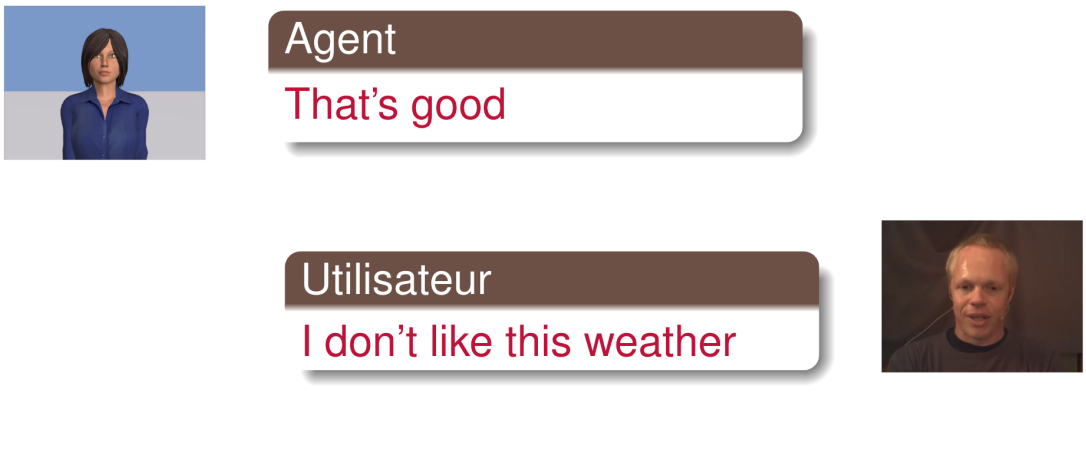
5.2.2 Statistiques

Une fois les traitements décrits ci-dessus effectuées, nous nous retrouvons avec un corpus de 958 PA utilisables : 145 négatives, 534 neutres et 279 positives provenant de 35 discussions différentes du corpus SEMAINE de McKeown et collab. (2012). Par la suite, nous allons travailler sur une tâche de classification binaire afin de commencer par une tâche plus simple. De ce fait, nous prenons uniquement en compte les PA avec des valences positives ou négatives, et n'utilisons pas les PA neutres. Nous obtenons un corpus, utilisé par la suite, composé de 424 PA non neutres.

Différentes statistiques sur les PA sont disponibles dans le tableau 5.3. On peut voir que les PA négatives contiennent en moyenne 5 mots de plus que les PA positives, et que cette augmentation est due en grande partie à l'utilisateur et non à l'agent. Ceci est un phénomène que l'on peut expliquer par le fait que l'attitude annotée est celle de l'utilisateur et non celle de l'agent. Cela peut aussi être dû au fait que l'agent suit un script et par conséquent, il est moins sensible aux variations selon l'humeur.

5.2.3 Limitations

Dans le corpus *SEMAINE-Attitude*, les transcriptions et *timecodes* utilisés pour la création des PA provenaient d'un traitement préalable mettant en jeu un système de reconnaissance de la parole (*Audio Speech Recognition* : ASR). Ces transcriptions ont servi à construire les différentes PA qui ont été annotées par Langlet et Clavel (2015b, 2014) et que nous avons utilisées pour créer le corpus final. Or, ces transcriptions et *timecodes* contenaient des erreurs. L'utilisation de méta-données provenant d'un traitement préalable a induit des erreurs qui se sont directement répercutées sur la qualité



(a) PA contenant une attitude négative de l'utilisateur

(b) PA contenant une attitude positive de l'utilisateur

FIGURE 5.1: Exemples de Paires Adjacentes annotés de *SEMAINE-Attitude*

de la base de données. Plus de détails sur les erreurs provoquant un décalage par rapport aux données réelles sont disponibles en annexe B.2.

5.3 MISE EN PLACE D'UNE BASE DE DONNÉES ANNOTÉE EN OPINIONS : SEMAINE-OPINIONS

Pour la poursuite de ce travail de thèse, nous avons souhaité continuer à utiliser la base de données SEMAINE collecté par McKeown et collab. (2012). Le type d'interactions contenues dans SEMAINE est décrit dans la sous-sous-section 4.2.2.1. Dans cette section, nous présentons le travail que nous avons effectué pour une ré-annotation en opinions de cette base de données, qui sera exempt des erreurs constatées dans la sous-section 5.2.3. Sur cette base de données déjà existante, nous avons effectué

TABLE 5.3: Tableau récapitulatif de la vérité terrain de *SEMAINE-Attitude* en fonction du label et du locuteur

Étiquette PA	**Locuteur**	**# mots par PA**						**# PA**
		μ	σ^2	**min**	**max**	**med**	Σ	
Positive	Agent	7.33	6.31	1	45	6	2045	279
	Utilisateur	11.80	12.85	1	92	9	3291	
	Tous	19.13	14.62	2	121	16	5336	
Négative	Agent	8.87	9.17	1	77	7	1286	145
	Utilisateur	15.57	18.53	1	105	11	2258	
	Tous	24.44	21.42	2	121	17	3544	
Toutes	Agent	7.86	7.44	1	77	6	3331	424
	Utilisateur	13.09	15.12	1	105	9	5549	
	Tous	**20.94**	**17.41**	**2**	**121**	**17**	**8880**	

les traitements qui suivent. Nous avons d'abord normalisé les transcriptions faites à la main et aligné le texte et l'audio mot par mot afin d'obtenir les *timecodes* précis de chaque mot (Sous-section, 5.3.1). Ensuite, nous avons créé un schéma d'annotation en opinion sur des tours de parole d'une discussion entière (Sous-section 5.3.2) et qui a été implémentée sur une plateforme web créée à partir d'une plateforme existante. Pour finir, nous avons fait annoter 79 discussions en opinion au niveau du tour de parole via le site participatif *Figure Eight*[2] via des retours avec les annotateurs et un dispositif de visualisation (Annexe A). Ce processus a permis d'augmenter la qualité des annotations au fil du temps, pour obtenir un corpus annoté en opinion au niveau des tours de parole et par rapport à la discussion entière . Nous présentons ce corpus en Sous-Section 5.3.3, ainsi que les statistiques sur les observations et sur les annotations. Nous avons calculé l'accord inter-annotateurs en utilisant le α de Krippendorff (2013). La variance des valeurs des alphas sur les différentes sessions est très grande, allant de 0.2 à 0.9.

Ce corpus n'a malheureusement pas pu être exploité dans le cadre de cette thèse faute de temps, et les expériences des chapitres 9 et 10 utilisent le corpus SEMAINE-Léger.

5.3.1 Travail préliminaire sur la base de données initiale

La base de données que nous avons utilisée est fournie avec l'ensemble des captations : différents **fichiers vidéos** provenant des différentes caméras, différents **fichiers audio** provenant des différents microphones, des **transcriptions manuelles** pour cer-

2. https ://www.figure-eight.com/

tains des fichiers, etc... On peut voir en figure 5.2 une photo prise lors d'un enregistrement de SEMAINE avec les multiples microphones utilisés lors de la collecte de la base de données.

FIGURE 5.2: Photographies d'un enregistrement du corpus SEMAINE où l'on peut voir à la fois la salle de l'utilisateur (gauche) et la salle de l'opérateur (droite)

Les fichiers fournis avec SEMAINE proviennent de 20 sessions différentes. Chaque session est composée de 4 discussions entre utilisateur et un opérateur jouant le rôle d'un des personnages de SEMAINE dans chacune des discussion. Une discussion reste inutilisable car l'utilisateur a choisi d'arrêter l'expérience avant la fin, rendant le nombre total de discussion a 79. Nous avons utilisé les transcriptions manuelles des 79 discussions ainsi que les enregistrements audio provenant des microphones montés sur les participants.

5.3.1.1 Normalisation des transcriptions

Afin d'utiliser pleinement le contenu textuel, nous avons utilisé les transcriptions manuelles fournies avec le corpus SEMAINE procuré par McKeown et collab. (2012). [3] Par rapport aux transcriptions automatiques, les transcriptions manuelles ont l'avantage d'être de meilleure qualité. Cependant, le problème de la non-normalisation du texte apparait généralement si des normes précises n'ont pas été érigées pour les transcripteurs. En effet, que ce soit parce que les transcriptions sont faites par plusieurs personnes différentes, ou bien à cause de typographies, les même phénomènes peuvent être retranscrits de manières différentes.

Dans cette partie, nous aborderons la normalisation que nous avons effectuée sur le texte, sur la ponctuation et sur les annotations para-linguistiques afin de normaliser l'ensemble des transcriptions, et de prendre pleinement partie de la diversité des mots-vecteurs disponibles. En effet, nous avons profité de l'utilisation d'un ensemble de mots-vecteurs appris au préalable sur une grande quantité de texte et disposant d'un vocabulaire imposant. **Les opérations que nous avons effectuées portent sur** : mots

3. https ://semaine-db.eu/

hors-vocabulaire (*Out Of Vocabulary* : OOV), répétitions de caractères, casse, chiffres, orthographes régionales et orthographe générale, mots coupées et annotations para-linguistiques.

Repérage des mots OOV Pour repérer les mots OOV, nous avons utilisé le vocabulaire des vecteurs word2vec disponibles, qui est la représentation distribuée utilisée comme descripteur textuel dans nos expériences (voir Chapitre 6 pour plus de détails). Après une segmentation du texte en symboles (nous utiliserons le mot *token*) à l'aide du CoreNLP (Schuster et Manning, 2016), nous avons commencé par regarder tous les mots OOV, c'est à dire qui n'avaient pas de mot-vecteur associé dans le vocabulaire des représentations word2vec. Le vocabulaire étant très riche et composé de mots avec des orthographes "exotiques", cette approche permet d'utiliser des orthographes diverses, afin de représenter chaque mot par une représentation au plus proche de ce que le transcripteur avait en tête. Par exemple, *awesoome* pourra avoir un vecteur différent de *awesome* : l'utilisation de son vecteur est préconisé car sa représentation contient l'information sémantique associée à cette élongation orale (intensification). Les mots n'ayant pas de vecteurs sont ensuite traités avec les procédés expliqués ci-dessous afin gérer des spécificités ou typographies.

Les répétitions de lettres Lorsque 3 lettres similaires ou plus se suivent nous réduisons le nombre à 2 lettres. Par exemple, le mot *awesooooooome* devient *awesoome*. Nous avons fait ce choix car s'il ne peut pas y avoir plus de 3 lettres similaires qui se suivent en anglais, des vecteurs comme *awesoome* étaient présent mais non *awesooome*.

Les répétitions de ponctuation Lorsque 4 signes de ponctuations similaires ou plus se suivent, nous réduisons le nombre à 3 signes. Cela permet d'utiliser ces signes de ponctuations comme descripteurs textuel et d'une manière homogène sur le corpus entier et ajouter des indications para-linguistiques dans la représentation du signal.

Les mots en majuscule Certains mots comme *WHAT*, peuvent avoir un vecteur associé. Ces vecteurs de mots avec une casse spéciale sont des représentations spécialisées comme ceux avec des répétitions de lettres (voir paragraphe plus haut). Néanmoins, la majorité des mots en majuscules n'ont pas de vecteur associé, mais lorsque l'on change la casse pour des lettres minuscules, on trouve un vecteur. Par exemple, le mot *CONTRADICT* n'a pas de vecteur alors que le vecteur de *contradict* est présent dans le vocabulaire. Il est à noter que l'on garde toutefois l'information de la casse en ajoutant un descripteur "mot en majuscule".

Les mots en minuscule Des fois, c'est une majuscule qu'il manque à la première lettre du mot pour lui trouver un vecteur associé. Par exemple, le mot *rapunzel* n'a pas de représentation alors que *Rapunzel* en a une, et fait référence au personnage du conte.

Les nombres Les nombres qui n'ont pas de vecteurs sont transformés en nombres semblables ayant des vecteurs associés. On trouve des vecteurs associés pour tous les nombres de 1 à 10, les dizaines, ainsi que les centaines.

Les orthographes régionales Les mots-vecteurs que nous utilisons ont été appris pour une orthographe de type américaine et non de type Grande-Bretagne. Pour cela, nous utilisons un dictionnaire de l'anglais britannique vers l'anglais américain de 1737 mots trouvés sur le net.[4] Dès qu'un mot de notre corpus a une entrée dans ce dictionnaire, nous changeons le mot pour sa version orthographique américaine. Par exemple, on trouve les mots : *humour*, *colour*, *baptise*, *learnt*,...

Les inachèvements Chose fréquente à l'oral, certains mots sont coupés quand le locuteur arrête de parler. Il est possible de repérer de tels phénomènes car ils sont souvent caractérisé par des points de suspension à la fin du mot. Cela permet de reperer des mots coupés comme : '*usua*', '*hopef*', '*biling*', '*sentim*', '*interrup*' ou '*whenev*'. Etant donné qu'il est impossible de deviner la fin du mot, une balise d'inachèvement est placé à cet endroit pour être utilisée comme descripteur.

Le correcteur orthographique Lorsqu'aucune des techniques ci-dessus n'a pu fonctionner pour trouver un vecteur associé à un mot, alors nous utilisons un correcteur orthographique[5] pour détecter des erreurs de typographie. Ce correcteur n'est pas sensible à l'orthographe régional. Par exemple, les mots *blessd*, *f*ck*, *frushtrating*, *stressin*, *kindsa* et *despressed* sont corrigés en *blessed*, *fuck*, *frustrating*, *stressing*, *kinda* et *depressed*.

Les mots dont on ne peut rien faire Une fois toutes ces opérations effectuées, il reste quelques mots sans vecteur qui sont impossibles à traiter de manière automatique. Généralement ces mots ne sont pas reconnus car ils ont au moins 2 types d'erreurs en cumulé. Par exemple, on trouve *Gelotophobe* (Majuscule + typographie), *reaslise* (orthographe britannique + typographie) , *bibble* (le correcteur le prend comme bon, *bibble* étant un logiciel de traitement d'images photographiques)

4. http ://www.tysto.com/uk-us-spelling-list.html
5. https ://github.com/phatpiglet/autocorrect

Les annotations para-linguistiques Le corpus contient des annotations para-linguistiques très variées, la manière de les écrire ayant été laissée à la discrétion du transcripteur. Comme nous n'étudions que les modalités orales et textuelles, nous avons choisi d'écarter toutes annotations visuelles (du type mouvement de la tête, sourire, ...). Pour cela, nous ne conservons que les annotations para-linguistiques contenant des chaines de caractères faisant référence à un phénomène oral : *'voice'*, *'laugh'*, *'whisper'*, *'tone'*, *'breath'*, *'swallow'*, *'sigh'*, *'say'*, *'said'*, *'noise'*, *'sniff'*, *'giggl'*, *'cackl'*, *'speak'*, *'tut'*. Les phénomènes para-linguistiques conservés sont du ressort de l'intonation, de la prononciation, du rire et du volume sonore.

5.3.1.2 Alignement des mots

Les transcriptions manuelles sont exactes par rapport à celles provenant de l'ASR qui contiennent de nombreuses erreurs. Cependant, les sessions ont été coupées au préalable à la main en quatre discussions. La concaténation de tous les fichiers audio de chaque discussion ne rend pas la session complète. Nous avons donc eu recours à un procédé décrit en annexe (Annexe C) afin de trouver les codes temporels (nous utiliserons le mot anglais *timecodes*) de chaque tour de parole par rapport à l'audio, correspondant à chaque discussion. Une fois les *timecodes* obtenus, nous avons pu isoler l'audio de chacun des tours de parole pour aligner le fichier avec son texte associé à l'aide de l'aligneur de Ochshorn et Hawkins (2017).

Le taux de mots non-alignés par rapport au nombre de mots total que l'on trouve dans le corpus est de 1.21% ce qui parait raisonnable pour notre application. Lorsqu'un mot n'est pas aligné, nous nous servons des *timecodes* des mots voisins présents dans le même tour de parole pour lui trouver des temps de début et de fin.

5.3.2 Schéma d'annotation

Nous voulons étudier la dynamique de l'opinion inter-locuteur et intra-locuteur dans le contexte particulier des interactions orales. Pour cela, nous avons choisi d'utiliser des annotations en opinion dans des interactions de parole à la granularité d'un tour de parole. Voici la présentation du schéma d'annotation que nous avons créé, implémenté, mis en ligne puis utilisé pour la collecte des annotations du corpus SEMAINE.

5.3.2.1 Enjeux et présentation générale

Créer un schéma composé de questions simples à poser aux annotateurs est une tâche difficile lorsque l'on s'intéresse à des phénomènes subjectifs tels que les opi-

nions. Les tours de parole contenant au moins 2 opinions de valence différentes nous ont semblé un cas important à cerner. Pour cela, nous avons créé un processus d'annotation basé sur plusieurs questions afin de rendre la tâche plus simple pour l'annotateur, diminuer sa charge cognitive et ainsi lui permettre d'être plus attentif à la détection des phénomènes recherchés. Nous avons repris l'idée d'un historique de la conversation et des annotations antérieures visible proposé par Langlet et collab. (2017) pour une tâche d'annotation de préférences, tout en changeant complètement le schéma d'annotation pour notre tâche.

Le schéma d'annotation utilisé pour l'annotation en opinion de notre corpus peut être résumé par le schéma-bloc de la figure 5.3. L'annotation se fait tour de parole par tour de parole, en ayant toujours devant les yeux le tour de parole précédent et le tour de parole actuel (voir Figure 5.4). Cela permet à l'annotateur de prendre en compte le contexte interactionnel dans son annotation. Une autre fenêtre est visible avec l'historique de la conversation ainsi que l'historique des annotations de l'annotateur. Pour chaque tour de parole, on compte une série de 1 à 3 questions qui se suivent (décrites Figure 5.3).

5.3.2.2 Une annotation par étapes

Avant de commencer à effectuer une tâche complexe comme l'annotation d'opinions, il est nécessaire de passer par différentes phases afin d'amener une rapide montée en compétence de l'annotateur. Ceci permet de diminuer la possibilité d'une mauvaise interprétation de la tâche à effectuer.

Phase d'instructions :

Premier contact La tâche d'annotation commence par une phase d'instruction. Cette phase d'instruction contient plusieurs niveaux. La première étape de cette phase d'instruction commence lors de la présentation de la tâche à effectuer via la plate-forme d'annotation collaborative *Figure Eight*.[6] La page de présentation est visible en annexe A et elle résume l'ensemble de la tâche à effectuer de manière simplifiée.

Instructions détaillées Une fois que l'annotateur a choisi d'effectuer la tâche, et qu'il s'est identifié sur la plate-forme en ligne, des instructions et informations plus complexes lui sont présentées. Ces instructions et informations portent sur :

- les différents locuteurs : discussion orale entre un agent virtuel et un utilisateur ;
- l'introduction de la notion de Paire Adjacente (PA) : deux tours de parole de locuteurs différents qui se suivent ;
- le schéma d'annotation présenté plus haut :

6. https ://make.figure-eight.com

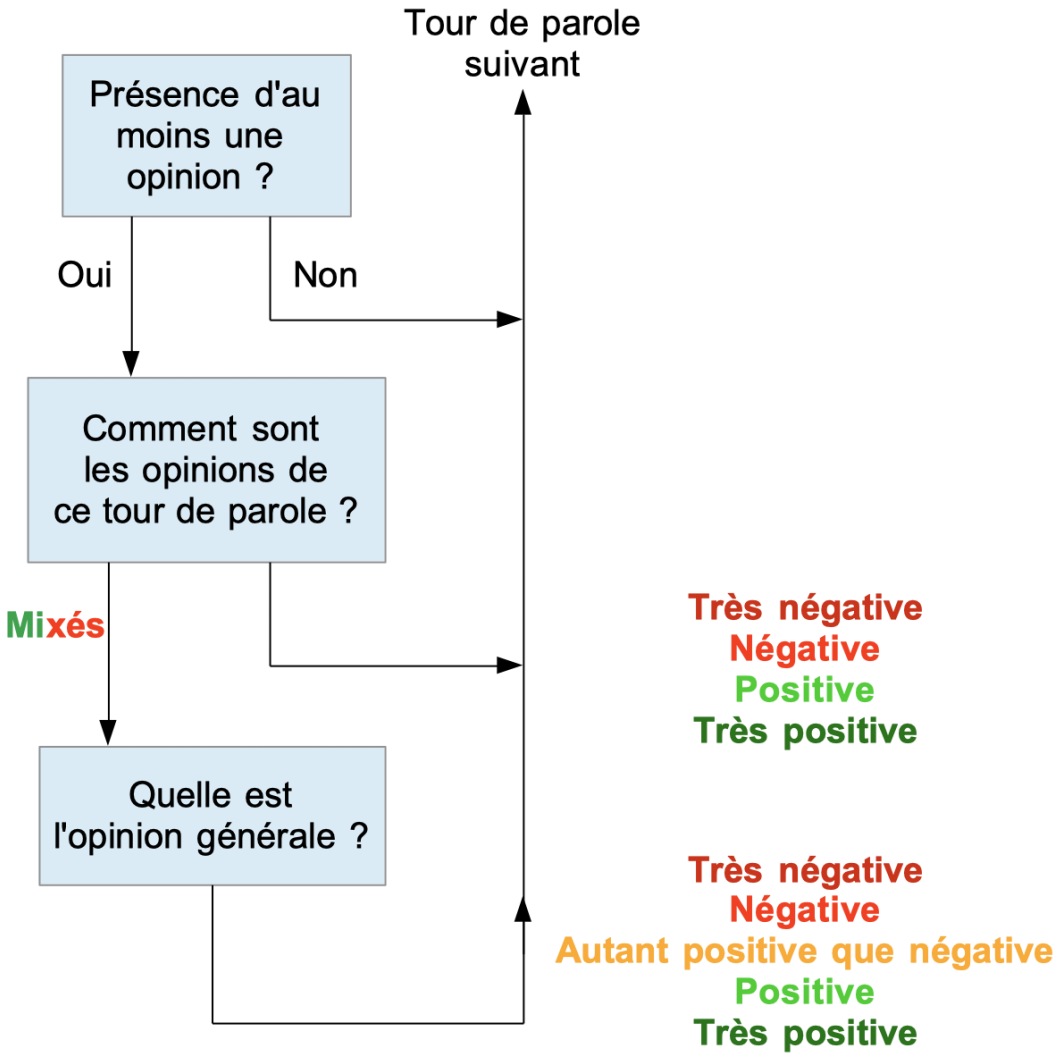


FIGURE 5.3: Diagramme en blocs du schéma d'annotation que nous avons utilisé pour annoter le corpus SEMAINE

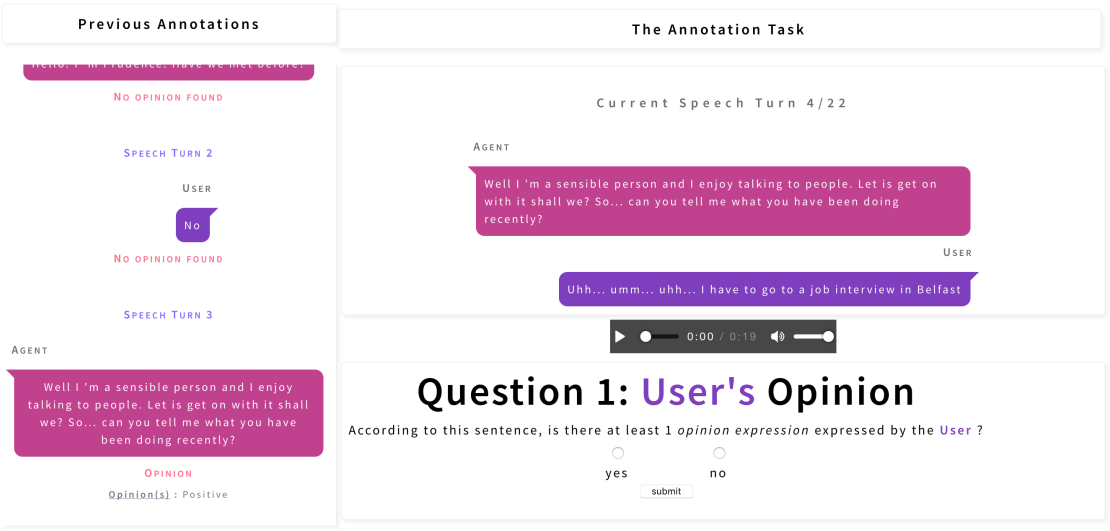


FIGURE 5.4: Exemple provenant de la plateforme d'annotation avec l'historique à gauche

5

- Question 1 : Le locuteur exprime-t-il une opinion ?
- Quelle est la valence de l'opinion générale ?
- S'il y a des valences différentes, quelle est la valence de l'opinion proéminente ?

— la définition d'une opinion, d'une non présence d'opinion, et des exemples associés ;

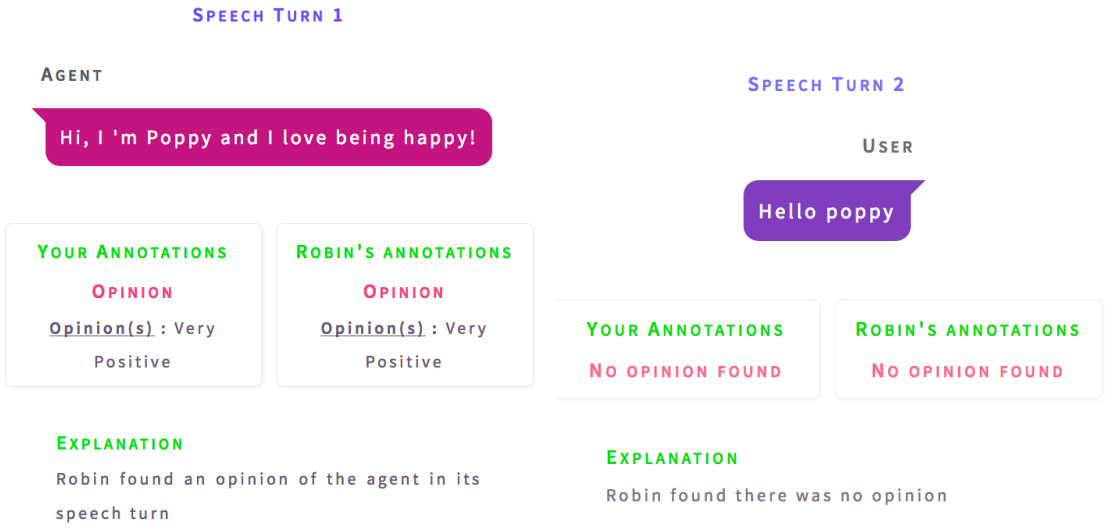
— la définition d'une opinion proéminente dans un tour de parole contenant des opinions de différentes valences, et des exemples associés ;

— l'introduction d'une phase d'apprentissage : Courte discussion ou l'annotateur apprend à prendre en main les outils et utiliser les concepts présentés plus haut ;

Des détails supplémentaires sur les instructions sont disponibles en Annexe A.

Phase d'apprentissage et phase d'annotation :

L'utilisation d'une phase d'apprentissage nous a été inspirée par le travail de Langlet et Clavel (2016). Nous avons créé une phase d'apprentissage préalable où l'annotateur pourra s'exercer sur une conversation factice composée de 8 tours de parole. Lors de cette phase d'annotation, il pourra également voir les annotations d'un annotateur ordinaire appelé "Robin". **Cette phase de transition permet à l'annotateur de s'entraîner et de se comparer avec des résultats que nous avons jugés comme valides au préalable**. Par exemple, nous avons souvent vu que les annotateurs ont du mal à faire la différence entre une émotion et une opinion, qui sont deux choses distinctes. De même, beaucoup d'annotateurs étiquetaient souvent comme opinion une question



(a) Exemple avec la présence d'une opinion (b) Exemple sans la présence d'une opinion

FIGURE 5.5: Exemples d'annotations de Robin avec les explications associées

de l'agent destinée à déclencher une opinion de l'utilisateur, mais n'étant elle-même nullement une opinion.

Pour finir, afin de ne pas stresser l'annotateur, il lui est bien répété qu'il n'est pas évalué sur la phase d'entraînement et que la réponse de Robin n'est qu'une réponse bonne parmi plusieurs possibles. Des exemples sont montrés dans la figure 5.5.

Une fois la phase d'apprentissage effectuée, l'annotateur commence la phase d'annotation réelle sur une des discussions du corpus SEMAINE. Cette phase d'entraînement est tout le temps la même. Un annotateur ayant déjà effectué l'entraînement peut passer cette phase à l'aide d'un code qui lui sera remis à la fin de l'annotation.

5.3.3 Présentation de notre base de données

Dans cette section nous présenterons en détail le sous-ensemble de SEMAINE que nous avons fait annoter et que l'on nomme dans ce manuscrit *SEMAINE-Opinion*. Pour rappel, le corpus SEMAINE est composé de sessions séparées en 4 discussions. Une session correspond à un participant côté utilisateur, et une discussion correspond à une interaction entre l'utilisateur et un agent. Dans SEMAINE, nous comptons 4 agents : Spike qui est colérique, Obadiah qui est morose, Prudence qui est pragmatique et Poppy qui est joyeux.

5.3.3.1 Généralités et taille de la base de données

Cette sous-section présente la base de données que nous avons faite annoter. Pour construire notre base de données annotée, nous avons choisi d'utiliser le sous-ensemble de SEMAINE qui possédait des transcriptions. La base de données SEMAINE compte 20 sessions avec les transcriptions associées qui sont disponibles : cela correspond à 80 discussions. L'une des discussions étant inutilisable car l'utilisateur a préféré arrêter l'expérience, **le corpus final *SEMAINE-Opinions* est composé de 79 discussions**. Le lecteur pourra trouver dans le tableau 5.4 des statistiques sur le nombre de mots et les temps de parole de chaque locuteur par tour de parole et par discussion, ainsi que les histogrammes associés en figure 5.6.

(a) Tableau récapitulatif des valeurs caractéristiques de *SEMAINE-Opinion* par discussion

Locuteur	**# Mots par Disc.**						**Temps de parole par Disc.(s)**					
	μ	σ^2	**min**	**max**	**med**	Σ	μ	σ^2	**min**	**max**	**med**	Σ
Agent	306	140	67	670	290	XXX	101	55	32	363	88	8 006
Utilisateur	630	343	96	1602	640	XXX	187	90	32	447	186	14 809
Tous	936	407	210	2220	953	XX	289	118	75	638	291	22 815

(b) Tableau récapitulatif des valeurs caractéristiques de *SEMAINE-Opinion* par tour de parole

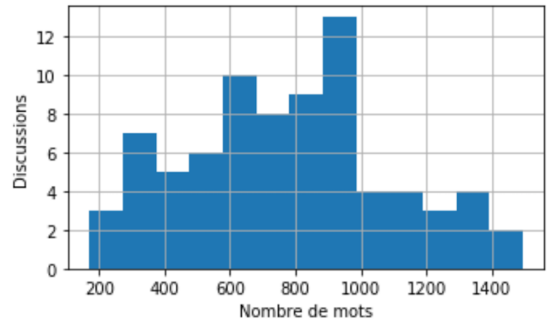
Locuteur	**# Mots par TP**						**Temps de parole par TP (s)**					
	μ	σ^2	**min**	**max**	**med**	Σ	μ	σ^2	**min**	**max**	**med**	Σ
Agent	8,6	8,2	1	106	6	XX	2,7	2,8	0,1	45,2	1,9	8 006
Utilisateur	15,66	18,31	1	103	9	XX	5,3	7,1	0,1	83,0	2,9	14 809
Tous	13,1	19,0	1	256	7	XXX	7,9	8,2	0,1	92,9	5,5	22 815

TABLE 5.4: Tableaux récapitulatifs des valeurs caractéristiques de *SEMAINE-Opinion*

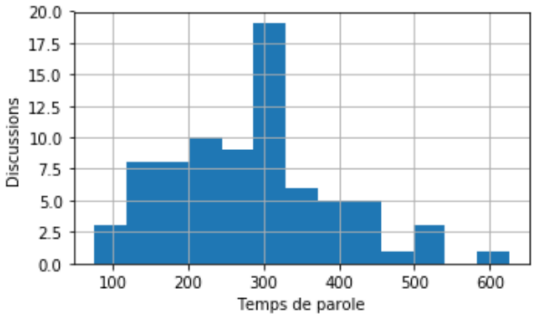
On peut observer dans le tableau 5.4 que les tours de parole de l'utilisateur sont plus longs que ceux de l'agent. Ceci peut être expliqué par le fait que l'agent est censé suivre un script et possède une personnalité peu complexe. Il est simplement censé déclencher des émotions chez l'utilisateur, qui est l'interactant devant s'exprimer le plus.

Les histogrammes en figure 5.6 montre une répartition classique du nombre de mots et du temps de parole par discussions. Cependant, les distributions des nombres de mots et temps de parole par tour de parole (TP) ont une longue queue, ce qui veut dire qu'il existe de très longues tirades dans certaines discussions, qui vont perturber la modélisation de la conversation.

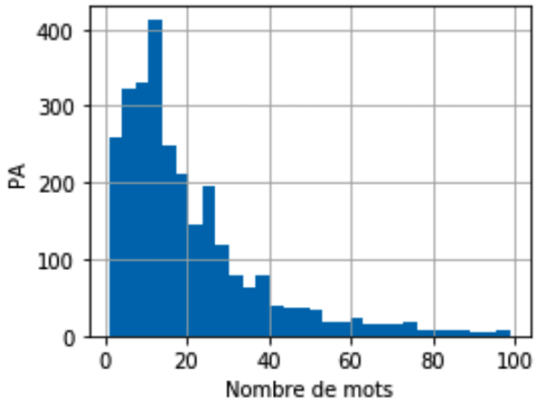
Une annotation de ce type permet d'obtenir une classification des PA à la fois par rapport à l'opinion de l'agent comme pour le corpus *SEMAINE-Léger*, mais aussi par rapport à l'opinion de l'utilisateur. Contrairement à *SEMAINE-Léger* où les PA étaient annotées indépendamment du contexte de l'historique du dialogue, la continuité dans



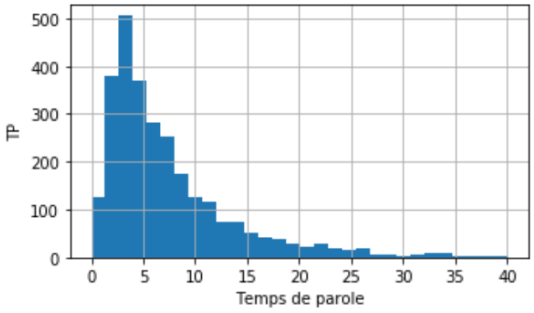
(a) Histogrammes du nombre de mots par discussion



(b) Histogrammes de la durée d'une discussion en seconde



(c) Histogrammes du nombre de mots par tours de parole



(d) Histogrammes de la durée des tours de parole en seconde

FIGURE 5.6: Histogrammes des tours de parole selon différentes valeurs

l'annotation d'une même conversation ouvre ici plusieurs possibilités. Premièrement, cet annotation rend possible l'apprentissage de modèles d'AA qui utilisent l'historique de la discussion. Ensuite, il devient possible d'effectuer un étiquetage de séquence au niveau des PA et même des tours de parole car il n'y a pas de trous dans l'annotation (contrairement au corpus CMU-MOSI (Zadeh et collab., 2016) où l'intégralité de la vidéo n'est pas présente). Pour finir, c'est à notre connaissance, la première annotation en opinion et par tour de parole d'une base de données multimodale interactionnelle.

5.3.3.2 Accord inter-annotateur et influence de la classe *Mixed*

Les opinions sont des phénomènes difficiles à caractériser et très subjectifs (Callejas et collab., 2014). Ainsi, l'obtention d'une vérité terrain fiable pour entrainer un modèle n'est pas une tâche simple. Obtenir des annotations cohérentes entre les annotateurs est une tâche dont la difficulté grandit avec la taille du segment à annoter.

Zadeh et collab. (2016) ont choisi de segmenter le discours en blocs fins afin d'avoir des segments de phrase contenant le minimum nécessaire d'information utile à l'annotation d'un sentiment. De cette manière, on obtient moins d'annotations contraires, et un accord inter-annotateur plus grand.

Dans cette thèse, le sujet d'étude est l'analyse des opinions dans des conversations et non des monologues, ce qui rend plus difficile d'effectuer cette segmentation car le locuteur n'est pas le seul maitre du discours. De plus, nous voulions que l'annotateur puisse travailler d'une traite sur la conversation entière. Il devenait impossible de lui faire annoter trop de choses à la fois sans faire exploser sa charge mentale et perdre en qualité d'annotation. Nous avons donc opté pour une segmentation par tour de parole, et non par phrase, afin d'éviter un travail trop intense à l'annotateur.

Pour s'affranchir du problème de la subjectivité des opinions, nous avons créé une étiquette spéciale pour les opinions les moins consensuelles : comme notre segmentation d'annotation est en tour de parole, certains tours de paroles contiennent différentes opinions et c'est dans ces cas-là que les annotateurs vont différer le plus. C'est aussi dans ces cas que les annotateurs vont hésiter entre plusieurs classes. Pour cela, nous avons incorporé une étiquette spéciale appelé *Mixed* utilisable par les annotateurs lorsque le tour de parole contient des opinions ayant des valences différentes. Cependant, comme nous voulions rester sur un problème de classification plus simple, les annotateurs utilisant l'étiquette *Mixed* se voient demandés ensuite quelle est l'opinion proéminente dans ce tour de parole. De cette manière, **nous pouvons savoir que ces tours de paroles peuvent être problématiques tout en utilisant le label de l'opinion proéminente pour effectuer une tâche de classification avec ces données**.

Accord inter-annotateur : choix du α de Krippendorff :

L'utilisation d'une mesure d'accord inter-annotateur est classique dans la validation de données ayant été codées par différents individus. Dans le cas de l'annotation d'opinions pour l'apprentissage automatique où la subjectivité tient une place importante, il est crucial de savoir si les annotations agrégées sur lesquelles nos modèles vont s'entrainer sont fiables et représentent une réalité consistante. Pour chaque discussion, nous comptons 3 annotateurs par session. Il est nécessaire de choisir une mesure représentative de l'accord entre les différents annotateurs. Les mesures décrites ci-dessous consistent à modéliser les variations dans les accords dûs au hasard, au type de données et aux différents annotateurs.

Le π de Scott (1955) modélise la distinction du hasard entre les catégories mais pas entre les annotateurs. Le κ de Cohen (1960), plus informatif, permet de faire cette distinction en prenant en compte les différences dans la distribution des valeurs entre les catégories pour les différents annotateurs (Lombard et collab., 2002) . Cependant, les π de Scott et κ de Cohen ne sont utilisables que pour des données avec 2 codeurs uni-

quement. Il est possible de calculer des valeurs en prenant les annotateurs 2 à 2 mais ce n'est pas conseillé car cela ne permet pas de prendre en compte la covariance générale des codeurs (Hallgren, 2012). Le κ de Fleiss (1971) est une généralisation du π de Scott pour plus de 2 codeurs, souffrant toujours du même défaut d'absence de modélisation des différents codeurs. Le α de Cronbach (1951) n'est pas considéré car, d'après Hughes et Garrett (1990), cet indice mesure uniquement la cohérence interne des annotations via une corrélation et retranscrit uniquement une mesure de covariance qui ne permet pas de modéliser un accord entre des annotateurs.

Pour cela, nous avons choisi d'utiliser le α de Krippendorff (2013), car c'est un coefficient qui est adapté à une comparaison de plus de 2 codeurs, contrairement au κ de Cohen et au π de Scott. Le α de Krippendorff a été conçu pour des études sur des codeurs analysant un contenu (Hallgren, 2012). Il permet de modéliser de nombreux aspects, à l'instar du κ de Cohen et contrairement au κ de Fleiss. Le α est aussi le coefficient utilisé par Zadeh et collab. (2017) pour calculer l'accord inter-annotateurs des annotations en sentiment de la base de données CMU-MOSI. Il calcule le pourcentage de désaccord $\frac{D_o}{D_e}$ entre les annotateurs, comme montré dans l'équation (5.1) :

$$\alpha = 1 - \frac{D_o}{D_e} \tag{5.1}$$

où D_o (équation 5.2) représente le désaccord observé et D_e (équation 5.3) représente le désaccord attendu par le hasard.

$$D_o = \frac{1}{n} \sum_{c \in R} \sum_{k \in R} \delta(c,k) o(c,k) \tag{5.2}$$

$$D_e = \frac{1}{n(n-1)} \sum_{c \in R} \sum_{k \in R} n_c n_k \delta(c,k) \tag{5.3}$$

où R est l'ensemble des réponses possibles, $o(c,k)$, n_c, n_k et n se réfèrent aux fréquences des valeurs dans des matrices de co-occurrences et $\delta(c,k)$ une métrique de calcul d'un désaccord entre deux réponses c et k.

Son calcul est long et laborieux (Lombard et collab., 2002) impliquant des matrices de co-occurrences, mais ce n'est pas un problème dans notre cas puisque nous calculons cet accord une seule fois hors-ligne. D'après Krippendorff (2004), une mesure de α de 0.667 est nécessaire pour avoir un accord satisfaisant. Pour plus de détails sur le calcul du α, les lecteurs sont invités à se référer à (Krippendorff, 2011). On utilisera de la manière les valeurs de α et 100α de manière interchangeable dans la suite du texte, la différence entre les deux étant assez grande pour ne pas porter à confusion (ex : 0.66 ou 66).

Analyse de la classe *Mixées* :

Nous pouvons voir dans le tableau 5.5 les différences d'accords inter-annotateurs en prenant en compte l'étiquette *Mixées* comparé avec celui calculé utilisant les annotations en opinions proéminentes comme étiquette. Les colonnes des valeurs agrégées (moyenne, médiane, écart-type, minimum et maximum) sont calculées depuis les α de chaque discussion et la colonne "Total" est calculée sur le corpus entier.

TABLE 5.5: Tableau comparatif des α de Krippendorff pour *SEMAINE-Opinion* avec l'utilisation de l'étiquette Valences Mixées *versus* Opinion proéminente

Locuteur	**α 4 classes**						**α 3 classes (Mixées –> Proéminente)**					
	μ	σ^2	**min**	**max**	**med**	**Total**	μ	σ^2	**min**	**max**	**med**	**Total**
Agent	53.5	18.6	15.1	95.0	55.7	∅	60.4	16.1	34.0	100	60.5	∅
Utilisateur	45.4	17.7	-7.1	89.4	48.0	∅	54.9	15.0	13.3	85.9	56.0	∅
Tous	52.8	13.1	25.0	91.7	50.4	57.8	60.9	11.0	39.1	90.6	58.1	66.3

Au vu de la baisse de l'accord inter-annotateur, il est difficile d'utiliser l'étiquette *Mixées*. L'utilisation de 4 classes donne un α de Krippendorff de 57.8, ce qui est faible, quoi que relatif car nous travaillons sur des opinions dans la parole qui est un phénomène subjectif difficile à quantifier pour des annotateurs non experts. Cependant, l'utilisation de l'opinion proéminente permet d'augmenter notre α à une valeur raisonnable de 66.3 sur tout le corpus. La valeur du α est supérieure pour les TP de l'agent que pour les TP de l'utilisateur : étant donné que l'agent est joué par un acteur ayant un rôle précis, on peut s'attendre à avoir des opinions plus tranchées et plus facilement identifiables par les annotateurs.

Les tours de parole contenant des opinions à valences mixées sont importantes. En analysant les résultats du tableau 5.6, il est possible de voir que l'accord inter-annotateur augmente lorsque l'on ne prend pas en compte chaque tour de parole ayant été étiqueté *Mixées* par au moins un des annotateurs. Les opinions des tours de parole pour lequel aucun des annotateurs n'a trouvé une classe mixte sont plus consensuels que les autres. Le α pour les TP non mixées uniquement atteint un score de 71.4 par rapport au score de 66.3 sur la totalité, et de 57.8 en comptant 4 classes différentes.

On peut aussi noter que pour les classes mixtes uniquement, l'écart est très fort entre le min et le max : ceci est simplement dû au faible nombre de TP annotés mixtes dans certaines discussions. En effet, s'il n'y a que 3 TP dans une session, il est facile d'arriver à un accord inter-annotateur très faible par rapport à une session ayant 15 TP de ce type. Curieusement, si les TP de l'agent obtiennent de meilleurs accords inter-annotateurs en général, ce sont ceux de l'utilisateur qui obtiennent de meilleurs scores sur les TP mixtes (Figure 5.6, tableau de droite). Une explication est que l'agent à moins de TP avec mixte car il doit suivre un script. Néanmoins quand il sort de son script, c'est qu'il est en difficulté dans la conversation.

TABLE 5.6: Tableau comparatif des α de Krippendorff pour les tours de parole de *SEMAINE-Opinion* ayant été étiquetés Valences Mixées par un annotateur au moins *versus* les autres tours de parole

Locuteur	α **Non Mixées uniquement**						α **Mixées uniquement**					
	μ	σ^2	**min**	**max**	**med**	**Total**	μ	σ^2	**min**	**max**	**med**	**Total**
Agent	64.2	16.2	27.5	100	63.7	∅	28.3	37.8	-31.6	100	19.7	∅
Utilisateur	57.2	18.3	5.3	100	59.5	∅	38.6	29.4	-27.3	100	38.4	∅
Tous	65.1	11.9	33.5	97.1	64.1	71,4	39.4	27.5	-27.3	100	39.7	43.7

5.3.3.3 Opinions annotées : agrégation et analyse

Maintenant que nous avons présenté et analysé les différentes mesures inter-annotateurs obtenus sur toute la base de données, nous allons nous concentrer sur les étiquettes qui seront utilisées par nos systèmes d'apprentissage supervisé. Ces modèles s'entrainent sur une vérité-terrain composée d'un unique label et non de trois, il est donc nécessaire de faire une agrégation des annotations. Une fois l'agrégation faite, nous effectuerons une analyse de la vérité-terrain afin d'être rassuré quant à la consistance des données. Nous présenterons ci-dessous des statistiques sur les annotations en opinion par tour de parole que nous avons récoltées, et agrégées.

Agrégation par vote majoritaire

Si certains travaux ont préféré garder l'information de grande variance entre les annotateurs (Dang et collab., 2018) et l'incorporer dans leurs modèles, nous sommes restés sur une approche traditionnelle. Faire un vote de majorité est une des manières d'obtenir des valeurs proches de la réalité. Pour cela, nous avons décidé d'agréger nos annotations à l'aide d'un vote majoritaire comme Wöllmer et collab. (2013a). Le vote de majorité consiste à prendre la valeur donnée par la plus grande partie des annotateurs. Dans le cas où nous avons 3 annotateurs, la majorité est obtenue lorsqu'il y a au moins 2 réponses similaires. Le cas particulier où chacun des annotateurs a donné une réponse différente n'est jamais arrivé.

Analyse de la vérité-terrain

Nous avons obtenu un accord inter-annotateur calculé à l'aide du α de Krippendorff (2013) ayant pour valeur 66.3, pour 3 annotateurs par discussion. Ce α est relativement faible mais tout à fait acceptable pour une tâche aussi subjective et difficile que l'annotation d'opinion. Par comparaison, Zadeh et collab. (2017) obtiennent une valeur de 0.77 sur CMU-MOSI, mais sur des données non interactionnelles donc

sans l'ambiguïté apportée par le contexte dialogique. De plus, notre segmentation en TP nous oblige à ne pas couper de longues tirades, ce qui augmente les possibilités de confusion dans l'annotation. Finalement, nos annotateurs utilisent uniquement l'audio et non la vidéo. Ces points facilite grandement la tâche d'annotation deZadeh et collab. (2017).

On pourrait penser que les opinions sont un phénomène rare dans le corpus et que la plus grande partie des annotations obtenues avec le vote de majorité seraient donc des opinions neutres, mais c'est tout l'inverse. On peut observer que sur la totalité du corpus, environ 52% des tours de parole contiennent des opinions : 20.7% en opinions négatives et 27.4% en opinions positives. Les analyses statistiques des annotations de vérité-terrains sont visibles en Tableau 5.7.

TABLE 5.7: Tableau récapitulatif de la vérité terrain obtenue après agrégation par discussion de *SEMAINE-Opinion*

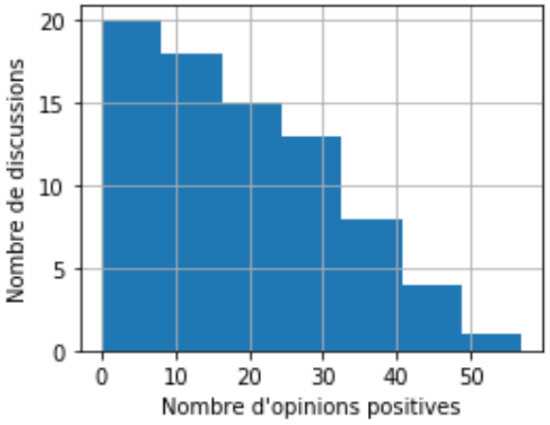
Locuteur	**# Opinions par Disc.**						**Opinions par TP (%)**		
	μ	σ^2	**min**	**max**	**med**	Σ	**-**	**+**	**~**
Agent	15.22	7.71	2	37	14	1203	22.84	21.14	56.02
Utilisateur	17.90	8.98	3	47	17	1414	18.47	33.75	47.78
Tous	33.12	15.14	8	82	33	2617	20,66	27.42	51.92

Par une analyse approfondie, on trouve plus d'opinions positives que d'opinions négatives. Ceci est en grande partie dû à l'utilisateur : on trouve souvent ce genre de biais dans les bases de données non actées ou les personnes ont tendance à être plus positives que négatives. On voit cependant que l'agent qui est joué par les opérateurs a autant d'opinions positives que négatives (22.8% contre 21.1%), ce qui est rassurant étant donné la stabilité des humeurs de l'agent sont fixées.

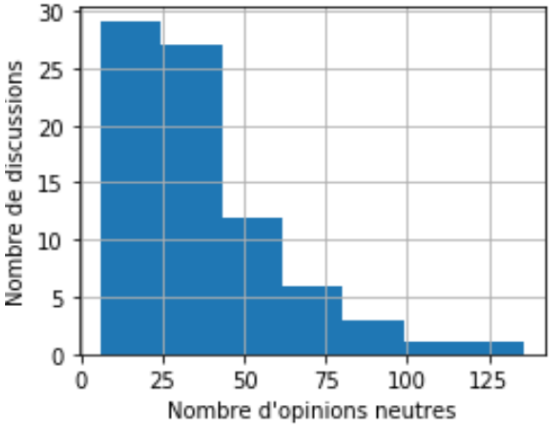
L'agent a un rôle d'influenceur sur l'utilisateur. Premièrement, l'agent a aussi plus d'opinion que l'utilisateur (56.0% contre 47.8%), ce qui est consistant avec la logique du scénario de SEMAINE où l'agent est censé déclencher des émotions et des opinions chez l'utilisateur. Deuxièmement, lorsque l'agent émet des opinions avec une certaine valence, celles de l'utilisateur auront tendance à être de la même valence. Ce genre de phénomène est classique et a été étudié à des niveaux plus complexes comme l'étude du mimétisme dans (Bilakhia et collab., 2015) ou de la contagion émotionnelle dans (Varni et collab., 2017). Sur les sessions de Spike qui est colérique et généralement désagréable avec les sujets de l'expérience, les opinions de l'utilisateur sont beaucoup plus négatives que dans les sessions de Poppy, qui est joyeux. Nous avons remarqué qu'une grande partie des opinions négatives venaient des sessions avec Spike, et c'est généralement des opinions négatives envers l'agent hostile au sujet. Finalement, on peut donc émettre l'hypothèse que l'agent utilise ses propres émotions et opinions

afin d'inciter l'utilisateur à en avoir, cela en plus de poser des questions promptes à déclencher des phénomènes affectifs chez l'utilisateur.

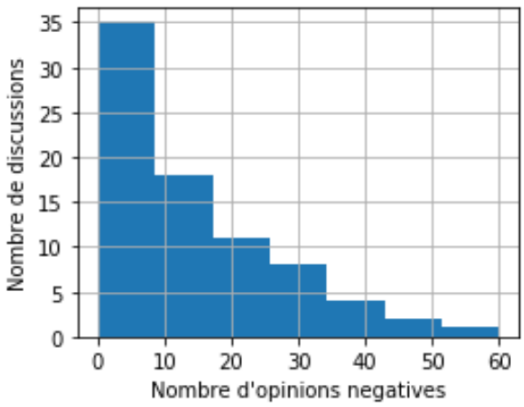
Pour finir, nous avons remarqué que les sessions avec des α faibles sont les sessions ayant le plus d'opinions implicites et difficiles à déterminer, l'annotation devenant encore plus subjective. Une opinion implicite indique, par le choix des mots employés dans un contexte particulier, un degré de subjectivité de la part du locuteur. Par exemple, quand l'agent dit à l'utilisateur qu'il maitrise sa situation telle ou telle situation et que celui répond "Oui autant que je peux" c'est une opinion implicite positive par rapport au fait de maitriser une situation.



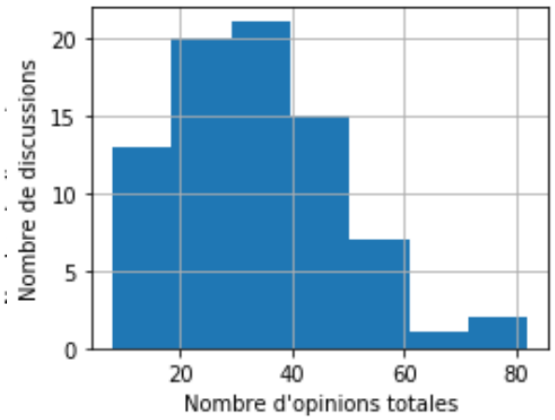
(a) Histogrammes du nombre de TP labellisés positif par discussion



(b) Histogrammes du nombre de TP labellisés neutre par discussion



(c) Histogrammes du nombre de TP labellisés négatif par discussion



(d) Histogrammes du nombre de discussion par nombre d'opinions

FIGURE 5.7: Histogrammes du nombre de discussion par catégorie des TP

On peut observer la répartition du nombre de discussions en fonction du nombre de TP positifs, négatifs ou neutres en Figure 5.7. On peut voir en Figure 5.7c que la

courbe du nombre de discussions en fonction du nombre d'opinions négatives diminue plus rapidement que celle en fonction du nombre d'opinions positives Figure 5.7a. Les interactions avec des opinions négatives ont tendance à être de moins bonne qualité que celles avec un échange d'opinions positives. En effet, comme dit plus haut, lorsque les participants n'échangent que des opinions négatives entre eux dans SEMAINE, c'est souvent dans le cadre d'une interaction avec Spike et synonyme de conflit (comme on peut le voir dans les exemples de la figure 5.8). Lorsque les participants échangent beaucoup d'opinions positives, l'interaction est de bonne qualité et a tendance à durer plus longtemps.

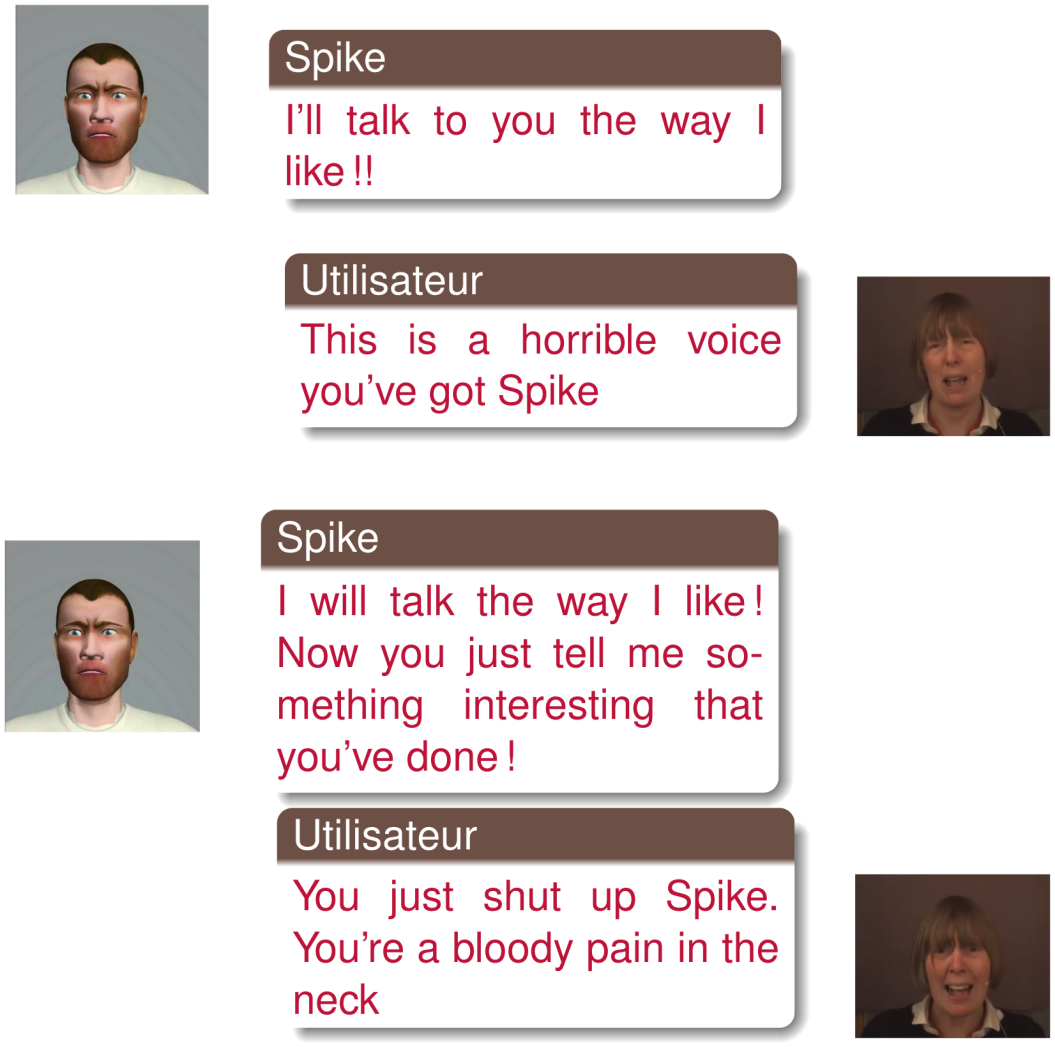


FIGURE 5.8: Exemples de conflits entre l'utilisateur et l'agent : beaucoup d'opinions négatives sont échangées, synonymes d'une très mauvaise interaction

5.3.4 Conclusion

En conclusion, nous avons collecté des annotations en opinions par tour de parole sur le corpus SEMAINE. Nous avons aligné l'audio avec le texte mot par mot en utilisant des transcriptions faites à la main contenant beaucoup de nombreuses informations comme la ponctuation et des phénomènes para-linguistiques. Le corpus est non symétrique car l'agent et l'utilisateur ont des rôles différents, l'agent servant à déclencer une émotion chez l'utilisateur. De plus, le corpus est riche en opinion puisqu'on compte au moins une opinion sur un peu moins de la moitié des TP. Finalement, ce corpus peut être utilisé avec les annotations en émotion continues du challenge AVEC 2011 de Schuller et collab. (2011b).

6 Représentation du signal de parole

Résumé du chapitre

— Nous avons construit un ensemble de descripteurs textuels composé de valeurs provenant de lexiques de subjectivité, de structures syntaxiques et de représentations distribuées.

— Nous avons construit un ensemble de descripteurs audio composé de descripteurs de la qualité de voix issus du modèle de Linjencrant-Fant, et de descripteurs classiques. Nous avons créé des représentations apprises à l'aide d'une méthode utilisant des auto-encodeurs séquentiels.

— Les pauses, en auto-interruptions, sont importantes dans la structure de la parole. Ainsi, nous avons utilisé ces pauses pour segmenter la parole, de manière automatique, en unités inter-pausales. Les descripteurs ont été intégrés sur ces intervalles à l'aide de 13 fonctionnelles statistiques pour l'audio et de la moyenne pour le texte.

La représentation du signal de parole est une tâche difficile. En effet, les modèles statistiques utilisés jusqu'à récemment n'étaient pas assez complexes pour pouvoir utiliser un signal audio brut en entrée et arriver à extraire d'eux-mêmes les informations nécessaires. Depuis de nombreuses années, la technique classiquement utilisée est d'extraire des descripteurs de la voix qui sont jugés importants pour la tâche que le modèle se propose de réaliser (Clavel, 2007).

Avec le temps, la puissance de calcul ainsi que la taille des bases d'apprentissage ont augmenté, ce qui permis de faire apparaître des approches utilisant de grands ensembles de descripteurs et laissant le choix de la sélection aux modèles d'apprentissage, aussi pour l'informatique affective (Schuller et collab., 2009b). Finalement, ces dernières années ont vu apparaître des approches se distinguant de ces pratiques, en apprenant directement les descripteurs depuis le signal de parole brut. Ces approches dites de bout-en-bout (*end-to-end*) qui sont essentiellement neuronales nécessitent une grande quantité de données annotées. Les approches *end-to-end* construisent une représentation du signal de manière intrinsèque durant l'entraînement du modèle à une ou plusieurs tâches précises.

Cependant lorsque l'annotation manque et que de grandes quantités de données non annotées sont disponibles, une alternative efficace est la création de représentation via des méthodes non-supervisées. Cette approche est une forme d'apprentissage par transfert. Il a été montré que l'apprentissage par transfert permettait d'atteindre de hautes performances sur de l'audio ou du texte. L'apprentissage non supervisé a été la norme pendant quelques temps pour les tâches utilisant des données textuelles, notamment avec l'utilisation de représentations distribuées au niveau des mots. Des représentations sont apprises sur de larges corpus et peuvent être utilisées pour une variété de tâches différentes. Pour le texte on teste la qualité des représentations obtenues sur des tâches d'analyse de sentiment, de classification et de traduction Kiros et collab. (2015), alors que pour l'audio ce sont des tâches d'analyse de scènes acoustique ou de détection de style musical Freitag et collab. (2017). Cette méthode allie les avantages de ne pas avoir besoin de beaucoup de données annotées lorsqu'on s'attaque à des tâches non triviales comme la détection d'opinions.

Dans cette thèse, des représentations sont créés avec des descripteurs jugés importants à nos tâches, et à l'aide de larges corpus non annotées, mais aussi des descripteurs plus classiques que ce soit pour le signal vocal (section 6.1) ou le signal verbal (section 6.2). Ces descripteurs sont extraits à des granularités différentes et il est nécessaire de les intégrer sur des segments pertinents avant des les utiliser comme variables d'entrées pour nos modèles d'apprentissage. L'utilisation des pauses pour segmenter le signal a été étudiée, afin d'utiliser ces auto-interruptions comme indices d'unités lexicales intéressantes pour l'analyse de l'opinion (section 6.3).

6.1 REPRÉSENTATIONS TEXTUELLES CHOISIES

Afin d'utiliser le signal textuel dans des modèles d'apprentissage, il est nécessaire de transformer des chaînes des caractères en vecteurs utilisables en entrée de ces modèles. Les représentations de *word embeddings* transforment un mot en un vecteur, ceux de *paragraph embeddings* transforment un groupe de mots en un vecteur. Certaines représentations peuvent contenir plus d'information utile que d'autres, elles peuvent être parcimonieuses ou distribuées, apprises de manière statistique ou bien créées à la main, et avec des dimensions variables.

Si on observe une quantité grandissante d'articles se concentrant sur comment modéliser statistiquement des données textuelles à l'aide de représentations distribuées, des représentations basiques donnent encore de bons résultats. Ces dernières ont l'avantage d'être directement efficaces avec peu de données. Nous avons commencé avec une représentation en Sac de N-grammes classique, mais trouvant que cette représentation manquait de sémantique nous avons changé pour une représentation distribuée contenant des informations sémantiques. Jugeant que les mots à forte valence étaient aussi très importants pour caractériser l'opinion du locuteur nous avons ajouté des lexiques de subjectivités. Afin d'aller plus en profondeur et d'utiliser la syntaxe de la phrase nous avons aussi ajouté des structures linguistiques associées avec le sentiment. Ceci forme ce qu'on appelle dans la suite notre Ensemble de Descripteurs Textuels (EDT). Nous avons finalement utilisé les annotations paralinguistiques disponibles en les regroupant en fonction de leurs similarités.

Nous avons utilisé 4 groupes de descripteurs textuels :

— *Les N-grammes* : Le Sac de N-grammes est une modélisation qui reste efficace pour les documents contenants un ensemble des mot-clés (N-gramme-clés) qui ne sont pas spécialement partagés entre les différentes classes que l'on cherche à discriminer une à une. Les paramètres ont été choisis empiriquement sur un ensemble du corpus ou bien pris dans des travaux antérieurs (Schuller et collab., 2009a). Ces descripteurs sont présentés en Sous-section 6.1.1.

— *Les représentations distribuées* : Avec cette représentation distribuée, les informations sémantiques contenues dans les vecteurs entraînés au préalable sur de grandes quantités de données permettent un apprentissage plus général. Souhaitant un modèle général, nous avons choisi d'utiliser le modèle de mot vecteur word2vec (Mikolov et collab., 2013a) pré-entraîné sur le corpus Google Press avec les paramètres classiques. [1] Ces descripteurs sont présentés en Sous-section 6.1.3.

— *Les lexiques de subjectivité* : La valence d'un document peut être directement extraite à l'aide d'une heuristique utilisant des valeurs spécifiques provenant

1. https ://code.google.com/archive/p/word2vec/

d'un lexique de subjectivité attribuées à chacun des mots. Nous avons utilisé le calculateur d'orientation sémantique de Taboada et collab. (2011), les scores de SentiWordNet (Baccianella et collab., 2010) et les valeurs de valence, activation et dominance du lexique ANEW (Warriner et collab., 2013). Ces descripteurs sont présentés en Sous-section 6.1.2.

— *Les structures linguistiques* : l'extraction des préférences d'un utilisateur peut se faire à l'aide de structures linguistiques. Nous avons utilisé des structures simples liées à l'appréciation afin d'améliorer les résultats de notre système. Ces descripteurs sont présentés en Sous-section 6.1.2.

— *Les descripteurs paralinguistiques* : Les informations paralinguistiques du locuteur peuvent indiquer un état émotionnel qu'il/elle n'évoque pas forcement à travers les mots. Ces annotations sont faites à la main durant la phase de transcription. Ces descripteurs sont présentés en Sous-section 6.1.4.

Nous avons fait des pré-traitements selon les différentes représentations, allant du filtrage des hapax, à l'utilisation d'un dictionnaire Britannique-Américain en passant par un correcteur orthographique (sous-section 5.3.1).

6.1.1 Sac-de-mots et Sac-de-N-Grams

Le Sac-de-mots (*Bag-of-Words* : BoW) introduit par Harris (1954) est une des manières les plus simples de représenter un texte en utilisant tout le vocabulaire employé. Cette méthode permet de représenter les textes par les mots qui le composent, nous la décrirons de manière plus précise ci-dessous. C'est un procédé qui reste beaucoup utilisé, car si très simple, il permet d'obtenir de bons résultats sur des bases de données de tailles raisonnables. Le BoW a été amélioré en Sac-de-N-Grammes (*Bag-of-N-Grams* : BoNG) par Dai et collab. (2003) pour une tâche de classification de texte. Le principe est le même que celui du BoW, mais caractérisant aussi les documents par les suites de N mots les composant (un gram peut aussi être au niveau du caractère).

Définition :

Pour un ensemble de documents D contenant un ensemble T de termes (mots ou N-Grammes), la méthode BoNG créée un vecteur par document $d \in D$, de la longueur de la taille du vocabulaire $|T|$. Chaque coordonnée $i \in [\![1;|T|]\!]$ d'un vecteur est attribué à un unique terme $t \in T$. Pour le modèle classique, la valeur de la coordonnée i représente le nombre d'occurrences du terme t dans le document $d \in D$, appelé la fréquence du terme (*term-frequency*) $tf(t,d)$.

Transformation TF-IDF :

Lors de l'utilisation du BoNG, il est possible d'utiliser d'autres valeurs que la fréquence du terme. Il est possible d'appliquer une transformation appelé TF-IDF (*Term Frequency-*

Inverse Document Frequency). Cela consiste à donner une valeur correspondant à l'importance du terme t dans le document d. En effet, dans l'ensemble des documents de D, certains termes apparaissent plus que d'autres et ne vont pas avoir une valeur significative pour le document. La TF-IDF permet d'obtenir une valeur plus représentative de l'importance d'un terme dans un document donné. Elle peut se calculer comme ceci :

Term-Frequency : $tf(t,d)$ représente le nombre d'occurrences du terme t dans le document d.

Inverse-Document-Frequency : $idf(t,D)$ représente l'inverse de l'occurrence du terme t dans le corpus D. Cela permet de pondérer le poids du mot dans le document par une valeur permettant de donner plus d'impact à des mots rares et non utilisés dans les autres documents, et de débruiter, en diminuant l'impact des mots vides ou anti-mots[2]. Il existe plusieurs IDF, dont $idf(t,D) = \log(\frac{|D|}{|\{d \in D : t \in d\}|})$, qui correspond a $-\log P(t|D)$. Cela fait un parallèle avec la loi de Zipf (Mandelbrot, 1957) qui stipule que l'occurrence d'un mot dans un corpus de texte diminue d'une manière logarithmique.

Il est possible de normaliser chaque vecteur par la norme L2 afin de s'abstraire des longueurs des documents variables.

Faiblesses :

Bien que le BoW soit une représentation fort utile par son ratio efficacité sur simplicité, le BoW a de nombreuses faiblesses. Avec cette représentation, il y a une perte de l'ordre des mots et par exemple les deux phrases "pas vraiment bon" et "vraiment pas bon" auront la même représentation. L'utilisation de N-Grammes permet de changer un peu ceci mais **les dépendances à termes plus long que** N ne sont pas analysées. Le BoNG analyse des chaînes de caractères, la représentation n'apporte pas plus d'information que ce qui est contenu dans la base de données, contrairement à un *embedding* distribué appris sur des données externes. Les mots n'ont pas de concept sémantique, qui peut les lier entre eux : ils sont représentés par de simples *tokens* (symboles) sans signification autre. Pour finir, comme le BoNG n'utilise pas de connaissance extérieure, il est tributaire de la bonne orthographe du transcripteur : il y a perte d'information à chaque erreur.

Calcul :

Les approches BoW et BoNG sont des représentations très brutes : modélisation via la chaine de caractère, très sensible aux typographies, insensible aux similarités des mots, n'ordonnant pas les mots dans la phrase, etc... Il est parfois nécessaire de faire plusieurs pré-traitements semi-automatiques. Ces procédés permettent de réduire le bruit en supprimant les informations inutiles à la tâche et de diminuer la complexité des don-

2. *stop-words* en anglais

nées textuelles à modéliser en réduisant l'espace de dimension. Les pré-traitements que nous avons effectués pour créer les BoNG sont résumés ci-dessous.

Mots vides Certains mots sont tellement communs qu'il est inutile de les indexer ou de les utiliser. Ces mots, spécialement avec l'utilisation d'un BoNG qui ne prend pas en compte l'ordre des mots, représentent principalement du bruit (ex : *and, or, with,...*).

Filtrage des hapax Les mots apparaissant trop peu fréquemment ne donnent pas assez d'information, et ne sont pas assez importants pour faire partie d'une structure à repérer. Ces mots sont appelés hapax.

Racinisation Il est parfois utile de grouper des mots afin qu'ils aient la même représentation, par exemple en groupant singulier et pluriel, car ces mots transportent le même concept. La racinisation est un procédé réduisant le mot à son mot-racine (*stem*) qui est proche de sa racine (*root*). La racinisation est généralement effectuée par un mécanisme simple fonctionnant sur des heuristiques.

6.1.2 Représentations textuelles basées sur une expertise linguistique

Les représentations linguistiques ont l'avantage de pouvoir injecter directement de la connaissance humaine telle qu'une expertise linguistique dès l'initialisation du système d'apprentissage : ces *features* ont été construites à la main et découlent d'une connaissance qui soit spécifique au domaine. Pour notre tâche, cela peut se composer de lexiques de subjectivité et d'opinion récoltés en amont (sous-sous-section 6.1.2.3), de structures linguistiques (sous-sous-section 6.1.2.4), d'un ensemble de règles basées sur la syntaxe de la phrase et possiblement construites à partir d'une connaissance de la tâche à effectuer (sous-sous-section 6.1.2.1). D'un côté, tout ceci permet d'obtenir une grande précision : lorsque des structures linguistiques complexes et précises apparaissent dans un signal textuel, il y a de grandes chances que le phénomène recherché soit présent. D'un autre côté, la structure spécifique de la syntaxe à retrouver est problématique car trop rigide pour une énonciation orale, offrant ainsi un rappel peu élevé. Nous présentons ici les représentations que nous avons utilisées dans nos modèles.

6.1.2.1 Représentations syntaxiques

Les descripteurs linguistiques basiques utilisent les fondations du langage et sont utilisés en linguistique computationnelle depuis des décennies, avant l'avènement des

modèles statistiques complexes nécessitant de plus grandes puissance de calcul et quantités de données. Ces descripteurs sont toujours importants, lorsque la tâche est atypique et les données non disponibles pour un entraînement massif préalable. Insérer dès le départ des structures linguistiques haut-niveau dans un modèle d'apprentissage permet à celui-ci de ne pas à avoir à tout ré-apprendre, mais à utiliser directement de la connaissance humaine pour renforcer son efficacité.

Type de phrase :
La catégorie syntaxique du mot (ou du constituant le plus profond du groupe de mots) dans l'arbre syntaxique. Nous nous restreignons à si la phrase est nominale ou verbale. Nous utilisons le CoreNLP de Schuster et Manning (2016) pour calculer l'arbre syntaxique.

Nature grammaticale :
La nature grammaticale (*Part-of-Speech* : POS) des mots est un descripteur important utilisé dans plusieurs tâches d'analyse de sentiments (Choi et collab., 2005, 2006; Poria et collab., 2015). Par exemple, un énoncé contenant plus d'interjections sera plus caractéristique d'une attitude affective. Nous utilisons le CoreNLP de Schuster et Manning (2016) pour calculer l'arbre syntaxique et les POS.

6.1.2.2 Représentations lexicales

Nous avons utilisé des représentations lexicales précises que nous avons jugé importantes. Plusieurs travaux utilisent les inverseurs de valence (Morency et collab., 2011) qui permettent de gérer les négations ou bien d'avoir un indice simple sur le nombre de négations. De même, les modificateurs de degré sont utiles pour augmenter ou diminuer la puissance d'un sentiment (Kennedy et Inkpen, 2006).

Inverseurs de valence :
Si une négation est présente dans une séquence de mots, cela peut être caractéristique d'une inversion dans la valence de l'opinion, changeant complètement la sémantique de la déclaration du locuteur. Ces négations proviennent d'un lexique simple construit à la main.

Modificateurs de degré :
Nous utilisons le nombre d'intensifieurs et d'atténuateurs à l'intérieur d'une séquence de mots car ils permettent de changer l'intensité d'une expression d'opinion, la renforçant ou la diminuant selon le type de modificateur. Les modificateurs de degré proviennent d'un lexique fourni par Taboada et collab. (2011).

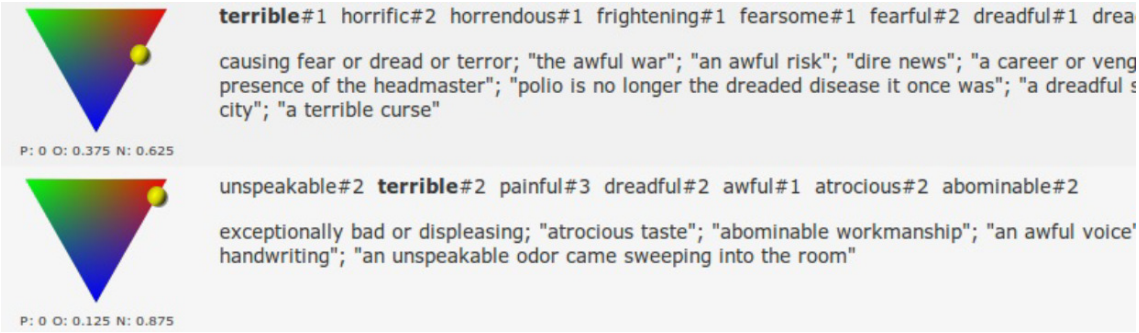


FIGURE 6.1: Exemple de deux *synsets* pour le mot "*terrible*"

6.1.2.3 Représentations lexicales subjectives

Les lexiques de subjectivité peuvent être construits à la main ou bien de manière semi-supervisé. Ce sont des ressources utiles pour l'amélioration des performances des algorithmes d'analyse de sentiment. En effet, il est clair que les valeurs de valence des différents mots utilisés ont un impact fort sur la valence générale exprimée par le locuteur. Les méthodes utilisant des lexiques sont plus robustes et ont de meilleures performances dans les études multi-domaines, car elles dépendent moins de la BDD utilisée pour la création du lexique. De plus, elles peuvent aisément être améliorées avec des ressources extérieures spécialisées.

SentiWordNet :

SentiWordNet de Baccianella et collab. (2010) est une amélioration de WordNet de Fellbaum (1998). Chaque *synset* est lié à un 3-uplet contenant des scores liés à la négativité, la neutralité ou la positivité du mot, procurant 3 descripteurs. Chaque variable est comprise entre 0 et 1, la somme des 3 valeurs étant égale à 1. Des mots peuvent être liés à plusieurs *synsets*. En Figure 6.1, deux significations et scores de sentiment sont disponibles pour le même terme (*terrible* dans le cas présent). Il est possible de remarquer que pour chacune des significations, la somme des scores est égale à 1. Dans le premier cas :

$$pos_{score} + obj_{score} + neg_{score} = 0 + 0.375 + 0.625$$

Lexique ANEW2013 :

Le lexique ANEW2013 de Warriner et collab. (2013) est composé de 13 915 mots et lemmes annotés par 1 827 individus de 16 à 87 ans, comptant 60 % de femmes. Ces valeurs ont été annotées sur différentes échelles : valence, activation et domination. Les scores sont séparés entre genre, niveau d'éducation, et âge de l'annotateur, ou bien tous mélangés.

Les auteurs ayant trouvé que les différences entre les groupes (genre, éducation,

...) étaient présentes mais limitées, nous avons décidé de ne pas prendre en compte les groupes. Nous utilisons la valeur générale, afin d'obtenir des descripteurs les plus généraux pour chaque mot dans notre lexique. Ce lexique nous fait obtenir 3 descripteurs.

6.1.2.4 Patrons et règles linguistiques

Nous avons utilisé des règles associées avec des valeurs provenant de lexiques de subjectivité. Ces règles permettent de former des structures linguistiques pertinentes pour l'étude du phénomène d'opinion. Nous nous sommes aussi intéressés à de simples structures syntaxiques pas forcement liées avec des valeurs de subjectivité.

Orientation sémantique :

Le SO-CAL de Taboada et collab. (2011) est un groupement de lexiques avec un ensemble de règles visant à calculer l'orientation sémantique (*Semantic Orientation CALculator* : SO-CAL) d'un texte. Les lexiques contiennent des mots subjectifs et des modificateurs de degré (amplificateurs et atténuateurs) avec des valeurs associées, ainsi que des inverseurs de valence.

La liste de mots initiale est composée d'adjectifs, de noms, de verbes et d'adverbes, chacun associé à une valeur entre -5 et +5. Le POS est liée à chaque mot dans un but de désambiguïsation : les mêmes termes avec différents POS peuvent avoir des scores distincts. La valeur de chaque mot peut augmenter, diminuer ou bien s'inverser si dans son contexte se trouve un intensifieur, un atténuateur ou un inverseur.

Le modificateur peut multiplier la valeur par un nombre entre 0.5 et 2, et l'inverseur change la valeur par ± 4 (selon le signe initial). Par exemple, dans la phrase "*It is not good*", le terme *not* annule le terme *good* qui a une valeur de SO de 3. La valeur de SO finale devient :

$$SO(\text{not great}) = SO(\text{not}) + SO(\text{good}) = -4 + 3 = -1$$

Il est important d'utiliser les négations, les intensifieurs et les POS des mots avec des structures linguistiques afin d'obtenir un score global représentatif de l'énoncé. La combinaison des modificateur de degré avec les inverseurs de valence fournit un outil puissant qui peut changer drastiquement la valence totale d'une opinion dans une phrase. Par exemple dans la phrase "*The movie ain't that bad*", on est bien en présence de l'intensifieur "*that*", de la négation "*ain't*" et du mot de valence négative "*bad*".

$$SO(\text{The movie ain't that bad}) = SO(\text{ain't}) + SO(\text{that}) \times SO(\text{bad}) = 4 - 1.5 \times 3 = -0.5$$

Structures simples :

Nous avons utilisé des structures simples, dans le but d'utiliser des premiers descripteurs syntaxiques qui sont souvent utilisés dans les expressions d'opinion.

Les structures relevées qui nous donnent 5 descripteurs booléens sont :

— un adjectif suivi d'un nom : "*awesome plot*" (présence ou non) ;
— un adverbe suivi d'un verbe : "*badly directed*" (présence ou non) ;
— un adverbe suivi d'un adjectif : "*really good*" (présence ou non) ;
— la présence d'une négation dans le contexte d'un mot de valence : "*not good*" (valence inversée de ANEW et SWN) ;
— la présence d'un modificateur de degré dans le contexte d'un mot de valence : "*slightly good*"(valence modifiée de ANEW et SWN).

6.1.3 Représentations apprises

Récemment, l'intérêt pour les méthodes non supervisées d'apprentissage de représentations distribuées a fortement grandi. Une représentation distribuée (Bengio et collab., 2003) est une représentation sous forme d'un vecteur dense de valeurs réelles encodant la sémantique de l'unité linguistique. Jusqu' il y a peu, ces genres de représentations étaient difficiles à obtenir car la puissance calculatoire demandée n'était pas acquise. De nos jours, l'intérêt général et les avancées en apprentissage profond ont permis de se soustraire à ces handicaps. Par exemple, un réseau obtenant des *embeddings* comme celui de Collobert et Weston (2008), a pris 2 mois pour s'entrainer en 2008 alors qu'aujourd'hui il ne lui faudrait même pas une heure sur une machine classique.

Dans nos travaux, nous avons utilisé la représentation word2vec de Mikolov et collab. (2013a).

Introduction à word2vec :

Word2vec est un modèle d'apprentissage permettant de représenter des mots sous forme de vecteurs denses à valeurs réelles dans un espace de faible dimension Mikolov et collab. (2013a). Le principe du modèle est que les mots sont définis par leurs champs sémantiques, c'est-à-dire par le contexte dans lequel ils sont employés. Le principe de l'apprentissage est d'utiliser tous les mots apparaissant dans le contexte proche de chaque mot, afin de pouvoir définir ce dernier de manière générale. Deux types d'apprentissages sont proposés :

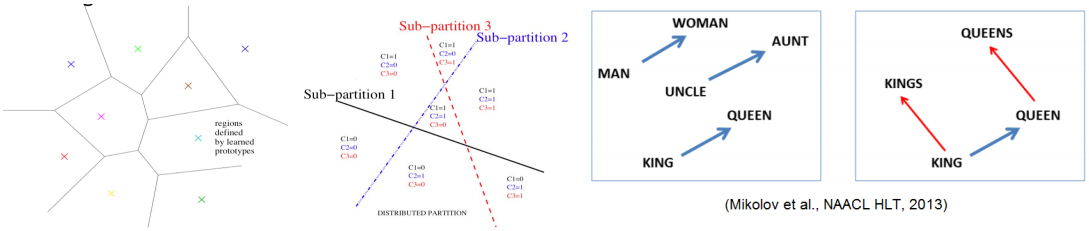
Le *Skip-Gram* consiste à apprendre des représentations de mots performantes à retrouver les mots qui apparaissent dans leur contexte gauche et droite.

Le *Continuous-Bag-Of-Words* consiste à apprendre des représentations de mots performantes à retrouver un mot, sachant les mots de son contexte gauche et droite.

Nous avons choisi d'utiliser word2vec car, lors du démarrage de cette thèse, ce modèle obtenait de meilleurs résultats sur les tâches d'analyse de sentiments Poria et collab. (2015) comparé a d'autres modèles comme GloVe de Pennington et collab. (2014) ou C&W de (Collobert et collab., 2011). Il est apparu empiriquement qu'un corpus d'entrainement plus grand et plus général permettait d'obtenir des vecteurs qui atteignaient de meilleurs résultats dans différentes tâches. L'entrainement est fait avec un *Skip-gram* que nous allons détailler ci-dessous

Entrainement : Le modèle *Skip-gram* et échantillonnage négatif :
Nous avons utilisé des vecteurs obtenus avec entrainement de type Skip-gram avec échantillonnage négatif. Les détails de l'entrainement sont données en annexe E.

Intérêts de cette représentation :
Le principe des représentations distribuées est d'utiliser les différentes propriétés communes des objets représentés afin de les caractériser. En utilisant la compositionnalité des différents objets l'avantage peut être exponentiel, ainsi illustré dans les sous-figures 6.2a et 6.2b où l'utilisation d'hyperplan permet de segmenter l'espace en une partition de taille exponentielle. Les représentations distribuées peuvent apprendre des relations, des similarités, et généralisent très bien (Hinton et collab., 1986).



(a) Partition locale (b) Partition distribuée (c) Relations dans l'espace word2vec

FIGURE 6.2: Différents intérêts des représentations distribuées

L'intérêt de cette représentation est l'utilisation du champ sémantique : les mots transportant les mêmes concepts sémantiques obtiendront des vecteurs similaires (voir Figure 6.3). Ainsi les vecteurs des mots *Paris*, *Londres* et *Berlin* seront proches dans l'espace word2vec car ce sont des capitales de grands pays européens. *Cat* et *dog* seront proches l'un de l'autre, mais *dog* sera plus proche de *wolf*. On constate généralement que les hyperonymes et les hyponymes sont proches dans cet espace, ce qui est logique car l'un contient l'autre.

Si la polysémie reste un problème, elle est en partie prise en compte car le mot sera apparu pour ces deux sens. Le vecteur utilisera aussi les deux sens, même si une désambiguïsation lexicale permettra d'avoir de meilleurs résultats.

Un intérêt de cette représentation est que notre classifieur n'apprend plus une chaine de caractères, mais une forme dans l'espace word2vec : il apprend des concepts et non

plus des mots. Si un mot peu utilisé ou utilisé uniquement dans l'ensemble d'apprentissage véhicule le même concept qu'un autre mot beaucoup utilisé, alors le classifieur sera capable de reconnaitre que le mot de l'ensemble de test est similaire à celui qu'il a déjà beaucoup vu.

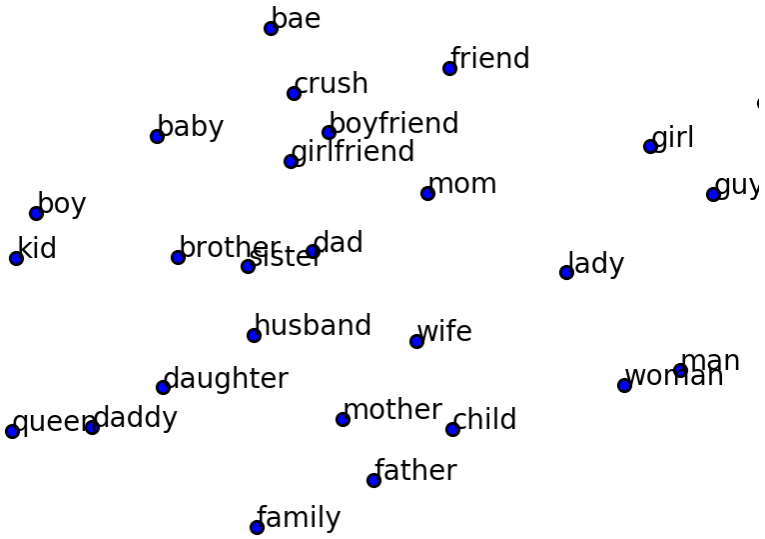


FIGURE 6.3: t-SNE pour visualisation de mots-vecteurs. Les mots les plus proches sémantiquement sont proches dans l'espace.

Implémentation utilisée :
Nous avons utilisé une version des mots-vecteurs pré-entrainé sur un corpus d'article *Google Press* composé de 100 milliards de mots [3] et 100 millions de documents, mis à disposition par les auteurs. Les mots-vecteurs ont été construits sur le modèle du *Skip-gram* avec une taille de dimension des mots-vecteurs de 300, une fenêtre de contexte de la taille de la phrase, entrainé avec un softmax hiérarchique

6.1.4 Représentations paralinguistiques

Dans les différents corpus que nous avons utilisés lors de cette thèse, des annotations paralinguistiques étaient présentes. Il a été trouvé dans de nombreux travaux que l'utilisation de ces descripteurs est efficace pour caractériser des traits de personnalité et des émotions.

Présence de disfluence verbale :
Nous avons choisi de noter la présence de disfluence verbale dans notre représentation

3. https ://code.google.com/archive/p/word2vec/

car il représente une non-fluidité du discours et peut être un indice d'opinion ou de doute. Nous avons créé un lexique de disfluences afin de pouvoir utiliser toutes les manières dont elles ont étés transcrites.

Représentations paralinguistiques transcrites :
Les annotations paralinguistiques n'étaient pas normalisées. Contrairement aux disfluences qui pouvaient être facilement normalisées, il était plus difficile de créer des sous-groupes pour les annotations transcrites par un ou plusieurs mots. On trouvait des annotations différentes qui reflétaient le même phénomène, comme : *laughs*, *laughing* et *laugh*. Il a donc été necessaire de leur appliquer un pré-traitement pour pouvoir les utiliser de manière efficace dans le modèle d'apprentissage.

Afin de normaliser les paralinguistiques, nous avons utilisé technique applicable pour toute sorte de données, elle permet de rassembler des expressions muti-mots en clusters. Nous avons rassemblés les annotations qui se rapprochaient le plus en fonction de la plus longue chaine de caractère en commun de leur racine.

Nous avons tout d'abord passé un algorithme de racinisation sur le texte de chaque annotation paralinguistique afin d'obtenir une écriture simplifiée composé uniquement des racines des mots. Puis toutes les annotations ayant une chaine de caractères commune de taille supérieur ou égale à 5 furent fusionnées. Finalement, les annotations ayant moins de 10 occurrences sur la BDD furent éliminées.

Par exemple, les annotations *lip smack*, *lip smacking*, *smacking lips* et *smacking* ou *coughing*, *cough* et *cough cough* ou encore *knocking*, *wood knocking* et *knock* furent fusionnées en trois uniques catégories.

6.1.5 Tableau récapitulatif

Dans un but de simplification, les différents descripteurs utilisés pour représenter le signal textuel sont résumés dans le tableau 6.1.

Nous avons fait le choix de prendre pour nos descripteurs un ensemble hybride permettant d'allier la précision des descripteurs experts avec la robustesse des descripteurs appris sur les données. Notre ensemble est constitué à la fois de descripteurs experts extraits à la main comme des lexiques de subjectivités (sous-section 2.1.1) et des descripteurs issus de règles syntaxiques (sous-section 2.1.2), combiné à des représentations distribuées apprises sur de grandes quantités de données (voir sous-section 2.1.3).

6.2 REPRÉSENTATION AUDIO CHOISIE

Dans ce travail nous avons choisi d'utiliser des descripteurs acoustiques classiques extraits ayant été éprouvés de nombreuses fois pour des tâches liées à l'analyse de sen-

TABLE 6.1: Descripteurs du signal textuel

Type	**Descripteur**	**Taille**	**Granularité**	**Exemple** : *alarm*
Sémantique	word2vec	300	mot	[0.45 0.2 ... 0.92]
Syntaxique	POS	7	mot	NN
Lexical	ANEW	3	mot	Valence : 3.9 Activation : 6.9 Domination : 6.6
Lexical	SWN	3	mot	Positif : 0.375 Neutre : 0.375 Négatif : 0.25
Syntaxico-lexical	Orientation Sémantique	3	structure	-2.5
Syntaxique	Négation	1	mot	non
Syntaxique	Intensifieur	1	mot	non
Paralinguistique	Paralinguistique	X	évènement	*laugh*, *caugh*, *smack*, etc...
Paralinguistique	Disfluence verbale	1	évènement	$\mathbb{1}$(mot = '*hum*')

timents à l'oral (sous-sections 6.2.1, 6.2.2 et 6.2.3), ainsi qu'une représentation apprise de manière non supervisée avec une méthode neuronale (sous-sections 6.2.5). Dans cette section nous présenterons les différents descripteurs audio que l'on a choisi et qui constitue notre Ensemble de Descripteurs Acoustiques (EDA) pour représenter le signal de parole.

Les descripteurs extraits peuvent être regroupés dans 3 catégories différentes :

— *Les descripteurs cepstraux et spectraux* : Ils sont caractéristiques des fréquences et de l'enveloppe spectrale du signal de parole en général. Nous avons calculés les 12 premiers MFCC ainsi que le 0-ième coefficient lié à l'énergie et les 3 premiers formants.

— *Les descripteurs de la prosodie* : Ils sont caractéristiques de la structure de la parole, et responsables de l'intelligibilité lorsqu'on modélise la parole. Ces descripteurs sont adéquates pour la modélisation de l'état émotionnel du locuteur. La fréquence fondamentale ainsi que la sonie (*loudness*) sont calculées.

— *Les descripteurs de qualité de voix* : Ils sont en majorité caractéristiques de la configuration glottale et par conséquent des changement physiologiques du locuteur. Les émotions étant fortement liées aux changements physiologiques comme le rythme cardiaque ou la tension dans le conduit vocal (Scherer, 2005), il est important de compter ces paramètres dans notre EDA. Ces paramètres, qui ont été utilisés pour de multiples études liées à l'analyse de sentiment, sont principalement tirés du modèle du flux glottal de Linjencrant-Fant (Fant, 1995) et liés aux instants de fermeture glottale. Nous avons calculé le Jitter, le Shim-

mer, le ratio de segments non voisés, le rapport harmonique sur bruit (*HNR*), le quotient quasi-ouvert (*QOQ*), le quotient du maximum de dispersion (*MDQ*), la pente d'après pic (*Peakslope*), le quotient d'amplitude normalisé (*NAQ*), le paramètre spectral parabolique (*PSP*), le coefficient de relaxation (*Rd*) et le spectre en bande réduite (*H1H2*).

6.2.1 Descripteurs cepstraux et spectraux

6.2.1.1 Formants

Les formants sont caractéristiques de l'intelligibilité, ce qui les rend très utiles dans la description du signal de parole pour des tâches liées à l'analyse de sentiments.

Définition :
Les formants sont les harmoniques du signal de parole, et définis par l'*Acoustical Society of America* pour un son complexe, comme "des valeurs de fréquence à l'intérieur duquel il y a un maximum absolu ou relatif dans le spectre sonore". Ils sont caractéristiques de l'enveloppe spectrale et ils décrivent les premiers maxima spectraux du signal vocal. Dans la théorie Source-Filtre du signal vocal, les formants sont les fréquences qui sont intensifiés par le filtre symbolisant le conduit vocal. On appelle FN le formant ayant la N^{ième} amplitude la plus forte (F1, F2, etc...).

F1, F2 et F3 sont importants pour déterminer la qualité phonémique d'un signal de parole donné, et sont représentatifs du son émis. Ils sont très utiles pour la détection de phonème et les systèmes de reconnaissance automatique de la parole par exemple. Les formants plus élevés comme F4 et F5 sont aussi connus pour être importants dans la détermination de la qualité de la voix.

Calcul :
Les formants sont calculés via un algorithme de suivi de formant à l'état de l'art (Bozkurt et collab., 2004). L'algorithme est basé sur un traitement de la dérivée négative de l'argument de la transformé en z (*chirp-z transform*).

6.2.1.2 Mel Frequency Cepstral Coefficients (MFCC)

Les MFCC sont des représentations classiques dans plusieurs domaines liées à l'analyse de la parole. Lorsqu'on regarde du point de vue physiologique, l'origine de la voix provient des vibrations des cordes vocales créant des ondes qui passent à travers le conduit vocal. Ceci peut être modélisé par une émission d'onde depuis une source et passant à travers un filtre. Le Cepstre d'un signal permet de séparer le signal en ses deux composantes grâce une transformation mathématique habile (voir Équation 6.1).

L'échelle MEL est ensuite utilisée afin d'obtenir des valeurs reflétant la perception humaine du son. L'ensemble de ces procédés font des MFCC de très bons descripteurs de la parole en générale et de très bon coefficients pour obtenir des informations sur l'état émotionnel d'un locuteur et c'est pour cela que nous les utilisons aussi.

Définition :

Les MFCC sont caractéristiques de l'enveloppe d'un signal. Le Cepstre d'un signal est l'inverse de la transformé de Fourier inverse du logarithme de la valeur absolue de la transformé de Fourier d'un signal :

$$(\tau) = \mathscr{F}^{-1}[\log(|\mathscr{F}[s(t)]|)](\tau) \tag{6.1}$$

où $\mathscr{F}[s(t)] = \int_{t\epsilon\mathbb{R}} s(t)\exp(-2i\pi f t)dt$ est la transformé de Fourier de s, et $\mathscr{F}^{-1}$ l'inverse de la transformé de Fourier.

Comme on peut voir le signal de parole comme une émission suivi d'un filtrage, sous avons $s(t) = g * h(t)$

Cela permet de séparer les composantes comme montré dans l'équation (6.2).

$$c(\tau) = \mathscr{F}^{-1}[\log|G(f)|](\tau) + \mathscr{F}^{-1}[log|H(f)|] \tag{6.2}$$

Où τ est un paramètre homogène au temps appelé quéfrence. Les coefficients cepstraux correspondant(s) aux basses quéfrences représentent en majeure partie la contribution du filtre D'Alessandro (2002).

Ensuite, on intègre avec l'échelle MEL qui est une échelle logarithmique caractéristique de la manière dont les humains perçoivent les sons (voir Équation 6.3).

$$m(f) = \frac{1000}{\ln(1+\frac{1000}{700})}\ln(1+\frac{f}{700}) \tag{6.3}$$

Le spectre $S(k)$ d'un signal $s(t)$ est intégré en bandes Mel afin d'obtenir un spectre avec une amplitude modifiée $\tilde{S_k}$,pour $k = 1...K$ qui représente l'amplitude de la $k^{ième}$ bande Mel.

Finalement, on applique une Transformé en Cosinus Discret (TCD) au spectre intégré en bande Mel (équation 6.4

$$\tilde{c}(\tau) = \sqrt{\frac{2}{k}}\sum_{k=1}^{K}\log(\tilde{S_k})\cos(\tau(k-1/2)\frac{\pi}{K}) \tag{6.4}$$

Calcul :

Nous avons choisi dans notre cas d'intégrer le signal à l'aide de 12 bandes Mel pour avoir 12 coefficients Mel. Pour obtenir les MFCC, l'enveloppe est calculée avec la méthode décrite dans Röbel et Rodet (2005), suivi d'une distorsion des fréquences avec la

méthode de Tokuda et collab. (1994).

6.2.2 Descripteurs de la prosodie

Ils sont caractéristiques de la structure de la parole, et responsables de l'intelligibilité lorsqu'on modélise la parole. Ces descripteurs sont adéquats pour la modélisation de l'état émotionnel du locuteur. La fréquence fondamentale ainsi que la sonie (*loudness*) sont calculées.

6.2.2.1 Le Pitch et les parties voisées et non voisées

Dans la voix, les parties voisées et non voisées d'un signal sont significatives de son contenu et des informations portées par celui-ci, qu'elles soient sur le contenu linguistiques ou sur l'état émotionnel du locuteur.

Définition :
En traitement automatique de la parole, la fréquence fondamentale (ou pitch) caractérise les parties voisées du signal de parole et est liée avec l'impression de hauteur dans la voix. Sur les parties voisées, le signal peut être modélisé comme une superposition d'un bruit blanc avec un signal périodique de période T. Le pitch F_0 est l'inverse de ce T, $F_0 = \frac{1}{T}$.

Notre représentation du signal contient aussi le ratio de parties voisées sur celui de parties non voisées.

Calcul :
Le calcul du pitch est effectué par l'algorithme de somme des harmoniques résiduels de Drugman et Alwan (2011). Cette méthode utilise la structure harmonique du signal résiduel de la prédiction linéaire afin d'estimer à la fois F_0 et les bornes du signal voisé. Cette méthode s'est montrée être très robuste au bruit.

6.2.2.2 Énergie (MFCC0)

L'énergie d'un signal est l'intégration du carré d'un signal sur le temps, aussi égal à l'intégration du carré de sa transformé de Fourier sur la fréquence (voir Équation 6.5).

$$E = \int_{\mathbb{R}} |s(t)|^2 dt = \int_{\mathbb{R}} |S(f)|^2 df \tag{6.5}$$

Définition :

Le coefficient MFCC0 caractérise l'énergie du signal et est l'équivalent de l'offset.

$$\tilde{c}(0) = \sqrt{\frac{2}{k}} \sum_{k=1}^{K} \log(\tilde{S}_k) \tag{6.6}$$

où S est la transformé de Fourier de s.

Calcul :
Même chose que pour les MFCC (voir sous-sous-section 6.2.1.2).

6.2.2.3 Sonie

La sonie est utilisée comme une alternative plus pertinente à l'energie du signal car représentant la perception. Dans le but d'approximer la perception non linéaire du son qu'ont les humain, un spectre auditif est appliqué dans la technique de Prediction Linéaire Perceptive de Hermansky (1990) qui est adoptée.

Définition :
La sonie correspond à la perception de la variation de l'amplitude de signal de parole causée par une énergie plus ou moins forte provenant du diaphragme et provoquant une variation de la pression de l'air sous la glotte. Ce descripteur permet de fournir une mesure de la force sonore perçue dans la voix du locuteur.

Calcul :
Un spectre non linéaire en bandes Mel est construit en appliquant 26 filtres triangulaires de 20 à 8000 Hz sur le spectre calculé sur des fenêtres de 25ms. Une pondération par une courbe isosonique est appliquée avant de sommer les racines cubiques de chaque bande pour obtenir le spectre auditif final. La sonie est calculée comme la somme sur toute les bandes du spectre auditif. Les détails se trouvent dans (Eyben et collab., 2015).

6.2.3 Descripteurs de qualité de voix du flux glottal

Le modèle du flux glottal est décrit de manière plus détaillé dans la sous-sous-section 2.2.1.3 traitant des descripteurs de la qualité de voix.

6.2.3.1 Le Paramètre Spectrale Parabolique (*Parabolic Spectral Parameter* : PSP)

Le PSP est une mesure liée au flux glottal qui permet une comparaison des flux glottaux en terme de diminution spectrale, même quand la F_0 est différente (Alku et collab.,

1997).

Définition :
Le PSP d'Alku et collab. (1997) est obtenue en approximant les fréquences faibles du flux glottal (voir Sous-sous-section 2.2.1.3) avec une fonction parabolique. Le PSP est une valeur numérique donnant une idée de comment se comporte la décomposition du spectre par rapport à la limite théorique correspondant à la décomposition du spectre maximale.

Calcul :
Après avoir calculé le spectre discret, il faut estimer celui-ci à l'aide d'un polynôme d'ordre 2 : $Y(k) = ak^2 + b$. Pour cela, il faut minimiser l'erreur sur un intervalle de fréquences discret où l'erreur est $E_N = \sum_{k=0}^{N-1}(X(k) - ak^2 - b)^2$. La fonction parabolique optimale est obtenue en cherchant les dérivées nulles de E par rapport à a et b. La valeur de N est trouvée quand l'erreur normalisé $NE = \frac{E_N}{\sum_{k=0}^{N-1} X(k)^2}$ devient supérieure à 0.01.

Finalement, le PSP est obtenu en normalisant la valeur de a par la décomposition spectrale maximale théorique $PSP = \frac{a}{a_{max}}$ (le a_{max} sort du modèle LF, voir Sous-sous-section 2.2.1.3).

6.2.3.2 Le rapport harmonique sur bruit (*Harmonic to Noise Ratio* : HNR)

Le HNR permet de caractériser la contribution sonore de la parole durant l'effort vocal. Le son de la voix perçu est dû à des oscillations irrégulières des cordes vocales et à du bruit ajouté.

Définition :
Le HNR représente le ratio de l'énergie des parties harmoniques du signal sur les parties bruitées et s'exprime en dB. Le HNR logarithmique est pallié a -100 dB afin d'éviter les valeurs très négatives.

Calcul :
Le HNR est estimé via la fonction d'autocorrelation à court terme (*AutoCorrelation Function* : ACF) (fenêtre de 60ms) comme le rapport de l'amplitude de l'ACF à la fréquence fondamentale et de l'énergie totale de la fenêtre exprimée en dB. C'est ainsi qu'il est exprimé par Schuller (2013) (voir Équation 6.7).

$$HNR_{acf,log} = 10log_{10}\frac{ACF_{T_0}}{ACF_0 - ACF_{T_0}}\text{dB} \tag{6.7}$$

Où ACF_{T_0} est le $0^{\text{ième}}$ coefficient de l'ACF (équivalent de l'énergie quadratique de la fenêtre).

6.2.3.3 Le quotient de quasi-ouverture (*Quasi-Open Quotient* : QOQ)

Le QOQ de Hacki (1989) qui décrit le temps d'ouverture de la glotte est un paramètre physiologique dans la production du signal vocal. Il est connu pour être utile à discriminer les voix tendues des voix soufflées (Scherer et collab., 2014).

Définition :
Dans le modèle LF qui est un modèle temporel de paramétrisation du flux glottal, il est possible de définir 2 paramètres basiques : le quotient d'ouverture (*Open Quotient* : OQ) et le quotient de fermeture (*Close Quotient* : CQ). Le quotient d'ouverte (respectivement de fermeture) est le rapport du temps entre la phase d'ouverture (respectivement de fermeture) de la pulsation glottale et la longueur de la période fondamentale (pour plus de détails, voir Sous-sous-section 2.2.1.3).

Calcul :
On mesure le QOQ en détectant le pic d'amplitude du flux glottal, puis en trouvant les 2 points de la courbe où l'amplitude est à 50% de l'amplitude maximale. La distance entre ces 2 points divisée par la période glottale locale donne le QOQ (voir Équation 6.8). Ainsi, le QOQ se calcule directement de l'estimation du flux glottal pour chaque instant de fermeture glottale.

$$QOQ = \frac{T_2 - T_1}{T_0} \tag{6.8}$$

Où T_2 et T_1 sont les temps où l'amplitude du flux glottal est à 50% de l'amplitude maximale, et T_0 la période fondamentale.

6.2.3.4 Le quotient d'amplitude normalisé (*Normalized Amplitude Quotient* : NAQ)

Le NAQ est un paramètre représentatif de la phase de fermeture glottale. Il est utile pour mesurer la qualité de la voix (quel est le ton de l'élocution) et peut se montrer efficace pour notre problème.

Définition :
Le NAQ a été développé par Alku et collab. (2002) et représente l'amplitude que prend le flux glottal (Équation 6.9). Il a été montré que le NAQ est robuste aux perturbations sonores et c'est pour cela qu'il a été utilisé par Campbell et Mokhtari (2003) pour l'analyse de la parole, qui est fréquemment bruitée.

$$NAQ = AQ * F_0, \quad \text{where} \quad AQ = \frac{f_{ac}}{d_{peak}} \tag{6.9}$$

Où AQ est le quotient d'amplitude, f_{ac} est l'amplitude maximale de flux glottal,

d_{preak} est la dérivée du flux glottal représentant sa diminution et F_0 la fréquence fondamentale.

Calcul :
Le NAQ est directement calculé depuis l'estimation du flux glottal pour chaque instant de fermeture glottale.

6.2.3.5 La différence en amplitude entre les premiers harmoniques (H1H2 et H1A3)

H1H2 est la différence d'amplitude (en dB) entre les deux premiers harmoniques du spectre en bande étroite de la source du modèle vocal (Titze et Sundberg, 1992). H1H2 est connue pour être fortement corrélée avec le OQ (voir plus haut) et est ainsi utile pour la discrimination d'une qualité de voix allant de tendue à haletante (Airas et Alku, 2007). À cela, on rajoute la différence entre l'amplitude de la fréquence fondamentale et celle de la deuxième harmonique appelée H1A3.

Définition :
H1H2 et H1A3 sont les amplitudes des premiers formants normalisés par l'amplitude du pic spectral à la fréquence fondamentale F_0. Plus précisément, ces valeurs sont calculées comme le rapport de l'amplitude du plus haut pic harmonique dans $[0.8F_i; 1.2F_i], i = 1,2$ sur l'amplitude du pic spectral à F_0.

Calcul :
Ces valeurs sont calculées de la même manière que les formants (voir sous-section 6.2.1.1).

6.2.3.6 Coefficient de relaxation (R_d)

Le coefficient de relaxation R_d est significatif de combien la voix est relaxée (comme son nom l'indique).

Définition :
R_d est le paramètre de l'enveloppe du modèle LF [4] de Fant (1995) (pour plus de détails, voir en annexe D), il est calculé avec une valeur de confiance associée à son extraction. Tahon et collab. (2012) ont étudié ce paramètre comme représentation du signal pour une tâche de reconnaissance de l'émotion et préconisent de l'utiliser uniquement pour

4. Si R_d et le NAQ paraissent similaires, la manière de les calculer est tout à fait différente (Scherer et collab., 2013). En effet, le NAQ est mesuré directement sur le flux glottal alors que R_d vient d'un modèle LF issu d'un apprentissage ayant des défauts. Par exemple, les amplitudes de la source vocale sont généralement pauvrement modélisées.

une valeur de confiance dans le paramètre R_d haute (>0.7).

$$R_d = \frac{UP.F_0}{E_e.110} \tag{6.10}$$

Où UP est l'amplitude et E_e la pente dans la figure D.1 en annexe D.

Calcul :
Le calcul de R_d se fait comme décrit dans Degottex et collab. (2011) à l'aide de l'opérateur de différence de phase du second ordre. Le principe général est d'utiliser la fonction de distorsion de phase.

6.2.4 Descripteurs de qualité de voix provenant d'ondelettes

6.2.4.1 PeakSlope

Le Peakslope de Kane et Gobl (2011) est une mesure de tension dans la voix, lorsque le descripteur atteint de hautes valeurs négatives. Il a été montré robuste aux bruits de bavardages ajoutés avec un ratio de signal sur bruit aussi faible que 10 dB.

Définition :

Le PeakSlope est la pente d'une droite obtenue à l'aide des valeurs maximums de la décomposition en ondelettes du signal de parole pour différentes fréquences.

Calcul :
On utilise l'équation de l'ondelette mère (Éq. 6.11) pour la décomposition du signal :

$$g(t) = -\cos(2\pi f_n t) \,.\, \exp(-\frac{t^2}{2\tau^2}) \tag{6.11}$$

où $f_s = 16kHz$ est la fréquence d'échantillonnage, $f_n = \frac{f_s}{2}$ et $\tau = \frac{1}{2f_n}$.
Le signal de parole est ensuite convolué avec $g(\frac{t}{s_i})$ pour obtenir la transformée en ondelette du signal $x(t)$ à l'échelle i :

$$\hat{x}_i(t) = x(t) * g(\frac{t}{s_i}) \tag{6.12}$$

où $s_i = 2^i$, pour $i = 0,..,5$ qui donne une banque de filtres en bandes d'octave autour des fréquence 8, 4, 2, 1, 0.5, 0.25 kHz. Le maximum local sur une certaine frame est ensuite calculé à toutes les échelles. Le Peakslope est le coefficient de cette régression linéaire (Figure 6.4).

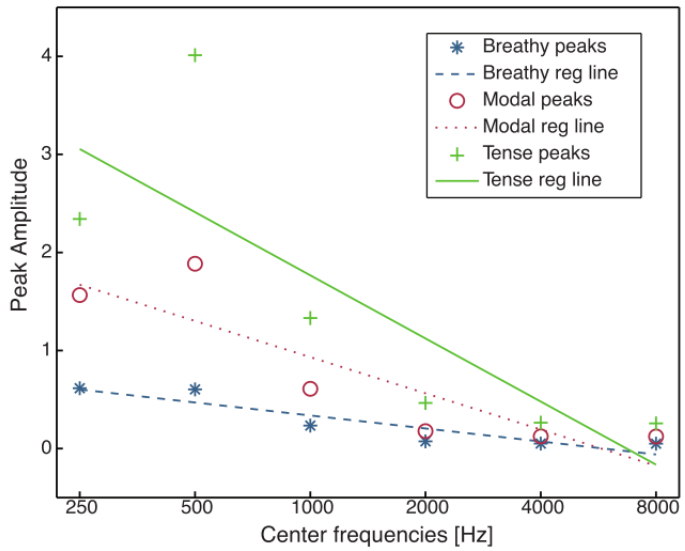


FIGURE 6.4: Amplitudes des pics par rapport aux transformées en ondelettes, et les régressions associées de Scherer et collab. (2013)

6.2.4.2 Quotient de Dispersion Maximum (*Maxima Dispersion Quotient* : MDQ)

Le MDQ de Kane et Gobl (2013b) est une valeur significative d'une voix tendue ou soufflée, selon la distance entre le maximum de la transformée en ondelette et l'instant de fermeture glottale.

Définition :

Le MDQ de Kane et Gobl (2013b) reflète la dispersion entre le maximum de la transformée en ondelette du signal et l'instant de fermeture glottale (*GCI* : Glottal Closure Instant). Les voix tendues ont une fermeture de glotte brusque, le maximum de la transformée en ondelettes est proche du GCI.

Calcul :

Premièrement, il est nécessaire de calculer les résidus de la prédiction linéaire avec un codage prédictif linéaire d'ordre permettant d'avoir 2 coefficients par formant , puis leurs transformées en ondelettes y_i.

Ensuite, la méthode SE-VQ de Kane et Gobl (2013a) permet de trouver les GCI, qui serviront pour la création de l'intervalle de l'équation (6.13).

$$int = [GCI(p) - T_0.c\,;\, GCI(p) + T_0.c] \tag{6.13}$$

Où p est l'index de la GCI et c une constante significative de la longueur de la fenêtre relativement à la période glottale dans lequel sera cherché le maximum m_i de la

transformée en ondelette y_i (Éq. 6.14).

$$m_i = \arg\max_{y_i}\{y_i(int)\} \tag{6.14}$$

C'est finalement la distance entre les m_i et le GCI qui seront utilisés pour obtenir le MDQ, avec une moyenne sur les différentes bandes et une normalisation par la période glottale (Eq. 6.15)

$$MDQ(p) = \frac{\frac{1}{K}\sum_{i=0}^{K-1}|GCI(p) - m_i|}{T_0(p)} \tag{6.15}$$

6.2.4.3 Extraction

Nous avons extrait ces descripteurs acoustiques toutes les 10 ms via les boites à outils COVAREP (Degottex et collab., 2014) et OpenSMILE (Eyben et collab., 2013) en utilisant des paramètres d'extraction classiques décrits dans (Eyben et collab., 2015) et (Degottex et collab., 2014).

Intégration : Afin d'obtenir des vecteurs de taille constante par observation, les descripteurs ont été intégrés temporellement à l'aide de 13 fonctionnelles étant : les centiles 10, 25, 50, 75 et 90, la différence entre les centiles 10 et 90, 25 et 75, la différence entre le maximum et le minimum, la moyenne arithmétique, la déviation standard, l'asymétrie, la pente et le biais de la régression linéaire. Ces intégrations sont faites sur les unités obtenues après les différentes segmentations que nous avons étudiés, allant d'une unité inter-pausale à un tour de parole.

6.2.5 Représentations apprises

Un autre moyen d'obtenir des représentations du signal de parole est d'entraîner un modèle à générer ces représentations de manière supervisée ou non-supervisée. Il est possible d'obtenir ces représentations de diverses façons comme via des méthodes neuronales Trigeorgis et collab. (2016) ou bien via une factorisation en matrices non négatives Bisot et collab. (2016).

6.2.5.1 Vue d'ensemble

N'ayant pas beaucoup de données labellisées, nous nous sommes concentrés sur un apprentissage non supervisé basé sur des auto-encodeurs qui s'est avéré utile pour de nombreuses tâches très différentes utilisant un signal acoustique. Dans cette sous-section, nous présenterons une méthode d'apprentissage non supervisé basée sur des RNN-LSTM auto-encodant des représentations temps-fréquences. Les avantages de

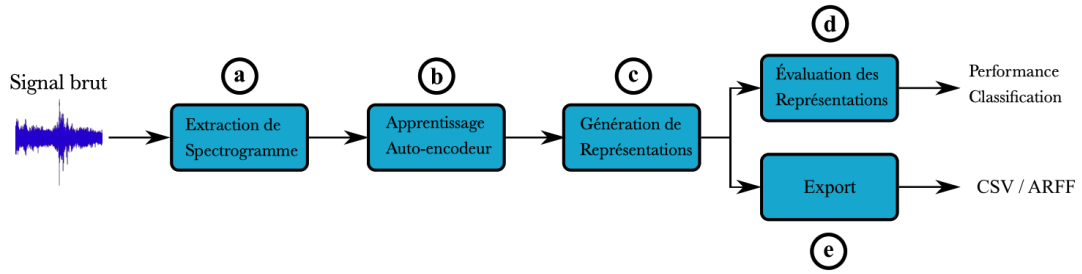


FIGURE 6.5: Processus de création des représentations apprises

cette méthode sont multiples. Premièrement, l'apprentissage se fait de manière non supervisée sur un grand corpus de données disponible, et le test sur un sous ensemble de données labellisées. Deuxièmement, on a une prise en compte la dynamique séquentielle de la parole grâce à des modèles séquentiels. Troisièmement, l'extraction des représentations est de bout-en-bout (*end-to-end*) sans utiliser des représentations complexes extraites au préalable.

Lors de la création des représentations, nous avons utilisé la boite à outils *AuDeep* de Freitag et collab. (2017) permettant de faire l'ensemble des opérations présentées ci-dessous. Le processus d'apprentissage et de génération est schématisé en Figure 6.5 et se compose des points suivants :

a. Extraction des spectrogrammes de chaque fichier audio
b. Entrainement de l'autoencodeur sur les spectrogrammes précédemment extraits
c. Génération des représentations de tous les fichiers audio à l'aide de la dernière couche cachée des RNN-LSTM
d. Possible évaluation des représentations générées à l'aide d'un ensemble de labels et d'une partition du corpus pour une validation croisée.
e. Exportation des représentations générées.

6.2.5.2 Création des représentations temps-fréquence

Le spectrogramme ou représentation temps-fréquence d'un signal est un diagramme représentant le spectre d'un phénomène, associant à chaque fréquence une intensité, et en fonction du temps. Pour le construire, on utilise une transformée de Fourier rapide (*Fast Fourier Transform* : FFT) avec une fenêtre que l'on fait glisser tout au long de la durée du son. L'intérêt est la possibilité de voir l'évolution spectrale au cours

du temps, cette évolution étant, par exemple, essentielle dans la définition du timbre d'une voix.

Différents paramètres sont disponibles pour la création des spectrogrammes :

- la largeur de la fenetre ;
- la taille du recouvrement entre deux fenêtre ;
- la taille maximal du fichier audio à prendre en compte ;
- la possibilité de fusionner les deux canaux audio ;
- le type de *padding* ;
- la coupe du signal audio (*audio chunking*).

Les paramètres que nous avons utilisés sont résumés dans le tableau 6.2. Toutes les combinaisons ont été utilisés.

6.2.5.3 Entraînement des Autoencodeurs

Les représentations ont été apprises de manière non supervisée à l'aide de RNN fonctionnant sur le principe d'un auto-encodeur de la représentation temps-fréquence. Cette approche permet la création de représentations neuronales à l'aide d'un modèle entrainé sur de grandes quantités de données non annotées, de manière non supervisée. Le modèle utilisé est une extension du réseau de neurones récurrent seq2seq présenté par Sutskever et collab. (2014). Le principe est d'utiliser un RNN comme encodeur de la représentation temps-fréquence, et un RNN décodeur pour la reconstruction de cette représentation.

La matrice temps-fréquence est coupée en fonction du temps afin d'être transformée en séquence de vecteurs qui servent d'entrée au RNN. La représentation interne finale du RNN est ensuite passé à travers une couche simple de neurones, puis est utilisé comme état initial d'un RNN et d'une couche linéaire servant à décoder la matrice (voir Figure 6.6). L'optimisation se fait en minimisant la fonction de coût classique qu'est l'erreur moyenne quadratique sur la reconstruction des représentations temps-fréquence. Afin d'accélérer l'entrainement, on utilise en entrée du décodeur la vraie séquence à re-coder, comme suggéré par Sutskever et collab. (2014).

6.2.5.4 Paramètres utilisés

Les différents paramètres utilisés pour l'apprentissage des représentations avec les auto-encodeurs sont résumés dans le tableau 6.2.

Les données que nous avons utilisés pour entrainer ces auto-encodeurs sont les mots et les tours de parole du corpus SEMAINE-Opinions (voir chapitre 5) selon la taille des échantillons que nous cherchions à générer.

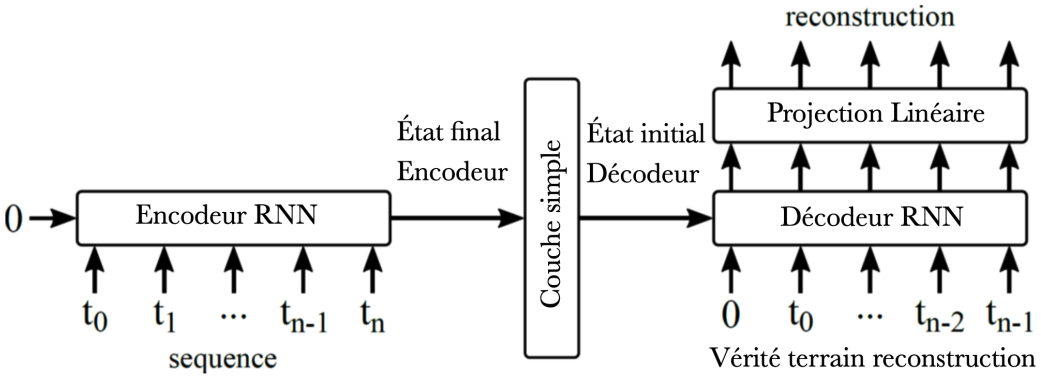


FIGURE 6.6: Modèle utilisé pour l'autoencodage des représentations

TABLE 6.2: Paramètres utilisés pour la création des représentations

Étape	**Paramètre**	**Valeur**
Spectrogramme	Fenêtre (s)	0,06 ; 0,08 ; 0,10
	Recouvrement (s)	0,02 ; 0,04
	Taille max audio (s)	6 ; 8 ; 10 (en partant de la fin)
	Fusion canaux	Somme
	Nombre coefficients Mel	64 ; 128
	Padding	Nul
Représentation	Type d'encodeur	Bi-LSTM
	Neurones encodeur	256
	Couches encodeur	2
	Type de décodeur	Bi-LSTM
	Neurones décodeur	256
	Couches décodeur	2
	Dropout	0,2
	Époques	5
	Taux d'apprentissage	0,001

6.2.6 Tableau récapitulatif

Nous avons choisi d'utiliser des descripteurs de la qualité de voix, ainsi que des descripteurs généraux comme les MFCC, la sonie, l'energie et pour l'articulation et la prononciation comme les formants. Dans un but de simplification, les différents descripteurs utilisés pour représenter le signal audio sont résumés dans le tableau 6.3.

TABLE 6.3: Descripteurs du signal acoustique

Type	**Descripteur**	**Taille**	**Pourquoi**
Cepstral	MFCC	12	généraux
Spectral	Formants	5	prononciation
Qualité de voix	PSP	1	indépendant de F_0
	HNR	1	contribution parole
	QoQ	1	voix tendue/soufflée
	NAQ	1	robuste au bruit
	PeakSlope	1	tension voix
	MDQ	1	tension voix
	R_d	1	relaxation voix
	H1H2	1	voix tendu/haletante
Prosodique	Pitch	1	hauteur voix
	Énergie	1	puissance émise
	Sonie	1	puissance ressentie

Nous avons fait le choix de créer notre ensemble de descripteurs acoustiques (EDA) avec des descripteurs experts (sous-section 2.2.1) de la qualité de voix (voir sous-sous-section 2.2.1.3) qui sont liées aux changements physiologiques du locuteur et ainsi qui pourraient donner des indications précieuse pour notre tâche. Nous avons ajouté des descripteurs de la prosodie (voir sous-sous-section 2.2.1.1) et spectraux/cepstraux (voir sous-sous-section 2.2.1.2) que nous avons jugés utiles. Nous avons aussi essayé des représentations générées avec des auto-encodeurs séquentiels (voir sous-section 2.2.3).

6.3 SEGMENTATION DU SIGNAL DE PAROLE ET INTÉGRATION DES DESCRIPTEURS

Les descripteurs présentés sont extraits à des niveaux différents : mot ou structure pour les descripteurs textuels et fréquence d'extraction de 10Hz ou plusieurs secondes pour les descripteurs audio. Nous souhaitons prendre en compte la séquentialité de la parole dans notre modèle, ainsi on utilise une séquence de vecteurs de descripteurs : une segmentation du signal de parole est nécessaire. Nous étudions des segmentations

TABLE 6.4: Fonctionnelles utilisées pour l'intégration des descripteurs audio

Fonctionnelles	**Groupe**
Quartiles 1 et 3, Amplitude entre les quartiles 1 et 3	Centiles
Médiane	Centiles
Centiles 10 et 90, Amplitude entre les centiles 10 et 90	Centiles
Écart-type , Asymétrie	Moments
Moyenne arithmétique	Temporelle
Régression linéaire pente, biais	Régression

à différents niveaux, et une intégration des descripteurs sur les intervalles obtenus est obligatoire.

6.3.1 Intégration des descripteurs

Nous avons extrait ces descripteurs acoustiques toutes les 10 ms via les boites à outils COVAREP (Degottex et collab., 2014) et OpenSMILE (Eyben et collab., 2013) en utilisant des paramètres d'extractions classiques décrits dans Eyben et collab. (2015) et Degottex et collab. (2014). Ces descripteurs ont enfin été intégrés temporellement à l'aide de 13 fonctionnelles étant : les centiles 10, 25, 50, 75 et 90, la différence entre les centiles 10 et 90, 25 et 75, la différence entre le maximum et le minimum, la moyenne arithmétique, la déviation standard, l'asymétrie, la pente et le biais de la régression linéaire. Ceci est résumé dans le tableau 6.4.

Avec la prise en compte des dérivées de chaque descripteurs et l'ajout de la moyenne des fenêtres voisées, on obtient finalement $(11+5+12)*13*2+1=729$ valeurs pour l'audio.

6.3.2 Segmentation en mots

La majorité des modèles séquentiels de la compréhension de la communication humaine multimodaux utilisent une segmentation en mots (Chen et collab., 2017; Sahay et collab., 2018).

6.3.3 Utilisation des pauses pour segmenter

La structure de la parole spontanée est telle que **de nombreuses phrases sont arrêtées en cours de route rendant difficile la segmentation d'un monologue en unités adéquates**. Nous avons alors choisi d'utiliser les pauses afin de segmenter chaque document en unités inter-pausales (UIP) à cause du rôle important que jouent ces auto-interruptions dans la segmentation du discours (Campione et Véronis, 2002). Une fois

cette segmentation effectuée, nous calculons les descripteurs sur chacun des segments et utilisons chaque séquence pour entraîner un classifieur qui pourra ensuite prédire le label le plus probable pour chaque nouveau document (voir Figure 7.2).

Pour traiter une longue séquence de parole sans interaction ni arrêt et aucune annotation nous avons choisi d'utiliser une variante des Champs Aléatoires Conditionnels (*Consitional Random Fields* : CRF), un classifieur discriminatif qui a prouvé son utilité dans bons nombres de tâches de traitement du langage naturel et de vision par ordinateur. Cette variante, appelé Champs Aléatoires Conditionnels Cachés (*Hidden Conditional Random Fields* : HCRF) a été utilisé avec succès pour analyser des séquences textuelles, audio ou vidéo et leur donner un label global. Les modèles à états latents ont déjà prouvé leur efficacité dans l'analyse de sentiment multimodale ou classification d'accord (Morency et collab., 2011; Bousmalis et collab., 2011).

Les états latents des HCRF sont utiles lorsque la représentation associée à une observation contient assez de données sur celle-ci pour pouvoir être associé à l'état latent pertinent. La représentation en IPU permet justement d'utiliser une segmentation optimale à la longueur caractéristique d'une structure d'opinion dans le langage (plus de détails dans la sous-sous-section 7.2.1.1). Les états cachés des HCRF permettent de modéliser une dynamique sous-jacente à l'opinion que l'on peut retrouver dans une critique oral de film et l'on suppose que la nature discriminative du CRF favorisera l'émergence de structures entre les descripteurs et les règles linguistiques au terme de la phase d'apprentissage.

7 Modèles d'apprentissage et base de données

Résumé du chapitre

— Les HCRF ont des fonctions caractéristiques de transition, de classe et d'états latents permettant de modéliser une séquence d'observation par une séquence d'états cachés. Ceci permet de faire une clusterisation linéaire des observations.

— Pour l'analyse de l'opinion intra-locuteur, les HCRF nous permettent de modéliser la structure locale et globale de la dynamique, et ainsi faire ressortir l'opinion générale dans des critiques de film à l'oral.

— Pour l'analyse de l'opinion inter-locuteur, nous avons étudié différentes segmentations, descripteurs et configurations du modèle HCRF pour classifier l'opinion du deuxième locuteur dans une paire de tour de parole et modéliser la dynamique interactionnelle.

7.1 HCRF : Champs Aléatoires Conditionnels Cachés

7.1.1 Modèle général

Le modèle de HCRF (Quattoni et collab., 2007) permet d'apprendre une fonction liant une séquence d'observations locales $\mathbf{x}_i = \{x_1,...,x_{L_i}\}$ de longueur L_i à un label $y_i \in \mathcal{Y}$. Chaque observation x_k est représentée par un vecteur de descripteurs d'observation $\phi(x_k) \in \mathbb{R}^d$, où d est la dimension de la représentation de notre signal. Pour chaque $\mathbf{x}_i$, une séquence de variables latentes non observées $\mathbf{h}_i = \{h_1,...,h_{L_i}\}$ est définie où $h_k \in \mathcal{H}$,$\mathcal{H}$ est un ensemble d'états finis (Quattoni et collab., 2007).

On utilise la probabilité postérieure $P(y|\mathbf{x},\theta)$ de l'Éq (7.1) pour inférer une décision concernant le label y :

$$P(y|\mathbf{x};\theta) = \sum_{\mathbf{h}} P(y,\mathbf{h}|\mathbf{x},\theta) = \frac{\sum_{\mathbf{h}} e^{\Psi(y,\mathbf{h},\mathbf{x};\theta)}}{\sum_{y',\mathbf{h}} e^{\Psi(y',\mathbf{h},\mathbf{x};\theta)}} = \sum_{\mathbf{h}} \frac{1}{Z(\mathbf{x};\theta)} e^{\Psi(y,\mathbf{h},\mathbf{x};\theta)} \tag{7.1}$$

où $\Psi(y,\mathbf{h},\mathbf{x};\theta) \in \mathbb{R}$ est une fonction potentielle mesurant la compatibilité entre un label, une séquence d'états cachés et les observations. La définition de la fonction potentielle $\Psi(y,\mathbf{h},\mathbf{x};\theta)$ se fait comme indiqué dans l'équation 7.2 comme la somme de 3 différentes composantes, chacune représentant des aspects différents du modèle. Les composantes φ_o, φ_s, et φ_t sont les sommes de ce qu'on appelle des fonctions caractéristiques f_o,f_s,f_t, nous les développerons dans la suite de cette section. La fonction potentielle pour un HCRF linéaire est définie comme ceci :

$$\Psi(y,\mathbf{h},\mathbf{x};\theta) = \varphi_o(\mathbf{x},\mathbf{h},\theta) + \varphi_s(\mathbf{h},y,\theta) + \varphi_t(\mathbf{h},y,\theta) \tag{7.2}$$

où

$$\varphi_o(\mathbf{x},\mathbf{h},\theta) = \sum_{j=1}^{L} f_o(x_j,h_j) = \sum_{j=1}^{L} \langle \phi(x_j) \,|\, \theta_o(h_j) \rangle \tag{7.3}$$

$$\varphi_s(\mathbf{h},y,\theta) = \sum_{j=1}^{L} f_s(y,h_j) = \sum_{j=1}^{L} \theta_s(y,h_j) \tag{7.4}$$

$$\varphi_t(\mathbf{h},y,\theta) = \sum_{j=2}^{L} f_t(y,h_j,h_{j-1}) = \sum_{j=2}^{L} \theta_t(y,h_j,h_{j-1}) \tag{7.5}$$

7.1.2 Les fonctions caractéristiques

Les trois sommes de fonctions caractéristiques φ_o, φ_s, et φ_t définis dans les équations (7.3), (7.4) et (7.5) ont des rôles particuliers dans la définition du HCRF, que nous allons présenter ici.

Les fonctions caractéristiques d'états latents f_o

Les fonctions caractéristiques d'états latents f_o dépendent uniquement du vecteur d'observation courant et de l'état caché courant. Un vecteur de poids $\theta_o(h_j)$ est créé pour chaque état caché $h_j \in \mathscr{H}$ et est de taille $\mathbb{R}^d$, le vecteur final $\boldsymbol{\theta_o}$ est donc de longueur $d \times |\mathscr{H}|$. Le produit scalaire représente **la compatibilité entre les descripteurs des observations** $\phi(x_j)$ **et l'état caché** h_j.

Les fonctions caractéristiques de classes f_s

Les fonctions caractéristiques de classes f_s dépendent de la classe globale y et de l'état caché courant h_j.

Le poids $\theta_s(y,h_j)$ représente la compatibilité entre une classe $y \in \mathscr{Y}$ et un état caché $h_j \in \mathscr{H}$, le vecteur de poids $\boldsymbol{\theta_s}$ est donc de longueur $|\mathscr{Y}| \times |\mathscr{H}|$.

Les fonctions caractéristiques de transitions f_t

Les fonctions caractéristiques de transitions f_t dépendent de la position dans la séquence d'états cachés (état caché courant et celui précédent) et de la classe. Le poids $\theta_t(y,h_j,h_{j-1})$ représente la compatibilité de la transition depuis un état latent $h_{j-1} \in \mathscr{H}$ vers un autre état latent $h_j \in \mathscr{H}$ (qui peut être le même) et ceci pour une classe $y \in \mathscr{Y}$. Le vecteur de poids $\boldsymbol{\theta_t}$ est donc de longueur $|\mathscr{Y}| \times |\mathscr{H}| \times |\mathscr{H}|$.

7.1.3 Entraînement

Le HCRF étant un modèle d'apprentissage supervisé, un ensemble d'entraînement labellisé est nécessaire afin d'apprendre les modèles du graphe. Notre ensemble d'entrainement sera noté $(\mathbf{x}_i, y_i), i = 1..n$ où n est la taille de l'ensemble d'entraînement.

Fonction de coût et régularisation

Le modèle est entraîné de manière classique en maximisant un coût de log vraisemblance L. Il peut être régularisé avec des pénalités de norme ℓ_2 et de norme ℓ_1 sur les paramètres (Éq (7.6)) (Quattoni et collab., 2007).

$$L(\theta) = \sum_i \log P(y_i|x_i,\theta) - \lambda_1||\theta|| - \lambda_2||\theta||^2 \tag{7.6}$$

L'utilisation de la norme ℓ_2 permet de régulariser les poids du modèle afin d'éviter le sur-apprentissage. D'une manière simple, en appliquant un coût quadratique sur l'amplitude des poids lors de l'apprentissage, on force le modèle à utiliser la majorité

des descripteurs, ce qui permet d'utiliser le maximum d'information disponible. De plus, on évite que le modèle atteigne une complexité inutile qui ne représente pas la réalité.

L'utilisation de la norme ℓ_1 permet de régulariser les poids en mettant à zéro une partie des poids les moins impactants, éliminant du même coup les descripteurs les plus bruitants et moins discriminants pour la tâche. D'une manière intuitive, une pénalisation de degré 1 sur l'amplitude des poids force le modèle à la simplicité car la boule ℓ_1 en dimension élevée a la majorité de sa densité sur les axes de $\mathbb{R}^n$. Une solution donnant le même coût final sera plus probablement placée sur les axes de la base correspondant aux poids les moins impactant, ce qui équivaut à une valeur nulle pour ces poids (si un point d'une droite est sur l'axe des y, alors ce points a une valeur de y nulle).

Algorithme d'apprentissage

Une descente de gradient est utilisée afin de chercher l'ensemble de paramètres optimum θ^* défini comme :

$$\theta^* = \arg\max_{\theta} L(\theta) \tag{7.7}$$

Il existe différents algorithmes que nous pouvons utiliser pour l'optimisation de la fonction de coût sur l'ensemble des données d'apprentissage. Par exemple, il est possible d'utiliser une méthode de gradient conjugué ou des méthodes quasi-Newton comme le BFGS (*Broyden–Fletcher–Goldfarb–Shanno* : BFGS). L'algorithme que nous avons utilisé est le L-BFGS (Limited-memory BFGS) qui est une approximation du BFGS. L'utilisation de cette méthode approximée permet une convergence beaucoup plus rapide du modèle.

7.1.4 Inférence

Étant donné une nouvelle séquence d'observations $\mathbf{x}$, l'inférence afin de trouver le label associé se calcule de plusieurs manières. Dans certaines de nos études, nous utilisons simplement la quantité associée avec la fonction L qui a servi lors de l'optimisation : l'équation (7.8) qui utilise le label maximisant la somme de tous les chemins (séquences d'états cachés) possibles.

$$y^* = \arg\max_{y} P(y|\mathbf{x},\theta^*) = \sum_{\mathbf{h}} P(y,\mathbf{h}|\mathbf{x},\theta^*) \tag{7.8}$$

Dans certaines de nos études, la décision se fait en utilisant le label y^* de la classe ayant la séquence d'états cachés la plus probable $\mathbf{h}^*$. Le couple $y^*,\mathbf{h}^*$ solution de l'équa-

tion 7.9 est trouvé à l'aide d'un algorithme de Viterbi.

$$y^*, \mathbf{h}^* = \arg\max_{y,\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}, \theta^*) = \arg\max_{y,\mathbf{h}} e^{\Psi(y,\mathbf{h},\mathbf{x};\theta)} \tag{7.9}$$

L'utilisation de la chaîne d'états cachés la plus probable permet de prendre une décision par rapport à une séquence précise que l'on peut visualiser et qui peut, possiblement, permettre une interprétation des observations.

7.2 MODÈLE D'APPRENTISSAGE POUR L'ANALYSE D'OPINION

Dans cette section nous allons aborder la méthodologie que nous avons suivie afin de créer les modèles d'apprentissage qui ont été étudiés durant cette thèse. Les modèles HCRF que nous avons créés permettent d'utiliser les différentes caractéristiques d'un signal de parole, qu'il soit émis dans un contexte interactionnel ou bien de manière discursive.

7.2.1 Modèle d'apprentissage intra-locuteur

Dans notre premier travail, la tâche à laquelle nous nous sommes attaqué est une tâche d'analyse d'opinion dans un long discours afin de faire une première version du modèle, permettant de modéliser la dynamique de l'opinion dans un long discours globalement annoté. Les HCRF tirent parti des différentes spécificités de la tâche, tant au niveau du type de données qu'au niveau du phénomène que l'on cherche a caractériser. Nous expliciterons les avantages des HCRF ci-dessous.

7.2.1.1 Un modèle interprétable

Un des aspects important du modèle d'apprentissage présenté est son interprétabilité (caractère interprétable). Chaque poids $\boldsymbol{\theta}$ du HCRF a une signification précise quant à sa place dans le modèle graphique. En analysant l'amplitude et le signe de l'ensemble des poids, on peut conclure à des compatibilités, incompatibilités ou indifférence entre les différents états, les labels globaux, et les descripteurs du signal de parole :

1 En analysant les poids de label global $\boldsymbol{\theta_s}$, il est possible de visualiser quel état caché est lié avec quel label. Cela permet de voir les affinités des états cachés avec chacun des labels.
2 En analysant les poids des observations $\boldsymbol{\theta_o}$, il est possible d'analyser chacun des états cachés h_i en regardant les descripteurs qui les activent ou désactivent le plus. Il est intéressant de savoir quelles sont les activations des différents

états afin de pouvoir caractériser la pertinence de ces états et de visualiser des exemples d'activations positives et négatives dans les données.

3 En analysant les poids de label global $\boldsymbol{\theta_t}$, on peut lier les différents états cachés entre eux selon le label global. Ceci permet de créer encore une distance entre chacun des états en utilisant leurs compatibilités réciproques selon chaque label.

Un des aspects les plus intéressants des HCRF, est le fait que les états cachés sont appris à l'aide d'un label global, mais ils agissent cependant au niveau des observations. Ils permettent d'effectuer une *clusterisation* des données de manière non supervisée à l'échelle des observations. Cette *clusterisation* est d'autant plus performante que les représentations utilisées en entrée sont adaptées à la tâche. En effet, si les compatibilités entre la séquence d'états cachés, les observations et le label global est non-linéaire, les compatibilités entre un état et une observation sont calculées via un simple produit scalaire. On a donc une séparation de manière locale linéaire, qui équivaut à une séparation de l'espace des observations à l'aide un hyperplan [1]. Les HCRF sont donc très sensibles à la qualité des représentations, et des moyens tels qu'un prétraitement par une méthode à noyaux comme dans (Song et collab., 2012b) permettent d'intégrer une non-linéarité.

Pour finir, il est à noter que cette interprétabilité est au niveau des poids du modèle. Ainsi, si des représentations ayant un sens abstrait sont utilisés, il devient plus difficile d'interpréter le modèle. Par exemple, si il n'y a aucune interprétabilité à l'analyse d'une simple dimension de représentation distribuée alors il sera difficile d'interpréter les poids associés.

7.2.1.2 Modélisation de la dynamique

Dans un énoncé, que ce soit un monologue ou un dialogue, il est important de prendre en compte la dynamique à la fois de manière locale et de manière globale. En effet, dans une même phrase, une suite de phrases ou un discours, si l'énoncé est correctement structuré, les transitions font sens et il est important de modéliser cette séquentialité. C'est ce qu'on appellera ici la **dynamique locale**. Les HCRF arrivent à modéliser cette séquentialité avec les poids de transitions $\boldsymbol{\theta_t}$. Ces poids servent à modéliser le caractère lisse et continu d'un discours, et ce au niveau des observations. Par exemple, il est peu probable de passer d'un état caché caractéristique de l'état émotionnel d'un locuteur joyeux à celui caractéristique d'un locuteur triste, comme il est

1. En effet, d'un point de vue reconnaissance de forme dans l'espace, un produit scalaire avec $\theta_o(h_j)$ équivaut à calculer la distance entre le vecteur d'observation et l'hyperplan ayant pour vecteur normal $\theta_o(h_j)$.

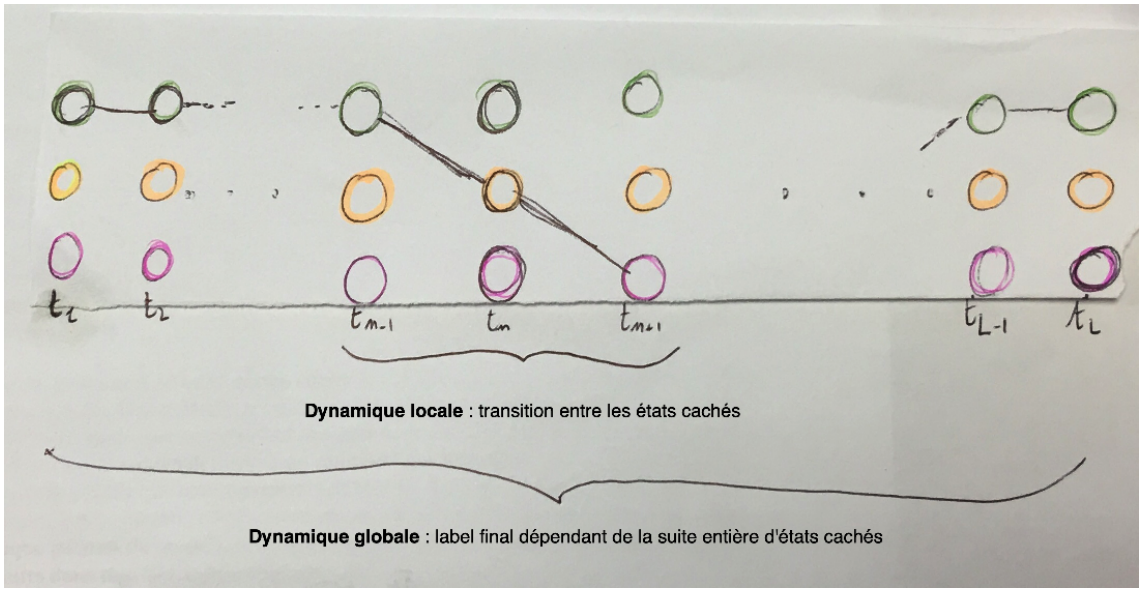


FIGURE 7.1: Dynamiques locales et globales dans les HCRF (le trait entre les états représente la navigation d'un état à l'autre)

peu probable de passer du rire aux larmes.

De plus, un discours est entier, et prend tout son sens avec l'accumulation des différentes parties articulées entre elles. Ainsi, il est aussi important de modéliser le discours en entier et pas uniquement la dynamique locale. Cette dynamique est de plus haut niveau, elle agit à l'échelle de la phrase, du paragraphe, et du discours entier. C'est ce qu'on appellera ici la **dynamique globale**. Les HCRF arrivent à modéliser cette unité par l'utilisation des poids de label global $\boldsymbol{\theta}_s$, allouant aux HCRF la capacité de pouvoir modéliser le discours par rapport au label global. En effet, les HCRF sont appropriés quand il y a une forte corrélation entre le label général et les labels plus fins des observations qui sont modélisés par les variables latentes (Täckström et McDonald, 2011). Si l'on s'attend en effet à voir des observations positives, négatives et neutres dans tous nos documents, il est aussi bien plus probable qu'il y ait plus d'observations positives que négatives dans un document positif.

Pour finir, le calcul de la probabilité du label final $P(y,\mathbf{h}|\mathbf{x},\theta^*)$ utilise la séquence entière des états cachés $\mathbf{h}$, liant ainsi dynamique globale et dynamique locale (voir figure 7.1).

7.2.1.3 Adapté à des corpus de tailles restreintes

Lors des différentes études de cette thèse, nous nous sommes intéressés à l'analyse d'opinions dans les interactions de parole. Lorsque ce travail de thèse a commencé, il n'y avait pas de bases de données collectée pour les opinions qui soit de taille significa-

tive. La première disponible a été CMU-MOSI de Zadeh et collab. (2018b), les base de données précédentes étant de tailles réduites. On citera par exemple ICT-MMMO de Wöllmer et collab. (2013b), que nous avons utilisée dans la première étude (chapitre 8). Pour cette raison, les modèles d'apprentissage nécessitant de grandes quantités de données ne nous semblaient pas adaptés.

Ainsi, il nous a paru important de souligner l'aspect peu gourmand en données des HCRF, par rapport à des modèles comme les réseaux de neurones. On verra au cours de nos expériences, à la suite de comparaisons avec des méthodes utilisant des RNN-LSTM que les HCRF arrivent à mieux modéliser le discours et permettent d'obtenir des performances plus élevées.

7.2.1.4 Segmentation automatique et discours

Le dernier aspect que nous allons aborder ne fait pas partie du modèle d'apprentissage en soi, mais nous avons choisi de le placer ici car il a un impact non négligeable sur la façon de modéliser le signal de parole. En effet, l'utilisation d'une segmentation automatique à base des pauses du locuteur permet :

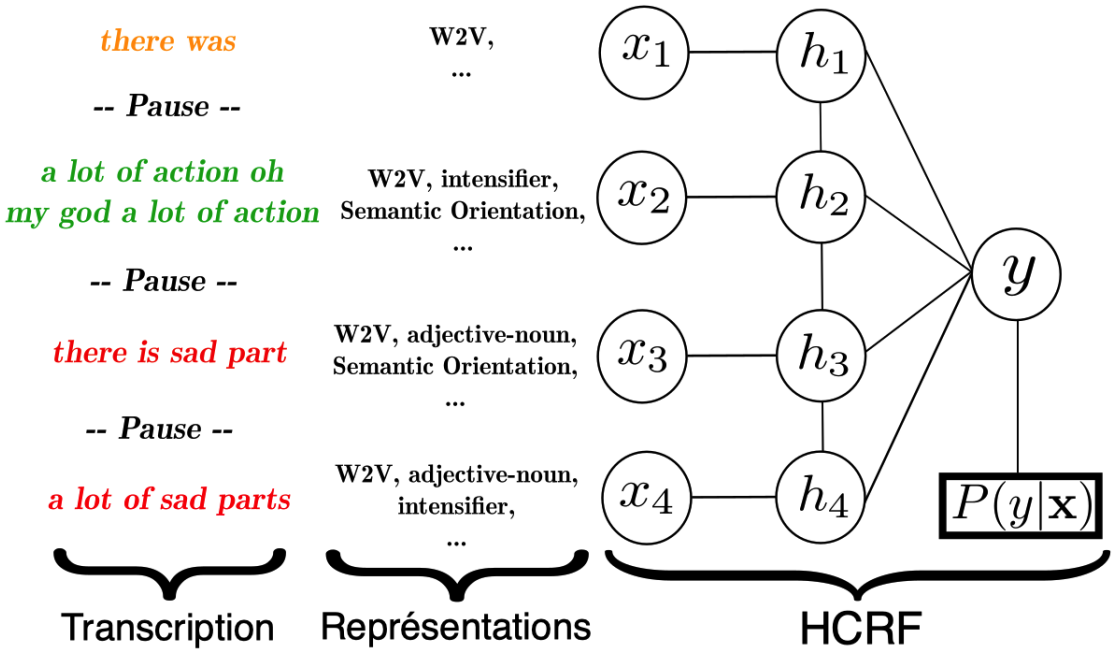
— de découper le signal textuel en unités pertinentes d'un point de vue linguistique ;

— de découper le signal audio en unités pertinentes d'un point de vue acoustique.

Le fait ne pas couper ni les structures linguistiques ni les structures acoustiques, tout en conservant un aspect séquentiel, choisi par le locuteur, a un impact crucial sur la modélisation de l'opinion.

La segmentation est une étape majeure pour le HCRF. Le HCRF est sensible à la qualité des observations : l'état caché h_j n'est pas lié aux observations $\mathbf{x}_{j-1}$ qui sont uniquement associées à l'état caché h_{j-1}. Contrairement aux RNN qui permettent de mettre à jour un vecteur latent au fur et à mesure que l'on avance dans la phrase, le HCRF linéaire ne permet pas une combinaison des observations au fur et à mesure, car chaque observation a son propre état caché associé. Par exemple, il devient beaucoup plus simple pour le modèle d'associer le n-gramme '*really liked*' à un état lui-même associé au label positif que d'utiliser un état séparé pour '*really*', puis un état pour '*liked*'. De même, lorsqu'on intègre les mots-vecteurs ensemble, il est utile de garder du contexte, qui est fortement impactant sur le sens donné par le locuteur (Polanyi et Zaenen, 2006). Ainsi, si on veut qu'un état soit lié, représentatif de quelque chose de précis, il faut qu'il puisse "observer" quelque chose de consistant, donc que les observations soient consistantes : la segmentation est une étape majeure, et utiliser des Unité Inter-Pausales (UIP) est particulièrement intéressant pour cela (exemple Figure 7.2).

FIGURE 7.2: Schéma de notre système



7

7.2.2 Modèle d'apprentissage interlocuteur

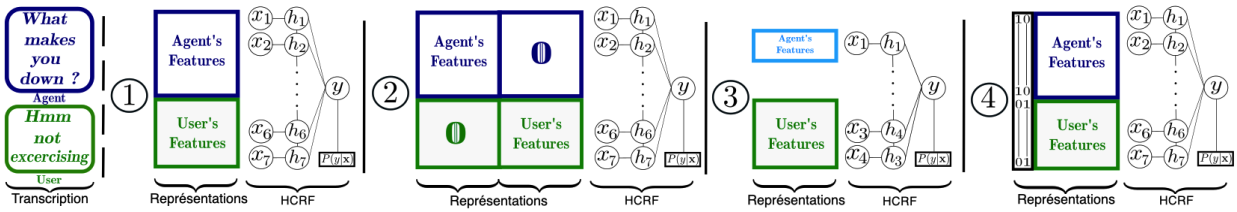
Pour l'étude des interactions, le contexte d'étude est totalement différent, ainsi nous proposons une autre série de modèles capables d'utiliser ce contexte à leur avantage. Les séquences à annoter sont beaucoup plus courtes et ce n'est plus un monologue mais une partie de discussion entre deux locuteurs. La tâche à effectuer étant une classification des opinions de l'utilisateur dans une interaction humain-agent, nous avons jugé important d'utiliser l'aspect asymétrique de la situation et de l'incorporer dans nos modèles de plusieurs manières.

Dans notre précédent modèle, nous prenions uniquement en compte le contexte du locuteur à analyser, ici nous avons créé un modèle qui prend en compte le contexte conversationnel afin de prendre une décision. Les enjeux sont de :

— prendre en compte le contexte dialogique, à l'aide du tour de parole précédent ;
— modéliser les deux tours de parole de manière dynamique ;
— utiliser l'asymétrie de l'interaction avec des représentations et segmentations spécialement dédiées ;

Dans cette section, nous présentons les différents modèles de HCRF spécialement conçus pour analyser l'opinion d'un locuteur au sein d'une interaction dyadique.

FIGURE 7.3: Les différentes configurations HCRF-X étudiées pour modéliser l'interaction humain-agent (plus de détails ci-dessous)



7.2.2.1 Présentation des différents systèmes d'interaction

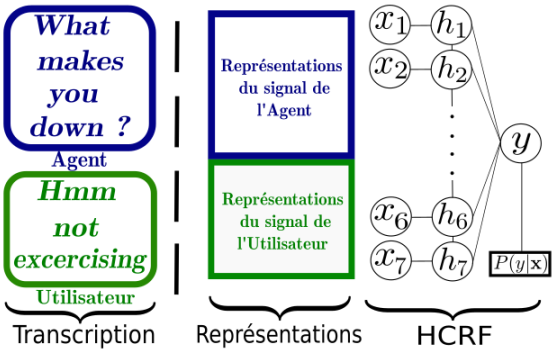
Dans nos différents modèles, nous avons essayé plusieurs façons de segmenter le texte et d'intégrer des descripteurs afin de prendre en compte le contexte interactionnel. L'objectif est d'analyser l'énoncé de l'utilisateur en s'appuyant sur des informations contenues chez l'agent. Détecter les attitudes de l'utilisateur lors d'une conversation humain-agent ne nécessite pas de prêter autant d'attention à l'énoncé de l'agent qu'à celui de l'utilisateur. Afin de mettre en évidence l'intérêt des informations contenues dans le tour de parole de l'agent, nous considérons comme système de base un modèle prenant en compte uniquement l'énoncé de l'utilisateur (modèle **HCRF-0**). Ensuite, nous avons étudié quatre différents modèles prenant en compte le contexte interactionnel de la PA [2] (voir Figure 7.3).

Comme première approche, nous avons utilisé une manière simple d'intégrer les descripteurs dans le HCRF : une segmentation mot par mot et un partage des mêmes poids θ_o entre l'agent et l'utilisateur (voir Figure 7.3, modèle **HCRF-1**). Motivés par l'idée que les rôles de l'agent et de l'utilisateur ne sont pas symétriques dans cette tâche, nous avons créé un modèle **HCRF-2** qui est entraîné avec différents poids θ_o pour les descripteurs respectifs de l'agent et de l'utilisateur. Dans le troisième modèle **HCRF-3**, nous avons décidé d'intégrer tous les descripteurs de l'agent sur son tour de parole afin d'obtenir un seul vecteur (contrairement à une séquence de vecteurs représentant la séquence de ses mots). Ceci permet de minimiser l'impact de l'énoncé de l'agent sur la décision finale tout en laissant cet énoncé influencer le début de la séquence d'états latents. Finalement, nous avons utilisé deux nouveaux descripteurs indiquant les locuteurs pour le quatrième modèle **HCRF-4**. Ceci joint les avantages des modèles 1 et 2 : les poids sont partagés mais on différencie les locuteurs.

De manière générale pour la majorité des modèles, nous utilisons une segmentation au niveau du mot, car la segmentation en pause est impossible : du fait de la nature particulière des dialogues du corpus, une segmentation en pause fournirait des séquences de tailles beaucoup trop courtes pour être utilisables par le HCRF.

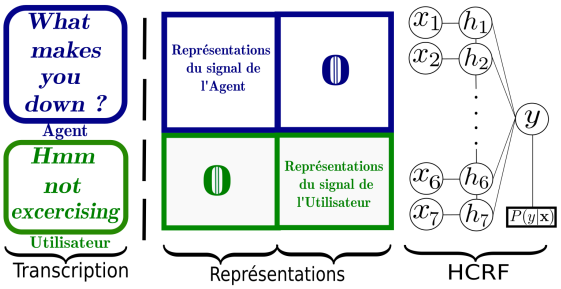
2. Paire Adjacente

7.2.2.2 HCRF-1



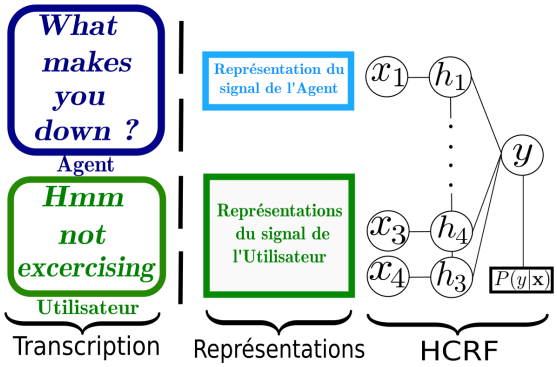
Le HCRF-1 est un modèle servant de système de base car il ne considère pas l'aspect interactionnel de la conversation. Les locuteurs ne sont pas différenciés, et aucun artifice n'est utilisé pour séparer le tour de parole de l'agent de celui de l'utilisateur. Les poids sont partagés entre les descripteurs de l'agent et l'utilisateur, qui sont considérés comme un seul locuteur. Cela permet d'apprendre plus facilement des structures spécifiques à l'opinion, et de pouvoir utiliser la connaissance relevant de cet apprentissage sur les deux locuteurs.

7.2.2.3 HCRF-2



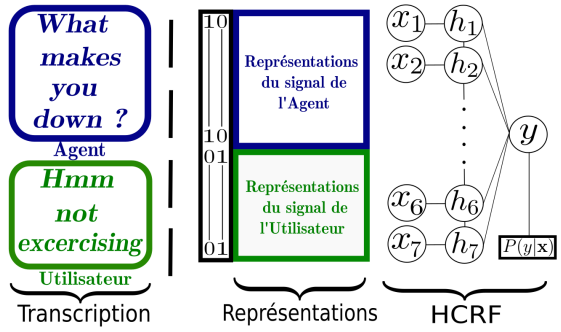
Le modèle HCRF-2 est un modèle spécialement conçu pour considérer l'interaction en utilisant des poids différents pour les descripteurs de l'agent et ceux de l'utilisateur. Ceci permet de bien différencier les deux locuteurs, en rendant leurs descripteurs indépendants. Néanmoins, on perd une information importante : les signaux de parole de l'agent et de l'utilisateur sont représentés avec des descripteurs similaires. Les mêmes descripteurs seront vus comme totalement différents. Si l'agent utilise le mot '*good*' et que l'utilisateur utilise aussi le mot '*good*', le modèle les considérera comme des mots différents.

7.2.2.4 HCRF-3



Le modèle HCRF-3 permet de représenter l'agent et l'utilisateur de manière asymétrique, ce qui est conforme aux données et à la tâche. Ce caractère asymétrique est mis en place avec l'utilisation d'une diminution des impacts de l'agent sur la décision finale, en intégrant son tour de parole en une seule observation. En effet, si l'agent est représenté par une unique observation, alors son impact dans la décision finale sera grandement diminué.

7.2.2.5 HCRF-4



Le modèle HCRF-4 permet de partager les poids entre l'agent et l'utilisateur, tout en différenciant les deux entités par l'ajout d'une indicatrice du locuteur actif. Ainsi, le système peut repérer qui est en train de parler, tout en utilisant le fait que les deux locuteurs utilisent le même langage pour parler

CONCLUSION

Dans cette partie, nous avons montré le classifieur que nous avons choisi d'utiliser dans la suite car :

- il permet de modéliser la séquentialité de la parole grâce aux poids de transitions au niveau local et au calcul d'inférence au niveau global ;
- il est interprétable en étudiant les différents poids, en particulier ceux des observations afin d'analyser les états cachés ;
- son apprentissage ne nécessite pas une grande quantité de données.

Finalement, nous avons adapté ce classifieur pour l'analyse d'opinions en segmentant de différentes façons l'énoncé du ou des locuteurs, le rendant sensible aux contextes particuliers de l'analyse des opinions intra-locuteur et inter-locuteurs.

Troisième partie

Expérimentations : Études sur l'analyse d'opinion

8

Expériences : Analyse de l'opinion intra-locuteurs

Résumé du chapitre

— Ces expériences ont été menées à l'aide des données textuelles uniquement. Les représentations construites par nos soins sont comparées avec Sac-de-N-grammes et les HCRF sont comparés avec une régression logistique et un LSTM.

— La configuration des états cachés du HCRF est toujours constituée de 3 types d'états : un état compatible avec le label positif, un état compatible avec le label négatif et des états non compatibles avec aucun des labels (dits neutres).

— En analysant les poids de ces états, on retrouve bien des descripteurs linguistiques et des représentations distribuées liés avec du contenu positif, négatif ou neutre des observations.

— Les états neutres permettent d'éloigner les observations neutres des états positifs et négatifs afin qu'elles n'influencent pas la décision finale.

— Les mots-vecteurs (appris sur de l'écrit général) n'ont pas le même sens dans un discours oral (*yeah*, *uhm*, ...) et dans un contexte de critique de film que dans du texte général. Ces incohérences sont prises en compte par les états neutres afin de diminuer leurs effets sur la décision finale.

8.1 DANS UN DISCOURS ENTIER : ÉTUDE SUR LE CORPUS *ICT-MMMO*

Dans cette première partie de thèse, nous nous sommes attachés à créer un système permettant de modéliser la dynamique de l'opinion à l'intérieur d'un long monologue non segmenté et non annoté. Nous allons ici parler du système proposé, ses résultats dans les expériences menées sur le corpus ICT-MMMO (décrit en sous-section 4.2.1 et section 5.1) et une analyse des performances abordant les descripteurs utilisés pour la linguistique et le modèle de Champs Aléatoire Conditionnels Cachés.

8.1.1 Protocole général de validation

Nous avons testé trois modèles avec des ensembles de descripteurs différents et des segmentations automatiques différentes dans le but de valider nos modèles.

Premièrement, nous avons créé un système de base pour la tâche abordée suivant un protocole similaire à celui de Wöllmer et collab. (2013a) sur la même base de données. Ce premier système est un modèle de régression logistique et avec une représentation en BoNG au niveau du document entier.

Deuxièmement, comme la régression logistique ne prend pas en compte la dynamique séquentielle de la parole et du discours, nous avons essayé un système de base plus élaboré permettant de gérer les données séquentielles : un réseau de neurones récurrent de type RNN-LSTM (Hochreiter et Schmidhuber, 1997). Comparé à ces modèles, les HCRF présentent l'avantage d'être interprétables dans la méthode par laquelle le modèle traite des données : de manière séquentielle tout en ayant la possibilité de modéliser la dynamique des phénomènes liés aux opinions (états émotionnels, position sociale adoptée, etc...) à travers l'utilisation des états cachés.

Troisièmement, nous avons comparé les représentations en BoNG à notre ensemble de descripteurs avec différents modèles d'apprentissage. Ceci a permis de bien valider l'apport de notre représentation du signal textuel .

Quatrièmement, nous avons examiné les performances que l'on pouvait obtenir avec différentes segmentations basées sur les pauses du locuteur. En créant des séquences d'observations de tailles variables, selon la longueur des pauses du locuteur, nous avons pu comparer les segmentations via les performances des systèmes.

Pour finir, afin de valider nos modèles, nous avons utilisé une validation-croisée à 10 tours, où les ensembles d'entrainement et de test sont disjoints. Chaque partition de test contenait les mêmes proportions de classes que celles trouvées dans la BDD entière. Les performances des systèmes sont calculées via différentes métriques : les scores F1 de chaque classe, l'exactitude (*Accuracy*), et le F1 global (voir annexe F).

8.1.2 Systèmes de base avec RL et LSTM

Méthodologie

Nous avons considéré des systèmes de base utilisant aussi bien des représentations simples que la représentation du signal textuel que nous avons créée et différentes segmentations (au niveau du document et au niveau de l'UIP). Ceci a permis de mesurer l'amélioration obtenue par l'utilisation de modèles HCRF. Nous avons utilisé une régression logistique avec une représentation de type BoNG comme Wöllmer et collab. (2013b) avec la même paramétrisation : utilisation de tri-grammes, racinisation de Porter (1980), tranformation TF-IDF, et normalisation au niveau du document. Cependant nous avons décidé de garder une taille de vocabulaire plus large, ce qui donnait de meilleurs résultats. Nous avons ensuite changé cette représentation pour un ensemble de descripteurs et de représentations plus sophistiqués (notre EDT dans le Tableau 6.1), qui contient une représentation distribuée des mots apprise de manière statistique (Mikolov et collab., 2013a) décrit dans la sous-section 6.1.3. Après une tokénisation [1] nous avons utilisé un correcteur orthographique [2] afin d'éliminer les erreurs de typographie des transcriptions et nettoyer le texte pour trouver les vecteurs de chaque mot (mots vides exclus). Nous avons suivi le protocole de Yang et collab. (2016) pour une tâche d'analyse de sentiment sur des critiques textuelles et nous avons fait la moyenne arithmétique des vecteurs de chaque mot contenu à l'intérieur de l'UIP. Cela permettait d'obtenir un vecteur de la même taille pour chaque UIP, que nous avons ensuite normalisé. Pour aider à la détection des opinions nous avons ajouté des descripteurs linguistiques et syntaxiques (décrits dans la sous-section 6.1.2). Nous avons utilisé les valeurs des descripteurs linguistiques présents pour chaque mot de l'UIP pour obtenir un score par descripteur et pour chaque UIP. Au niveau des segmentations, nous avons utilisé les pauses d'au moins 150, 300 ou 500 ms afin de segmenter le monologue du locuteur en énonciations.

Par rapport au réglage des hyperparamètres de la régression logistique, nous avons effectué des entrainements avec des valeurs de C, qui correspond à l'inverse du coefficient de régularisation, dans {0.1, 0.5, 1, 10, 100}. Nous avons utilisé l'implémentation de *scikit-learn* de Pedregosa et collab. (2012) pour la régression logistique.

Pour le RNN-LSTM, nous avons utilisé l'implémentation de Keras (Chollet, 2015) avec un nombre d'états cachés variant dans {64, 128, 256}, une régularisation avec une valeur de *dropout* (Srivastava et collab., 2014) de U et W (voir Hochreiter et Schmidhuber (1997)) dans {0.1, 0.2, 0.3} (si plus élevé, les performances chutent) et un nombre d'époques dans {4...10}. Nous avons utilisé l'entropie croisée comme fonction de coût

1. Nous avons utilisé le CoreNLP de Standford (Schuster et Manning, 2016)
2. https ://github.com/phatpiglet/autocorrect

et Adam comme algorithme d'optimisation (Kingma et Ba, 2014).

Résultats

Les résultats des systèmes de base sont répertoriés dans la première partie du tableau 8.2, à l'aide des scores $F1$ et de l'exactitude. Dans ce tableau, nous retrouvons différentes métriques : les scores F1 de chaque classe, et le taux de reconnaissance ou *Accuracy* (voir en annexe F pour plus de détails). Il est possible de voir que les meilleurs résultats sont obtenus avec notre ensemble de descripteurs. C'est un résultat inattendu étant donné que nous faisons la moyenne arithmétique des mots-vecteurs du document entier long en un seul vecteur, mais nous avons remarqué que la performance vient des autres descripteurs liés aux sentiments ou linguistiques. Les résultats du RNN-LSTM ne sont pas meilleurs pour la classe négative. Bien qu'il ait le potentiel de capturer une certaine dynamique, le réseau de neurones a besoin de plus de données que ce qui est disponible dans le corpus pour être pleinement efficace. Nos résultats sont en accord avec ce qu'on peut trouver dans la littérature. En utilisant la représentation classique en BoNG, Schuller et collab. (2009a) obtenaient un score F1 de 79% dans une tâche d'analyse de sentiument sur une base de données de critiques textuelles de films (*Metacritic*). Perez-Rosas et collab. (2013b) obtenaient un score F1 de 65% pour une tâche d'analyse de sentiment sur le corpus *MOUD* de critiques Vlogs en espagnol.

8.1.3 Modèles HCRF

Méthodologie

Nous avons entraîné les modèles HCRF avec le *wrapper* Matlab de *HCRF Library* Morency (2007) et utilisé un algorithme de type L-BFGS durant l'entraînement afin de maximiser la log-vraisemblance. Par rapport aux hyperparamètres des modèles, nous avons testé différentes valeurs qui sont résumées dans le tableau 8.1. La taille de la fenêtre de contexte signifie le nombre d'observations que l'on a concaténé à gauche et à droite autour de notre observation centrée. Lorsque l'observation n'avait pas de voisins, un padding avec des valeurs nulles a été utilisé. Nous avons fait des tests préliminaires avec plus d'états cachés et des plus hautes valeurs de coefficient de régularisation, mais les résultats n'étaient pas meilleurs et le temps d'entraînement significativement plus long.

TABLE 8.1: Ensemble des hyperparamètres testés lors de l'entraînement des modèles HCRF

Paramètre	Nom	Valeurs testées
λ_2	Coefficient de régularisation	{0.01, 0.05, 0.075, 0.1, 0.25, 0.5, 1}
$\lvert\mathscr{H}\rvert$	Nombre d'états cachés	{2, 3, 4, 5}
w	Taille de la fenêtre de contexte	{0, 1, 2}

TABLE 8.2: Scores de F1 et d'*Accuracy* (taux de bonnes réponses) avec les différents descripteurs, paliers de segmentation pour les pauses et modèles

Descripteurs	Modèle	F1+	F1-	F1	Acc
Positif	Naïf	78	0	50	63
BoNG	RL	84	69	78	79
Notre EDT	RL	83	72	79	79
Notre EDT (150ms)	LSTM	84	68	78	78
BoNG (150ms)	HCRF	84	67	78	79
Notre EDT (150ms)	HCRF	85	72	80	80
Notre EDT (300ms)	HCRF	**86**	**75**	**82**	**82**
Notre EDT (500ms)	HCRF	82	67	77	77

Résultats

Les résultats obtenus avec les HCRF et les systèmes de référence sont résumés dans le Tableau 8.2. La meilleure configuration est obtenue avec 5 états cachés, aucun contexte et un paramètre λ_2 de régularisation ℓ_2 égal à 10. Comme espéré, le HCRF améliore les résultats par rapport à la RL (F1 montant de 79 à 80) avec les descripteurs BoNG. Les meilleurs résultats sont obtenus avec une amélioration du score F1 pour la classe négative (8 points). Le palier de segmentation des pauses à 300ms apporte aussi une légère amélioration sur la classe négative alors qu'un palier de 500 ms diminue les performances. Cependant il est à noter qu'une validation croisée à 10 tours n'apporte pas assez d'informations pour conclure sur une significativité statistique d'une différence de performance ($p = 0.15$ pour la classe négative).

8.1.4 Discussion et analyse des résultats

Dans cette sous-section, on se concentre sur notre modèle le plus performant à 5 états cachés.

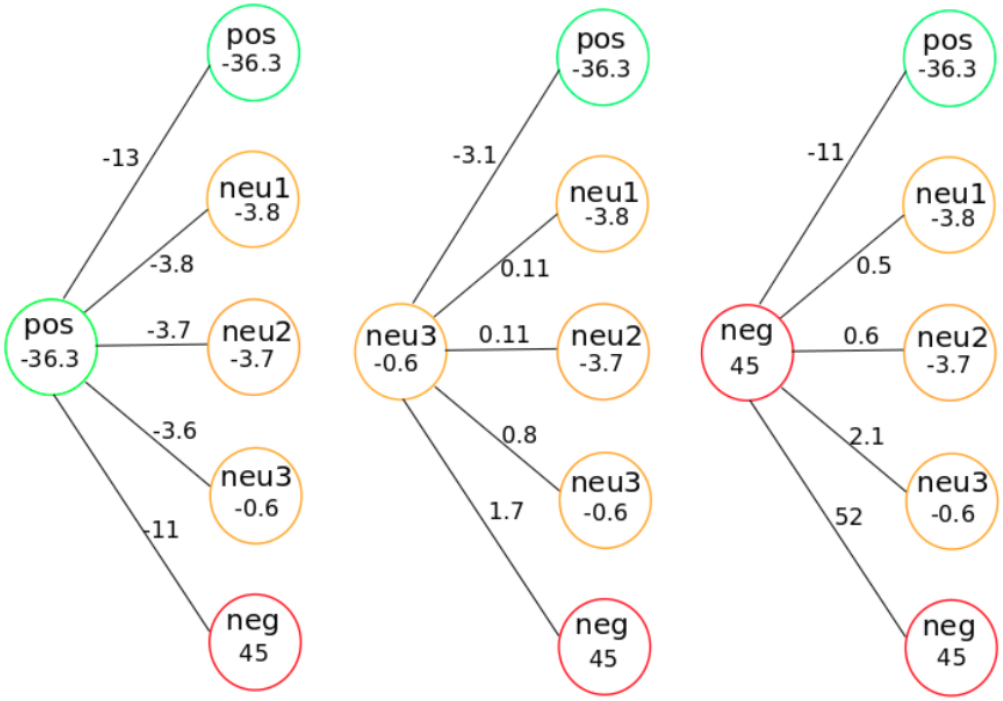


FIGURE 8.1: Poids de transition entre les états cachés (×100 pour plus de lisibilité)

8

États cachés et transitions :

Après chaque entraînement d'un HCRF il y a pour chaque label y au moins un état h_y avec une compatibilité $\theta_s(y,h_y)$ fortement positive avec le label y et fortement négative avec l'autre label $\overline{y}$ ($\theta_s(\overline{y},h_y)$ très négatif). On appellera ces états '*état négatif*' (*Neg*) et '*état positif*' (*Pos*) par abus de langage. Les transitions entre deux états compatibles avec des labels différents sont très peu probables. Les trois autres états sont considérés 'neutres' (*Neu1*, *Neu2*, *Neu3*) avec de faibles amplitudes pour les poids des fonctions caractéristiques de transitions et de compatibilités. Ces états peuvent être utilisés comme transition entre un état positif et un état négatif. Une représentation de plusieurs états avec les différents poids de transitions entre chaque état est visible en figure 8.1.

Mots d'activations :

Dans le Tableau 8.3, nous présentons les exemples les plus pertinents des descripteurs les plus compatibles pour chaque état caché (sélectionnés parmi les 30 poids les plus importants). Dans la première colonne nous pouvons voir que les descripteurs linguistiques et paralinguistiques ont une importance moindre pour les états neutres : le seul descripteur ayant un poids positif pour tous les états neutres est '**chuckling**'

TABLE 8.3: Exemples intéressants de descripteurs avec de fortes compatibilités avec chacun des états (paralinguistique entre *)

États	**Descripteurs linguistiques et paralinguistiques**	**Mots correspondants aux vecteurs fortement compatibles**
Pos	adj, disfluency, conjunction, intensifier, *lip smacking*, ...	*honors, fearless, awesome, fantastic*
Neu1	*chuckling*	*um, Uh, ah, dunno, nada*
Neu2	*chuckling*	*um, Uh, ah, dunno, nada*
Neu3	∅	*Thanks, justin, sean, michael, Sorry*
Neg	negation, *falling intonation*, interjection, *word elongation*, ...	*miserably, disappointing, yelling, failure, lack*

alors que *Pos* et *Neg* ont de nombreux et variés descripteurs linguistiques et paralinguistiques avec des poids positifs élevés.

Du côté du plongement de mots, notre système n'apprend plus des mots spécifiques comme avec le BoNG mais des formes dans l'espace 300-dimensionnel word2vec en utilisant l'information contenue dans les mots-vecteurs. Pour l'analyse des descripteurs de l'espace word2vec, regarder les dimensions qui ont les poids les plus importants est abstrait. Pour cela, nous cherchons les mots-vecteurs contenus dans notre corpus qui activent le plus chaque état. Afin d'analyser les activations spécifiques à chaque état dans l'espace de dimension 300 de word2vec, regarder les poids un à un indépendamment ne permet pas d'analyse globale. Pour pallier ce problème, nous avons donc regardé les mots qui étaient les plus compatibles avec chaque état caché. Pour cela, nous avons calculé le score que donnait chaque mot de notre vocabulaire lorsqu'il apparaît comme observation en faisant le produit scalaire du mot avec le vecteurs des poids associés à la représentation distribuée.

Dans la seconde colonne du Tableau 8.3, nous pouvons voir les mots d'activations avec de fortes valences, tels que '*disappointing*', '*miserably*' et '*awesome*'. On n'apprend plus à reconnaître des mots mais des endroits de l'espace word2vec qui correspondent à des ensembles de mots partageant de caractéristiques (entre autres) sémantiques en commun comme les champs lexicaux ou des concepts. Ces caractéristiques sémantiques contenues à ces position de l'espace correspondent bien aux états positif ou négatifs. Cela permet au système de savoir retrouver si un mot est positif même si le système ne l'a vu que quelquefois dans des documents négatifs, ce qui ne serait pas le cas avec une représentation en BoNG. De même, si le système voit un mot sémantiquement proche de mots à connotations positives, il saura le reconnaître comme positif. Une visualisation en nuages de mots des UIP du corpus les plus compatibles avec chaque état est visible en figure 8.2.

Un des inconvénients est une trop grande sensibilité aux typographies : en effet

TABLE 8.4: Exemples de différences dans les valeurs des fonctions de descripteurs

Mots	État positif	États neutres	État négatif
lantern	0.23	0.46	-0.41
Lantern	4.70	-0.40	-3.53
Uh	-2.57	2.8	-3.86
Yeah	0.98	2.21	-5.49
Yes	-0.06	1.21	-3.14
Thanks	2.25	3.48	-7.41

on peut voir dans le tableau 8.4 que les vecteurs des mots *lantern* et *Lantern* n'ont pas les mêmes effets sur les états cachés. Ceci est dû à la différence sémantique entre une simple lanterne et le héros Marvel la Lanterne Verte. Cependant, les transcripteurs n'ont pas toujours mis les majuscules.

Rôle des états neutres : Adaptation au domaine :

Apprendre un modèle de plongements de mots nécessite une quantité importante de données textuelles disponible, et c'est pour cette raison que nous avons choisi d'utiliser un modèle pré-entraîné. Le système semble prendre en compte le problème de décalage entre la sémantique contenue dans les vecteurs et leur véritable sens lorsque les mots sont utilisés dans le corpus. Les états cachés neutres du HCRF semblent prendre en compte les mots-vecteurs problématiques afin qu'ils n'affectent pas les états liées avec le label global du document.

Adaptation à l'oral :

Il est intéressant de noter que même si les mots-vecteurs utilisés ont été appris sur des données textuelles générales, le vocabulaire contient des mots de parole spontanée tels que '*uhm*' ou '*dunno*'. Cependant, leurs vecteurs ne correspondent pas sémantiquement à ceux qui auraient été appris sur un monologue audio comme les critiques orales analysées dans ce travail. Par exemple, un '*uhm*' écrit tend à avoir un effet stylistique négatif tandis que son équivalent oral est une simple hésitation et est possiblement neutre. Un autre exemple est la différence entre '*yes*' et '*yeah*' : ce dernier n'étant pas commun à l'écrit où il reflète une pensée plus positive (voir Tableau 8.4).

Il est à noter que pour 2 des 3 états neutres (*Neu1* et *Neu2*), lorsqu'on prend les 20 mots les plus compatibles, on observe un lien fort avec le langage urbain propre à l'oralité : '*Uh*', '*nada*', '*dunno*', '*hah*', '*somethin*', '*d*ck*', '*wh*re*', '*vagina*', '*yay*, '*ass*', '*howdy*'

et '*shee*'.

Adaptation aux critiques de film :

Nous nous plaçons ici dans un domaine particulier qui est la critique de film. Comme toutes les tâches spécifiques, on compte un vocabulaire particulier. Comme pour les mots oraux à l'écrit (voir paragraphe précédent) certains vecteurs de mots qui seront utilisés posséderont une place dans l'espace qui ne reflétera pas le sens donné par le locuteur. Par exemple, d'autres mots qui sont spécifiques au corpus comme '*Hi*' ou '*Thanks*' ("*Thanks for watching me guys*") sont associés à une valence positive via leur vecteur word2vec entraîné sur de l'écrit. Par conséquent, l'information contenue dans les mots-vecteurs ne correspond pas forcément au discours du locuteur. De même, on peut aussi noter que les noms propres comme *justin, sean, michael*[3] sont très compatibles avec les états neutres (voir tableau 8.3).

Conclusion

Pour conclure cette partie, on peut voir les états neutres comme des états absorbant les valeurs aberrantes des mots-vecteurs par rapport aux données. Ils permettent une adaptation de mots-vecteurs généraux appris sur une grande quantité de texte, au domaine des critiques de film à l'oral.

3. sans majuscule dans notre corpus

(a) État **négatif**

(b) État **neutre**1

(c) État **neutre**2

(d) État **neutre**3

(e) État **positif**

FIGURE 8.2: Visualisation en nuage de mots des UIP ayant les vecteurs les plus compatibles avec chaque état

9

Expériences : Analyse de l'opinion inter-locuteurs

Résumé du chapitre

— Des expériences ont été menées à l'aide des données textuelles uniquement. Les représentations construites par nos soins sont comparées avec des représentations en Sac-de-N-grammes et les HCRF sont comparés avec une régression logistique et un LSTM.

— Le meilleur modèle est celui utilisant une information sur la nature du locuteur qui parle, avec des poids partagés entre les features des différents locuteurs.

— L'utilisation de l'algorithme de Viterbi pour prendre en compte uniquement le label de la séquence la plus probable permet d'augmenter les résultats et de visualiser la séquence d'états cachés optimale.

Dans ce chapitre, nous présentons nos travaux pour la modélisation de la dynamique de l'opinion dans le cadre d'une interaction humaine au niveau d'une paire de tours de parole (Paire Adjacente : PA). Nous allons ici parler du système proposé, ses résultats dans les expériences menées sur le corpus SEMAINE-Léger et une analyse des performances abordant les descripteurs utilisés pour la linguistique et le modèle de Champs Aléatoire Conditionnels Cachés.

9.1 DANS UNE PAIRE ADJACENTE D'UNE CONVERSATION : ÉTUDE SUR LE CORPUS SEMAINE-LÉGER

Nous avons testé 4 modèles avec différents ensembles de descripteurs et segmentations. Premièrement, nous avons créé un système de base pour notre tâche à l'aide d'un modèle de régression logistique (RL) couplé à des descripteurs type BoNG suivant le protocole de Wöllmer et collab. (2013b) pour une tâche de classification binaire d'analyse de sentiment. Ensuite, comme dans l'expérience précédente en Sous section 8.1.2, nous avons remplacé les BoNG par notre ensemble de descripteurs (voir section 6.1.5). Afin de prendre en compte l'asymétrie de l'interaction nous avons testé 4 configurations différentes pour les descripteurs (voir sous-section 9.1.2). La RL ne prenant pas en compte la dynamique des observations, ce qui est important pour la parole. Nous avons alors utilisé comme deuxième système de base un modèle séquentiel à l'état de l'art : un réseau de neurones récurrent (RNN-LSTM) (Hochreiter et Schmidhuber, 1997). Comparé à ces modèles, les HCRF autorisent une plus grande interprétation tout en nécessitant moins de données pour l'entraînement. De plus, ils permettent de modéliser la dynamique des phénomènes liés a l'opinion (états émotionnels, attitudes sociales, etc...) à travers les états latents.

Afin de valider nos modèles, nous avons utilisé une validation croisée à 10 tours où les ensembles d'entraînement et de tests sont disjoints et ne contiennent pas d'exemples provenant d'une même session.

9.1.1 Base de données

La base de données que nous avons utilisée pour cette expérience est le corpus annoté en attitudes que nous avons appelé *SEMAINE-Léger* (présenté en section 5.2). Des statistiques récapitulatives sur la base de données sont disponibles dans le tableau 9.1.

Pour une tâche de classification binaire, cela correspond à 424 PA et **8880** mots (observations x_k). Pour rappel, les annotations d'attitudes (définition en sous-section 1.2.1) ont été faites sans l'information audiovisuelle.

TABLE 9.1: Statistiques sur le corpus *SEMAINE-Léger* utilisé pour une analyse d'opinion inter-locuteurs

Locuteur	**Mots**	**Mots/tour de parole**				
		Moyenne	Médiane	Variance	1er décile	9eme décile
Agent	3331	8	6	7	1	16
Utilisateur	5549	13	9	15	2	26
Tous	8880	21	17	17	7	38

9.1.2 Systèmes de base avec RL et LSTM

Méthodologie :

Nous avons considéré plusieurs systèmes de base avec un ensemble de descripteurs textuels simples et avec notre ensemble de descripteurs. Ces systèmes ont été testés à différents niveaux de représentation textuelles (au niveau de la PA ou utilisant chaque mot comme observation). Nous avons premièrement utilisé une régression logistique (RL) comme système de base simple en expérimentant différentes stratégies pour prendre en compte les interactions entre l'agent et l'utilisateur. Pour la RL, nous avons testé 4 configurations. Nous avons intégré les descripteurs sur toute la PA, sur l'énoncé de l'agent ou sur celui de l'utilisateur, ou sur les deux contenant les résultats.

Nous avons ensuite utilisé un ensemble de descripteurs plus sophistiqué, avec une représentation utilisant un modèle sémantique distributionnel de Mikolov et collab. (2013b) décrit dans la sous-section 6.1.3. Après une tokenisation [1] nous avons utilisé un correcteur orthographique [2] afin de gérer les nombreuses erreurs contenues dans les transcriptions manuelles et ainsi nettoyer le texte avant d'importer le vecteur du mot associé. Le corpus utilisé ayant été collecté au Royaume-Uni, les transcriptions des conversations contiennent des orthographes britanniques. Cependant, comme les vecteurs utilisés pour la représentation distribuées sont uniquement des orthographes américaines, nous avons utilisé un lexique pour passer de l'orthographe britannique à l'orthographe américaine. [3]

Si l'utilisation de la moyenne pour intégrer les mots-vecteurs est un système classique (Yang et collab., 2016), nous avons poussé plus loin en essayant différentes stratégies pour intégrer les mots-vecteurs sur la PA à l'aide de 4 fonctionnelles (moyenne arithmétique, médiane, maximum et minimum). Ceci nous a permis d'obtenir un vecteur de la même taille qu'un vecteur d'un mot, avant de normaliser sur chaque dimension. Nous avons pris l'agrégation permettant d'obtenir le meilleur résultats afin de

1. Nous avons utilisé le CoreNLP de Standford Schuster et Manning (2016)
2. https ://github.com/phatpiglet/autocorrect
3. http ://tysto.com/uk-us-spelling-list.html

maximiser les performances de la régression logistique. A cela, et pour aider la détermination de l'attitude nous avons ajouté les *descripteurs linguistiques et lexicaux* (comme décrit dans la section 6.1). Pour le LSTM, nous avons concaténé les descripteurs de l'agent et de l'utilisateur en segmentant mot par mot, comme pour le HCRF-1 (voir fig. 7.3).

La méthodologie utilisée pour trouver les meilleurs hyperparamètres a été la même que précédemment. Par rapport au réglage des hyperparamètres de la régression logistique, nous avons effectué des entrainements avec des valeurs de C, qui correspond à l'inverse du coefficient de régularisation, dans {0.1, 0.5, 1, 10, 100}. Nous avons utilisé l'implémentation de *scikit-learn* de Pedregosa et collab. (2012) pour la régression logistique. Pour le RNN-LSTM, nous avons utilisé l'implémentation de Keras (Chollet, 2015) avec un nombre d'états cachés variant dans {64, 128, 256}, une régularisation avec une valeur de *dropout* (Srivastava et collab., 2014) de U et W (voir Hochreiter et Schmidhuber (1997)) dans {0.1, 0.2, 0.3} (si plus élevé, les performances chutent) et un nombre d'époque dans {4...10}. Nous avons utilisé l'entropie croisée comme fonction de coût et Adam comme algorithme d'optimisation (Kingma et Ba, 2014).

Nous avons fait tous les efforts nécessaires pour entraîner le RNN-LSTM (voir ci-dessus), et avons envisagé l'apprentissage par transfert (*transfer learning*) suivi d'un réglage fin des hyperparamètres (*fine-tuning*). Les résultats étaient décevants avec un entrainement sur IMDb. Nous avons conclu que les données manquaient car la tâche considérée est atypique. La comparaison avec d'autres modèles utilisant du contexte comme celui de Ghosh et collab. (2017) a été écartée du fait de la taille réduite de notre base de données.

Résultats :

Les résultats des systèmes de base sont listés dans la première partie du tableau 9.3 recensant les scores F1 et le taux de reconnaissance (*Accuracy*). Dans ce tableau, le $F1$ global (moyenne harmonique du rappel et de la précision) est la moyenne arithmétique des $F1$ de chaque classe ($F1+$ et $F1-$) et l'$Accuracy$ est le pourcentage de vraies prédictions (voir en annexe F pour plus de précisions). Comme prévu, la représentation en BoNG ne convient pas pour des documents courts (Benamara et collab., 2016) tels que l'énoncé de l'utilisateur ou la PA. Le modèle BoNG est très peu efficace pour la classe négative. L'utilisation de notre ensemble de descripteurs contenant des descripteurs liés aux sentiments et des descripteurs linguistiques permet une amélioration sur ce point. Les résultats du RNN-LSTM sont meilleurs pour la classe négative. Bien qu'il ait le potentiel de capturer une certaine dynamique, le réseau de neurones requiert plus de données que disponible dans le corpus considéré pour être pleinement efficace.

TABLE 9.2: Ensemble des hyperparamètres testés lors de l'entraînement de nos modèles

Paramètre	Nom	Valeurs testées
λ_2	Coefficient de régularisation	{1, 5, 75, 0.1, 0.25, 0.5, 1}
$\lvert\mathcal{H}\rvert$	Nombre d'états cachés	{2, 3, 4, 5, 6, 7}
w	Taille de la fenêtre de contexte	{0, 1, 2}

9.1.3 Modèles HCRF

Méthodologie :

La présence d'états cachés dans les HCRF les rend aptes à modéliser une dynamique sur des phénomènes connexes aux opinions comme les émotions ou les atttitudes sociales d'un locuteur. De plus, à l'aide de notre ensemble de descripteurs incluant descripteurs liés aux opinions et une représentation distribuée, on s'attend à ce que le modèle exploite les informations contenues dans les vecteurs pré-entraînés et exploite plus profondément les concepts employés par les locuteurs. Nous avons utilisé les 4 configurations présentées en figure 7.3 et expliquées sous-section 7.2.2.1. Le réglage des hyperparamètres et l'implémentation du modèle suivent la même méthodologie que précédemment. Nous avons entraîné les modèles HCRF avec le *wrapper* Matlab de *HCRF Library* Morency (2007) et utilisé un algorithme de type L-BFGS durant l'entraînement afin de maximiser la log-vraisemblance. Par rapport aux hyperparamètres des modèles, nous avons testé différentes valeurs qui sont résumés dans le tableau 9.2. La taille de la fenêtre de contexte est le nombre d'observations que l'on a concaténé autour de notre observation centré. Lorsque l'observation n'avait pas de voisin, une marge (*padding*) avec des valeurs nulles a été utilisée. Nous avons fait des tests préliminaires avec plus d'états cachés et des plus hautes valeurs de coefficient de régularisation, mais les résultats n'étaient pas meilleurs et le temps d'entraînement significativement plus long.

Résultats :

Les résultats des modèles HCRF sont résumés dans le tableau 9.3. Utiliser la classe de la séquence globale la plus probable permet d'améliorer les résultats, comparé à une approche classique (voir sous-section 2.3).

Le meilleur score F1 a été atteint à l'aide de la quatrième configuration avec 6 états cachés possibles ($\lvert\mathcal{H}\rvert$), un coefficient de régularisation ℓ_2 égal à 5 et une fenêtre de contexte de taille 2. Cette configuration prend en compte la dynamique de l'interaction en distinguant les différents locuteurs par un descripteur indiquant le rôle de la

TABLE 9.3: score F1 et d'Accuracy avec différents descripteurs et modèles

Descripteurs	Modèle	F1+	F1-	F1	Acc
Vote majoritaire	Factice	79	0	40	66
BoNG	RL	73	36	55	62
Nos descripteurs	RL	70	39	55	60
Nos descripteurs	LSTM	83	64	74	77
BoNG	HCRF-1	80	55	67	72
Nos descripteurs	HCRF-0	82	67	74	77
Nos descripteurs	HCRF-1	84	74	79	80
Nos descripteurs	HCRF-2	84	73	78	80
Nos descripteurs	HCRF-3	**86**	**73**	**79**	**81**
Nos descripteurs	HCRF-4	**86**	**75**	**80**	**82**

personne qui parle (voir fig. 7.3). Notre ensemble de descripteurs permet d'améliorer le score F1 sur la classe négative pour le HCRF-1 passant de 55 à 67. Nous comparons notre approche à un modèle classique d'analyse de sentiment prenant uniquement en compte le locuteur analysé et pouvons voir une nette amélioration par rapport au modèle HCRF-0 : le score F1 passe de 74 à 80. Les HCRF sont particulièrement bons sur les documents négatifs par rapport aux systèmes de base.

Bien qu'une validation croisée à 10 tours sur un corpus de ce type ne permette pas de conclure qu'une différence de performance soit statistiquement significative l'une par rapport à l'autre, nous prenons les résultats comme très prometteurs et des efforts sont actuellement faits afin d'annoter plus de données et obtenir une validation plus solide.

9.1.4 Discussion et analyse des résultats

Le meilleur système est celui utilisant un descripteur indiquant quel est le locuteur en train de parler (agent ou utilisateur).

$$\Theta_{s,neg} = \begin{pmatrix} -1.4 \\ -1.3 \\ \mathbf{15.3} \\ -1.4 \\ \mathbf{-16} \\ -1.4 \end{pmatrix} \qquad \Theta_{t,neg} = \begin{pmatrix} -0.2 & -0.2 & 1.3 & -0.2 & -1.7 & -0.2 \\ -0.2 & -0.2 & 1.3 & -1.7 & -0.2 & -0.2 \\ 1.3 & 1.3 & \mathbf{11} & 1.3 & -0.9 & 1.3 \\ -0.2 & -0.2 & 1.3 & -0.2 & -1.7 & -0.2 \\ -1.6 & -1.6 & -0.7 & -1.6 & \mathbf{-7.4} & -1.6 \\ -0.2 & -0.2 & 1.3 & -0.2 & -1.7 & -0.2 \end{pmatrix}$$

(a) Poids d'état et de transitions avec le label **négatif**

$$\Theta_{s,pos} = \begin{pmatrix} -0.2 \\ -0.2 \\ \mathbf{-13} \\ -0.2 \\ \mathbf{20} \\ -0.2 \end{pmatrix} \qquad \Theta_{t,pos} = \begin{pmatrix} -0.1 & -0.1 & -1.5 & -0.1 & 1.5 & -0.1 \\ -0.1 & -0.1 & -1.5 & -0.1 & 1.5 & -0.1 \\ -1.5 & -1.5 & \mathbf{-6.4} & -1.5 & -0.4 & -1.5 \\ -0.1 & -0.1 & -1.5 & -0.1 & 1.5 & -0.1 \\ 1.5 & 1.5 & -0.6 & 1.5 & \mathbf{13} & 1.5 \\ -0.1 & -0.1 & -1.5 & -0.1 & 1.5 & -0.1 \end{pmatrix}$$

(b) Poids d'état et de transitions avec le label **positif**

FIGURE 9.1: Vecteurs de poids d'état et matrices de poids de transitions

9.1.4.1 États cachés et transitions :

Rappel :
Pour tous les modèles, lorsque nous avons analysé les poids du système, nous avons toujours remarqué qu'on obtenait après l'entraînement 3 types d'états différents :

- Un état très compatible avec le label positif et très incompatible avec le label négatif, qu'on appellera *Pos* par la suite
- Un état très compatible avec le label négatif et très incompatible avec le label positif, qu'on appellera *Neg* par la suite
- Le reste des états étant ni compatible ni incompatible avec chaque label, qu'on appellera *Neu* par la suite

Une grande compatibilité (respectivement incompatibilité) entre un état et un label signifie un poids d'état θ_s de grande amplitude et de signe positif (respectivement négatif). Les états neutres qui ne sont ni compatibles ni incompatibles avec les 2 labels ont des poids θ_s ayant des valeurs proches de zéro. Dans la suite nous parlerons d'états neutres, positif et négatif.

Pour plus d'information sur ces poids θ_s, le lecteur est renvoyé à la sous-section 7.1.2. Un exemple de matrice de poids après entraînement est donné en figure 9.1.

On verra plus loin dans la sous-section 9.1.4.5 que les poids associés avec ces états étaient aussi très spécifiques. Les poids d'observations θ_o ayant les plus hautes valeurs

avec l'état "positif" (respectivement "négatif") sont les valeurs positives (respectivement négatives) des lexiques de subjectivité. Pour les états neutres, ce sont les valeurs objectives (par exemple si un mot est généralement neutre, comme "tableau").

9.1.4.2 Confrontation au modèle de base

Tout d'abord, il est important de remarquer que le modèle HCRF-0 n'utilisant que le tour de parole de l'utilisateur n'est pas très performant sur les exemples négatifs : utiliser le contexte d'une PA est donc jugé sensiblement important. Ce résultat n'est pas surprenant dans le sens où l'annotation a été effectuée par un humain voyant une Paire Adjacente et non un simple tour de parole. Sans visibilité de la première partie de la paire adjacente, il peut être difficile pour le système de donner une prévision alors qu'aucune information d'opinion n'est visible dans le tour de parole de l'utilisateur.

9.1.4.3 Partage des poids

Lorsque l'on compare les différents modèles d'interactions (descriptions des différents modèles dans la sous-section 7.2.2.1), on observe que le modèle 2 qui apprend des poids différents par locuteur obtient des résultats médiocres comparé aux autres modèles. Il semble donc que partager les poids entre les 2 locuteurs soit la meilleure option. Le corpus est trop petit pour apprendre 2 fois plus de poids, expliquant les résultats médiocres du modèle numéro 2.

9.1.4.4 Séparation des locuteurs

Il est important de différencier les 2 locuteurs et l'utilisation d'un descripteur spécial ou d'une intégration particulière des descripteurs comme dans les modèles 3 et 4 est un plus et permet d'obtenir des scores plus élevés. En effet, nous allons montrer que ces deux modèles permettent d'incorporer le tour de parole de l'agent sans pour autant lui donner autant d'importance que dans le modèle 1.

Le modèle 3 permet de donner plus d'importance à l'utilisateur qu'à l'agent en intégrant les représentations de l'agent en une seule observation (voir sous-section 7.2.2.1). Le modèle 4 utilise quant à lui une indicatrice permettant au classifieur de savoir qui parle. L'indicatrice de l'utilisateur est très compatible avec les états positifs et négatifs et très incompatibles avec les états neutres (voir dans les tableaux 9.4 et 9.5). En d'autres termes, l'indicatrice permet de faire pencher le système vers les états positif et négatif qui influenceront la décision finale lorsque l'utilisateur est en train de parler, tout en laissant l'agent influencer la décision finale au préalable.

9.1.4.5 Activations des états cachés

On a dit précédemment dans la sous-sous-section 9.1.4.1 qu'on obtenait après l'entraînement un état positif, un état négatif et des états neutres d'après leurs compatibilités avec les labels positifs et négatifs venant des poids θ_s et θ_t. Nous allons ici nous intéresser aux poids d'observations θ_o de ces différents états. Pour voir les paramètres les plus importants pour chaque état nous pouvons regarder les amplitudes des différents poids. Une forte amplitude positive (respectivement négative) signifie que le paramètre est très compatible (respectivement incompatible) avec l'état.

Nous pouvons séparer les paramètres en 2 types : les mots-vecteurs et les autres paramètres. En effet, les mots-vecteurs sont des représentations de mots en dimension 300, il y a donc 300 poids différents associés à cette représentation. Pour les autres paramètres, les poids représentent des occurrences discrètes, comme par exemple l'utilisation d'un adjectif, et un seul poids est associé par occurrence.

Paramètres simples à un poids : En analysant les poids d'observations θ_o des descripteurs linguistiques, nous pouvons observer que les descripteurs états dits positifs (respectivement négatifs, neutres) qui sont compatibles avec les labels positifs (respectivement négatifs, neutres) sont représentatifs de valeurs de valence positives (respectivement négatives, neutres).

Pour l'état positif : Les valeurs positives venant du SentiWordNet, du SO-CAL et du ANEW, la présence de verbes et de noms communs sont des observations qui poussent le HCRF à aller dans l'état positif associé à ces observations. Ce sont des valeurs liées à des énoncés subjectifs positifs.

Pour l'état négatif : Les valeurs négatives venant du SentiWordNet et et du lexique de subjectivité ANEW, la présence de négations et d'adverbes sont des observations qui poussent le HCRF à aller dans l'état négatif associé à ces observations.

Pour les états neutres : Les valeurs neutres venant du SentiWordNet, du SO-CAL, et du lexique de subjectivité ANEW, ainsi qu'une valeur de dominance élevée du lexique ANEW sont des observations qui poussent le HCRF à aller dans l'état neutre associé à ces observations.

Mots-vecteurs : Comme expliqué dans la sous-section 8.1.4, les dimensions de l'espace word2vec étant abstraites, nous avons calculé le score que donnait chaque mot de notre vocabulaire lorsqu'il apparaît comme observation en faisant le produit scalaire du mot avec le vecteur des poids associés à la représentation distribuée. Une visualisation en nuages de mots des mots les plus compatibles avec chaque état est visible en figure 9.2.

(a) État **négatif**

(b) État **neutre**

(c) État **positif**

FIGURE 9.2: Visualisation en nuage de mots des mots ayant les vecteurs les plus compatibles avec chaque état

TABLE 9.4: Les descripteurs avec de fortes **compatibilités** avec chacun des états

États	**Descripteurs linguistiques**	**Mots correspondant aux vecteurs fortement compatibles**
Pos	SWN.pos, ANEW_pos, SO_pos, is_user, postag=VB, postag=NN	*delighted, thank, congratulations, fantastic, enjoy, loves, smiling, sunny, entertained*
Neu	SWN.obj, ANEW_neut, SO_neut, ANEW_Dom	*anybody, you, yourself , anyone, anymore, somebody, do, anything, nobody, em*
Neg	ANEW_neg, SWN.neg, negation=True, is_user, adv=True	*disorganized, venting, frustration, ineffective, incident, frustrated, feelings, defeatist, angry, unmotivated, harsh*

Pour résumer :

Les états neutres absorbent un peu de ce que dit l'agent, l'indicatrice de l'utilisateur forçcant à aller vers des états positifs ou négatifs qui sont impactants pour la décision finale (les valeurs des θ_s étant grandes). Les états neutres sont activés par des mots de valence neutre, ces mêmes mots étant incompatibles avec les états positifs et négatifs. Le HCRF permet une clusterisation des mots du vocabulaire avec les états latents, comme vu dans le chapitre précédent. Dans cette expérience, le HCRF tend vers une clusterisation où les mots neutres et les moments où l'agent parle sont dans la même catégorie.

9.1.4.6 Étiquetage de Viterbi

L'étiquetage de Viterbi est un étiquetage prenant la classe de la séquence d'états cachés la plus probable, liée au label le plus probable $y*$ de l'équation 9.1. En utilisant cet étiquetage, on améliore nettement les performances sur la classe négative, surpassant le score F1 du LSTM par 11 points (de 64 à 75).

$$y*, \mathbf{h*} = \arg\max_{y,\mathbf{h}} P(\mathbf{h}|y,\mathbf{x},\theta) \tag{9.1}$$

Nous avons remarqué que lorsque l'on s'intéressait à l'étiquetage de Viterbi (voir sous-section 7.1.4), il était possible de visualiser le changement de valence dans la phrase du locuteur au cours de l'énonciation (voir Figure 9.3). De plus, il est important de noter que le poids de l'indicatrice était hautement plus compatible avec les états positifs et négatifs qu'avec les états neutres, expliquant les améliorations du mo-

TABLE 9.5: Les descripteurs avec de fortes **incompatibilités** avec chacun des états

États	**Descripteurs linguistiques**	**Mots correspondant aux vecteurs fortement incompatibles**
Pos	SWN.obj, ANEW_neg, ANEW_neu, SO_neu, postag=EX, postag=CD	*disorganized, complain, circumstances, explain, anything, terribly, happening, anyone, seriously, extreme, anybody*
Neu	SWN.obj, ANEW_neut, SO_neut, ANEW_Dom, postag=WP	*optimism, gloomy, disappointment, frustration, optimistic, hopeful* sadness congratulations
Neg	ANEW_pos, ANEW_Dom, SO_pos, postag=TO, SWN.obj	*delighted, thank, lookin, love, prefer, holidaying, queen, fantastic, enjoy*

Agent : That's good
User : I don't like this weather

FIGURE 9.3: Un exemple d'étiquetage de PA (vert/rouge signifie que l'état est compatible avec label positif/négatif)

dèle HCRF-4 sur HCRF-1. Ceci peut s'expliquer par le fait qu'il est difficile de trouver les poids importants pour chaque état car les agents sont émotionnellement colorés et ne sont pas forcement émotionnellement alignés avec l'attitude des utilisateurs. Cette valence autre que la valence à détecter peut être perçue comme du bruit dans certains exemples. L'indicatrice peut être vue comme un artefact permettant de gérer ce bruit dans un modèle utilisant des poids partagés.

9.1.5 Conclusion et futurs travaux

Dans ce chapitre, nous avons présenté des modèles HCRF utilisant un contexte interactionnel dans le but de détecter une attitude chez un utilisateur au sein d'une interaction humain-agent. Notre ensemble de descripteurs inclut une représentation distribuée de mots, des règles linguistiques et des lexiques de subjectivité. L'utilisation de classifieurs HCRF permet d'apprendre des représentations linguistiques locales de chaque mot de la transcription afin de modéliser un processus dynamique. Les modèles étudiés utilisent le contexte interactionnel afin d'analyser de manière plus subtile l'utilisateur dans une interaction humain-agent. Nous avons testé plusieurs configurations afin de prendre en compte les informations contenues dans le tour de parole de

l'agent pour l'analyse de l'attitude de l'utilisateur.
Dans nos travaux futurs, il serait utile d'étudier des descripteurs spécifiques à l'agent comme des actes de dialogue et des structures linguistiques plus complexes distribuées sur les différents locuteurs. Il serait aussi intéressant d'utiliser une méthode plus élaborée pour l'intégration des mots-vecteurs sur les tours de paroles de l'agent, par exemple en utilisant une méthode de *paragraph embedding*. On remarque aussi que les états cachés permettraient l'association d'une observation à un état, cependant, ils sont limités par la longueur des observations. L'utilisation de règles distribuées sur plusieurs observations permettent au modèle d'intégrer un intensifieur suivi d'un mot de valence, possibilité qui n'est pas offerte par les mots vecteurs. Pour utiliser plus de contexte, il serait possible d'utiliser une modification des descripteurs par rapport au contexte, comme proposé par Poria et collab. (2017b). Finalement, nous souhaiterions aussi augmenter la taille du corpus afin d'obtenir des résultats significatifs.

10 Expériences multimodales : verbal et vocal

Résumé du chapitre

— Sur les expériences précédentes, les modèles multimodaux utilisant notre ensemble de descripteurs audio, une représentation apprise à l'aide d'un auto-encodeur séquentiel et une fusion de l'audio et du texte à l'aide d'un LSTM n'obtiennent pas des scores plus importants que les modèles textuels. La quantité des descripteurs audio est une première piste quant à la non augmentation des performances des systèmes multimodaux.

— Une sélection de paramètres a été faite avec un modèle de *Tree Boosting* sur le corpus SEMAINE-Léger. La sélection des 10, 20 et 30 paramètres les plus importants ne permet pas d'augmenter les résultats des modèles multimodaux qui tendent vers ceux des modèles textuels.

— Une autre sélection de paramètres a été faite avec l'utilisation d'une pénalisation de type *Elastic Net* sur les HCRF sur le corpus ICT-MMMO. Les apprentissages faits avec environ 60 descripteurs sélectionnés ne permettent pas une amélioration des résultats.

— Les faiblesses des modèles utilisant l'audio doivent être dues à la nature des représentations, incompatibles avec les HCRF. Si le HCRF est meilleur sur le texte, on observe qu'un modèle plus complexe permettant de faire des liens entre les représentations dans le temps comme un LSTM obtient de meilleurs résultats.

Dans les parties précédentes, nous avons testé nos modèles sur des données textuelles provenant de transcriptions de parole orale. Nous avons pu voir que nos modèles permettaient de :

— faire une adaptation de domaine de l'écrit à l'oral au niveau du style de parole ;
— modéliser la dynamique de l'opinion dans un discours ou un tour de parole ;
— intégrer les différents locuteurs afin de pouvoir utiliser le contexte interactionnel et ainsi obtenir de meilleures performances.

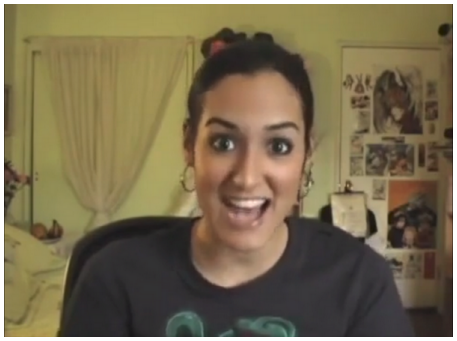
Une amélioration évidente de nos modèles est l'utilisation d'une modalité supplémentaire afin de pouvoir prendre en compte de plus riches **structures multimodales définissant l'opinion**. Ces structures plus riches contiennent plus d'information et permettent une désambiguïsation dans certains cas (par exemple lors de l'utilisation du sarcasme), et une confiance plus grande dans la valeur prédite. En effet, il y a plusieurs intérêts à utiliser un système multimodal : l'emploi simultané de différentes modalités peut être bénéfique du fait des principes de consensus et de complémentarité, décrits par Xu et collab. (2013), qui sont évoqués ci-dessous.

Principe de consensus Le principe de consensus vise à maximiser l'accord des différentes vues. Lorsque les informations contenues dans chacune des modalités sont caractéristiques d'une classe particulière de notre jeu de données, l'utilisation simultanée de ces deux modalités permet au système de faire une prédiction plus précise et de réduire son intervalle de confiance. Dans l'exemple de la figure 10.1a nous pouvons voir un exemple où les descripteurs de différentes modalités sont caractéristiques de la même classe.

Principe de complémentarité Le principe complémentaire stipule que, dans un environnement multi-vues, chaque vue des données peut contenir des informations que d'autres vues n'ont pas. Par conséquent, plusieurs vues peuvent être utilisées pour décrire de manière complète et précise un phénomène particulier. Par exemple, dans un cas où notre système unimodal hésiterait sur la prédiction finale, les différentes modalités peuvent apporter des informations qui sont complémentaires. Cela peut permettre de faire pencher la balance pour une des classes quand une modalité ne donne pas d'indice fort. Cela peut aussi permettre de détecter l'ironie dans le cas où les informations des modalités ne seraient pas consensuelles, avec une phrase positive et une intonation négative typique du sarcasme (Chen et collab., 2017), comme dans l'exemple en Figure 10.1b.

Dans tous nos travaux, nous nous sommes attachés à utiliser des données provenant de discussions orales. La modalité audio est donc présente dans chacune de nos bases de données et nous pouvons l'utiliser.

(a) Exemple de phrases où il y a consensus dans les informations des différentes modalités



"***I really liked the movie, big time, totally!***"

(b) Exemple de phrases où les différentes modalités sont complémentaires (Zadeh et collab., 2017)

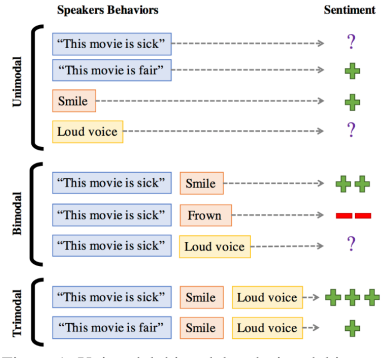


FIGURE 10.1: Exemple de consensus ou complémentarité

Dans ce chapitre nous allons présenter les différents résultats que nous avons obtenus en ajoutant le signal audio de plusieurs manières.

Dans la première section nous comparerons les performances des modèles utilisant uniquement l'audio, uniquement le texte et les deux modalités à la fois avec une fusion précoce, c'est-à-dire en concaténant les descripteurs en amont du classifieur. Nous avons testé différentes manières d'intégrer l'audio en représentations pertinentes, à l'aide des pauses du locuteur et des tours de parole. Les tâches d'analyse d'opinion inter-locuteur (sous-section 10.1.2) et intra-locuteur (sous-section 10.1.1) seront successivement abordées, et les modèles analysés. En découvrant que les performances de nos systèmes ne faisaient que baisser avec l'ajout de la modalité audio, nous avons ensuite comparé différentes méthodes de sélection de descripteurs, en section 10.2, afin de baisser l'importance du signal vocal tout en l'intégrant dans les données. Est ensuite étudiée une sélection de descripteurs, avec un *Gradient Boosting* sur des arbres de décision (sous-section 10.2.1), puis une sélection de descripteur via une

régularisation de type *Elastic net* appliquée à des HCRF (sous-section 10.2.2).

10.1 Fusion bimodale précoce et segmentation automatique de l'audio basée sur les pauses

Les phénomènes liées aux opinions n'ont pas la même temporalité en fonction de la modalité d'expression. En effet, la longueur caractéristique des structures de ces phénomènes n'est pas forcement la même pour l'audio et pour le texte. Cependant, une segmentation en pauses du locuteur a déjà été utilisée pour la modalité audio d'une tâche d'analyse de sentiment dans un long monologue (Wöllmer et collab., 2013b). Cette segmentation permet d'obtenir des unités pertinentes pour l'étude que nous menons. Pour l'étude de l'opinion inter-locuteur, nous étudions une fusion simple du texte et de l'audio en concaténant les descripteurs obtenus par des intégrations sur des unités inter-pausales. Comme pour le travail que nous avons mené sur le texte dans le chapitre 8, une segmentation en fonction de la longueur de la pause est étudiée. Pour l'étude de l'opinion inter-locuteur, les tours de parole des locuteurs sont plus courts que pour l'étude de l'opinion intra-locuteur (long monologue). Ceci rend la segmentation en pause moins efficace, diminuant fortement la taille des données (voir chapitre 9). Nous avons choisi de segmenter en fonction des mots la transcription et avons étudié plusieurs segmentations en fonction des pauses pour le signal vocal.

Dans la section actuelle, nous étudions les performances de nos modèles utilisant l'audio ainsi que l'impact de son ajout sur les performances de nos systèmes utilisant uniquement le texte. Nous pourrons donc comparer nos systèmes unimodaux utilisant le texte et l'audio avec des systèmes bimodaux utilisant ces deux modalités simultanément. Dans la première partie (sous-section 10.1.1) on s'intéressera à une analyse de l'opinion intra-locuteur comme ce qu'on a pu voir dans le chapitre 8. Dans la deuxième partie (sous-section 10.1.2) nous étudierons et analyserons les performances de notre système sur une tâche d'analyse d'opinion inter-locuteurs, comme ce qu'on a pu voir dans le chapitre 9.

10.1.1 Étude sur l'analyse d'opinion intra-locuteur

Dans cette sous-section, nous présenterons les résultats et analyses de nos modèles pour une tâche de classification binaire en sentiment sur les critiques de films de type vlog du corpus *ICT-MMMO* (présenté dans la sous-section 4.2.1). Plusieurs segmentations sont expérimentées, comme lors de l'étude unimodale utilisant uniquement la transcription. Le principal résultat ressortant de cette étude est que l'utilisation du signal vocal n'améliore pas les résultats par rapport à l'utilisation du signal verbal uniquement, et diminue même les performances du modèle. Après une présentation du

protocole et des résultats, on fera une courte analyse des performances des modèles utilisant les descripteurs acoustiques.

10.1.1.1 Protocole

Dans cette première étude multimodale, nous avons utilisé nos modèles de HCRF pour une tâche de classification binaire en sentiment sur les vidéos de type vlog du corpus *ICT-MMMO* (présenté dans la sous-section 4.2.1). Nous avons encore une fois testé plusieurs segmentations automatiques basées sur les pauses du locuteur. Le protocole d'entraînement a été strictement le même que dans la partie précédente utilisant uniquement le texte.

Une différence notable est l'utilisation d'une fusion tardive **pour la régression logistique uniquement**. Dans ce cas-ci, nous avons suivi le protocole de Wöllmer et collab. (2013b) et trouvé empiriquement les poids de pondérations α et $1-\alpha$ ($\alpha \in [0;1]$) entre les prédictions des classifieurs unimodaux permettant d'obtenir le score global le plus élevé.

10.1.1.2 Résultats

Les résultats des différents modèles unimodaux et multimodaux sont visibles dans le Tableau 10.1. On peut observer que l'utilisation de l'audio pour la tâche de classification de sentiment sur le corpus *ICT-MMMO* n'amène pas des résultats aussi bons qu'en utilisant le texte. Plus important que ça, on observe même une diminution des performances lorsque l'on compare le modèle unimodal utilisant le texte avec le modèle multimodal utilisant le texte et l'audio.

Les résultats concernant la régression logistique avec l'audio seul correspondent à nos attentes. En effet, on intègre les descripteurs avec des fonctions statistiques sur l'intégralité du discours, qui dure plusieurs minutes. Par conséquent, les représentations obtenues du signal n'intègrent pas pas la séquentialité de la parole. Les phénomènes liés aux opinions ont une temporalité importante à modéliser afin de produire des représentations pertinentes pour la tâche.

Pour les modèles séquentiels, les résultats sont plus surprenants. Lors de l'utilisation du signal verbal, les HCRF permettaient d'obtenir de meilleurs résultats que le LSTM, notamment pour les labels négatifs : on remarque un score F1 de 75% avec les HCRF par rapport à un score F1 de 68% avec un LSTM. On peut observer que cet écart de performance est inversé lors de l'utilisation du signal audio, et on observe en plus un nouvel écart pour les labels positifs. En effet, pour l'audio uniquement, le F1 néga-

tif est de 59% pour le HCRF par rapport à 65% pour le LSTM et le F1 total est de 72% pour le HCRF contre 77% pour le LSTM. On a donc de meilleurs résultats avec le LSTM qu'avec le HCRF sur l'audio.

Nous venons de voir que l'utilisation d'un HCRF donne de bons résultats pour le texte mais pas pour l'audio et l'inverse pour le LSTM. Cependant, les résultats du LSTM et du HCRF sont toujours moins bons en utilisant l'audio qu'en utilisant le texte uniquement. Par rapport aux expériences multimodales, on observe une diminution des performances par rapport au texte et une augmentation des performances par rapport à l'audio : tous les modèles sont moins bons en utilisant l'audio et le texte avec une fusion précoce. La régression logistique passe d'un F1 global de 79% à 73%, le LSTM de 78%+ à 78%- [1] et le HCRF, qui est le modèle fonctionnant le mieux sur le texte, passe de 82% à 76%.

Pour finir, il est important de noter une amélioration avec l'utilisation d'une fusion tardive pour la régression logistique qui passe d'un score F1 de 79 pour le texte à un 81 pour le texte et l'audio. L'utilisation de l'audio permet d' augmenter les performances lorsque les modèles appris restent propres à chaque modalité. Dans ce travail de thèse nous nous intéressons à une manière de détecter des structures multimodales et rejetons la possibilité d'utiliser un système avec une fusion tardive qui traite de manière séparée les différentes modalités et les rend indépendantes entre elles.

10.1.1.3 Analyse

La représentation du signal audio, plus brute et bas-niveau que celle du texte, on a deux fois plus de descripteurs (730 vs 340). En analysant les modèles multimodaux, on se rend compte que l'audio qui a beaucoup plus de descripteurs joue un rôle plus important que le texte dans la décision finale (tableau 10.2). Or l'audio étant moins performant en unimodal, l'impact se fait de manière négative sur les prédictions finales, possiblement à cause de l'abondance des descripteurs audio bas-niveau.

Cependant, même si le bruit apporté par la représentation du signal vocal peut être fort au niveau des états cachés, les prédictions s'équilibrent quand même sur la moyenne de la vidéo (exemple dans le tableau 10.3). On peut donc conclure que l'audio ne vient pas bruiter le texte jusqu'à ce que son effet disparaisse et c'est pour cela que les résultats du modèle multimodal restent meilleurs que le modèle unimodal audio.

Concernant les impacts mutuels des modalités, on voit dans le tableau 10.4 un exemple d'UIP où l'audio et le texte sont consensuels. À titre indicatif, en regardant la vidéo, on remarque que le pitch diminue et la voix est soufflée, marqueurs d'un

1. Les 2 valeurs sont 78%, cependant on voit le F1 négatif qui baisse un peu

TABLE 10.1: Scores de F1 et d'*Accurracy* (taux de bonnes réponses) avec les différents descripteurs, paliers de segmentation pour les pauses et modèles

Modalité	Modèle	Segmentation	F1+	F1-	F1	Acc
∅	Positif	∅	78	0	50	63
Audio	RL	vidéo entière	67	47	61	60
	LSTM	IPU (150ms)	82	62	76	76
		IPU (300ms)	83	63	76	77
		IPU (500ms)	**83**	**65**	**77**	**77**
	HCRF	IPU (150ms)	78	59	72	71
		IPU (300ms)	79	62	74	73
		IPU (500ms)	79	57	72	72
Texte	RL	vidéo entière	83	72	79	79
	LSTM	IPU (150ms)	84	68	78	78
		IPU (300ms)	84	68	78	78
		IPU (500ms)	83	65	77	78
	HCRF	IPU (150ms)	85	72	80	80
		IPU (300ms)	**86**	**75**	**82**	**82**
		IPU (500ms)	82	67	77	77
Audio-textuelle	RL (précoce)	vidéo entière	77	66	73	72
	RL (tardive)		84	73	81	80
	LSTM	IPU (150ms)	83	62	76	77
		IPU (300ms)	**84**	**65**	**78**	**78**
		IPU (500ms)	84	63	77	78
	HCRF	IPU (150ms)	82	65	76	76
		IPU (300ms)	82	65	76	76
		IPU (500ms)	83	63	76	76

TABLE 10.2: Poids des features les plus élevés pour les états positifs et négatifs toutes modalités confondus.

(a) Features les plus compatibles avec l'état caché **positif**

Représentation	Poids
phrase_type=VP	0.046693
postag=J	0.034319
intens=True	0.026922
postag=P	0.025930
F1_range25-75	0.025511
disfluency=True	0.024203
ΔLoudness_skew	0.022325
peakSlope_range25-75	0.021746
ΔMCEP_1_quant90	0.021717
F1_skew	0.020393

(b) Features les plus compatibles avec l'état caché **négatif**

Représentation	Poids
negation=True	0.053605
phrase_type=VP	0.037421
postag=R	0.033629
sentisynset.obj	0.032938
MFCC_8_skew	0.028809
ΔMFCC_11_mean	0.027731
ΔLoudness_median	0.026359
ΔMFCC_0_skew	0.024682
ΔMFCC_5_skew	0.022990
ΔpeakSlope_rangetot	0.022918

TABLE 10.3: Impact moyen des observations sur une vidéo par modalité et par état caché ($\sum_{j\in\{\text{Neg,Neu,Pos}\},x_a\in\text{audio}}\theta_o(h_j,\phi(x_a))$ et $\sum_{j=1\ldots L,x_t\in\text{texte}}\theta_o(h_j,\phi(x_t))$)

Modalité	Neg	Neu	Pos
Texte	0.210384	-0.444727	0.234352
Audio	0.331505	-0.171863	-0.159656

TABLE 10.4: Exemples de l'importance de chaque modalité sur la compatibilité d'une UIP avec les différents états cachés. UIP "*I just didn't feel like it.*" du fichier 1DmNV9C1hbY

État caché	Représentations	Impact des descripteurs
Neg	Tout	1.29966
Neu		-0.529525
Pos		-0.770019
Neg	W2V	0.394785
Neu		0.00295381
Pos		-0.397695
Neg	Linguistique	0.338279
Neu		-0.364461
Pos		0.026185
Neg	Audio	0.566595
Neu		-0.168018
Pos		-0.398509

sentiment négatif et le signal verbal est aussi négatif. En regardant la représentation du signal vocal pour cette UIP, on peut voir que les valeurs de la médiane du ΔH1H2, la moyenne du PSP et l'amplitude du MDQ sont importantes. Ces trois descripteurs représentent des indices de la qualité de la voix, permettant de discriminer une voix tendue, soufflée ou haletante.

Si des fois l'audio est consensuel et permet de retrouver le bon sentiment, ce n'est pas tout le temps le cas. Dans d'autres cas, l'audio induit en erreur. En effet les phénomènes liés aux opinions ne sont pas forcément les mieux détectables avec l'audio et ainsi les résultats sont toujours moins élevés qu'avec le texte pour des modèles unimodaux (Zadeh et collab. (2018c); Chen et collab. (2017) utilisent uniquement 5 descripteurs acoustiques par exemple). Dans la figure 10.2 se trouve un exemple d'UIP pour lequel le système donne un faux négatif. Le HCRF n'arrive pas à hiérarchiser les représentations et les modalités et à privilégier le texte quand il le faut. Ceci pose problème lorsque le texte est complémentaire à l'audio comme dans l'exemple ci-dessous (Figure 10.2).

Le LSTM permet d'obtenir les meilleurs résultats sur l'audio. Cependant, l'ajout de la modalité audio fait baisser les performances. Ayant plus de paramètres, il parvient mieux que le HCRF à modéliser les dépendances temporelles entre les différentes représentations du signal audio. Nous supposons que le fait que la représentation du signal audio ait deux fois plus de variables que le texte influe sur les performances du HCRF.

"***nasty. shocking. jawdropping***"

Modalité	Neg	Neu	Pos
Texte	0.016955	-0.555410	0.538425
Audio	0.201055	-0.166829	-0.034257

"***good.***"

Modalité	Neg	Neu	Pos
Texte	-0.890485	-0.304423	1.194984
audio	0.189300	-0.164616	-0.024727

FIGURE 10.2: Exemples de faux-négatif où le locuteur a le regard dans le vide, parle lentement et l'audio est compatible avec l'état négatif

Contrairement aux HCRF, le LSTM arrive à bien tirer parti de l'audio. Le HCRF modélise les variables de manières indépendantes et sans interactions entre elles à l'intérieur d'un état. Or la dynamique du signal vocal est plus complexe que celle du signal verbal et la structure que l'on essaie de reconnaître est complexe dans l'espace et un simple hyperplan ne suffit plus à associer ou non une observation à un état.

Une idée pourrait être de prendre en compte plus de contexte pour l'audio que pour le texte. Cependant, les expériences n'ont pas montré que le contexte de l'audio permettait d'obtenir de bons résultats... Nous supposons que c'est donc une structure complexe à distinguer, le problème ne pouvant pas se résoudre par une simple augmentation de la taille des IUP.

10.1.2 Étude sur l'analyse d'opinion inter-locuteurs

Dans cette sous-section, nous nous intéressons à l'ajout de l'audio pour l'analyse de l'opinion inter-locuteurs. Nous présentons les résultats et analyses des différents modèles implémentés pour une tâche de classification binaire en *attitude* dans des interactions multimodales humaines du corpus *SEMAINE-Léger* (présenté dans la section 5.2).

10.1.2.1 Protocole

Nous avons testé des segmentations différentes à la fois pour les modèles séquentiels et aussi pour la régression logistique que nous utilisons comme système de base. Pour la régression logistique, nous avons choisi de segmenter sur la paire adjacente entière et par rapport aux tours de parole de l'agent et de l'utilisateur.

Pour les modèles séquentiels comme les HCRF et LSTM, la segmentation en UIP a été écartée comme dit précédemment (sous-section 7.2.2) car elle diminue fortement la taille des données du fait de la courte durée des tours de parole. Par exemple, on ne souhaite pas avoir une paire adjacente constituée de deux observations uniquement. Cependant, nous avons testé plusieurs segmentations possibles pour l'intégration des descripteurs audio.

La première est au niveau du mot, afin de pouvoir concaténer cette représentation avec le vecteur du mot associé. Comme pour les études précédentes, les autres intégrations utilisent les pauses des locuteurs, et segmentent le tour de parole en UIP afin d'intégrer l'audio sur ces intervalles. Chaque mot de l'UIP se voit affecté des descripteurs audio intégrés sur son UIP de provenance. Nous utilisons différents paliers comme ce qui a été fait dans le chapitre 8. Comme pour les études précédentes, nous avons étudié une segmentation sur les pauses du locuteur et séparé les tours de parole en unités inter-pausales afin d'intégrer l'audio sur ces intervalles.

Nous avons aussi décidé d'utiliser d'autres descripteurs audio que l'ensemble que nous avons créé. Pour cela, nous avons entrainé un auto-encodeur sur les différentes paires adjacentes et tours de parole du corpus SEMAINE entier (selon le protocole de (Freitag et collab., 2017), décrit dans la sous-section 6.2.5). Nous avons utilisé ces représentations apprises comme représentation du signal audio.

Nous avons aussi choisi de nous comparer à un modèle à l'état de l'art pour l'analyse multimodale de sentiments, qui est le Tensor Fusion Network (TFN) de Zadeh et collab. (2017) (voir chapitre 7). Nous avons sélectionné une portion des descripteurs utilisés par notre modèle pour avoir les mêmes que Zadeh et collab. (2017). Ces descripteurs extraits via COVAREP sont : MDQ, PeakSlope, Rd, QOQ, NAQ, PSP, le ratio de contenu voisé sur non-voisé, ainsi que les 12 premiers MFCC et la fréquence fondamentale. Ils sont intégrés avec la moyenne uniquement, suivant le protocole des auteurs du TFN. Nous avons utilisé word2vec pour initialiser la couche d'embedding du RNN traitant le texte. Les auteurs originaux ont utilisé GloVe, qui est jugé équivalent à word2vec dans de nombreuses études. Nous choisissons word2vec dans le but de pouvoir comparer leur modèle avec les nôtres, et non pas comparer des descripteurs.

Finalement, nous avons aussi fusionné le texte et l'audio au moyen d'un réseau de neurones récurrent de type LSTM, avant d'utiliser les représentations apprises avec un HCRF. Pour le LSTM-HCRF, nous avons utilisé une approche en deux temps. Premiè-

rement, nous avons pris la configuration optimale du LSTM, d'après nos expériences précédentes. Ensuite, nous avons entraîné le LSTM seul sur le corpus pour trouver le label associé à l'AP. Finalement, nous avons entraîné un HCRF en couche finale de ce LSTM en utilisant la séquence d'états cachés que celui-ci produit. Pour ce deuxième entraînement, nous avons gelé les poids du LSTM. Nous avons testé les mêmes configurations de HCRF que pour les expériences précédentes.

10.1.2.2 Résultats

Nous avons fait plusieurs expériences concernant l'intégration des descripteurs. Le tableau 10.5 contient les résultats des expérimentations avec différentes segmentations et compare les résultats des systèmes utilisant différentes modalités. Le tableau 10.6 présente des résultats des expériences avec des modèles utilisant différentes représentations du signal audio : notre EDA, les représentations provenant d'un LSTM, et les représentations apprises avec un auto-encodeur sur la PA entière.

Concernant les résultats de la régression logistique, on a vu dans le chapitre 9 que les performances de ce modèle sur du texte étaient médiocres, atteignant des score F1 de 55 au maximum. L'ajout de l'audio permet d'augmenter faiblement ces résultats, mais en ne prenant uniquement que les descripteurs intégrés sur le tour de parole de l'utilisateur. Ceci est étrange, car quand on ajoute en concatenant ceux de l'agent les résultats baissent. On conclura par un écart non significatif dû à la faible taille de la base de données.

En utilisant le signal audio seul de manière unimodal ou bien mélangé au texte de manière multimodale, le HCRF obtient à chaque fois de meilleurs résultats que le LSTM. Cependant, il semble que le LSTM réussit à utiliser l'information contenue dans le signal vocal car l'ajout d'une nouvelle modalité permet au LSTM d'augmenter ses résultats par rapport au texte ou à l'audio seul, avec un F1 passant de 74 à 76. Ces résultats sont néanmoins toujours inférieurs à ceux du HCRF utilisant le texte ou le texte et l'audio qui ont des scores F1 de 80 et 77 respectivement.

Comme lors de l'étude de l'analyse l'opinion intra-locuteur, l'ajout du signal vocal diminue les performances des HCRF, le score F1 passant de 80 à 77. Finalement, on peut observer que l'intégration de la modalité audio sur des unités inter-pausales détériore systématiquement les performances du système, et ce pour les HCRF et LSTM (sauf un cas).

Dans le tableau 10.6, on peut voir que l'utilisation d'auto-encodeurs qui fonctionnent bien pour de l'analyse de scènes sonores ne donnent pas de bons résultats sur ce type de tâche. Nous avons aussi testé ces représentations au niveau des mots, sans résultats probants, confirmant notre intuition de la non efficacité de ce type d'apprentissage de

TABLE 10.5: Scores de F1 et d'*Accurracy* (taux de bonnes réponses) avec les différents modalités, modèles et segmentations

Modalité	Modèle	Segmentation	F1+	F1-	F1	Acc
∅	Positif	∅	79	0	40	66
Audio	RL	AP entière	68	45	56	59
		TP Utilisateur	72	48	60	64
		AP+TPs	68	47	57	60
	LSTM	mot	78	38	58	67
		IPU (150ms)	76	43	59	66
		IPU (300ms)	75	37	56	64
		IPU (500ms)	77	42	59	67
	HCRF-X	mot	79	54	66	71
	HCRF-X	IPU (150ms)	76	47	62	67
	HCRF-X	IPU (300ms)	73	43	58	64
	HCRF-X	IPU (500ms)	72	47	60	64
Texte	RL	AP entière	68	42	55	58
		TP Utilisateur	70	39	55	60
		AP+TPs	70	39	55	60
	LSTM	mot	83	64	74	77
	HCRF	mot	**86**	**75**	**80**	**82**
Audio-textuelle	RL	AP entière	73	41	57	63
		TP Utilisateur	70	50	60	63
		AP+TPs	66	47	57	58
	LSTM	mot	84	68	76	79
		IPU (150ms)	82	65	74	77
		IPU (300ms)	81	63	72	75
		IPU (500ms)	82	60	71	75
	HCRF-X	mot	85	69	77	80
	HCRF-X	IPU (150ms)	83	66	75	77
	HCRF-X	IPU (300ms)	85	62	74	78
	HCRF-X	IPU (500ms)	84	67	76	79

représentations pour notre tâche.

Il est intéressant de noter que l'utilisation d'un HCRF sur les états cachés d'un LSTM ayant appris la même tâche permet de faiblement augmenter les résultats par rapport au LSTM et au HCRF utilisant directement les descripteurs audio. Les résultats de l'implémentation du Tensor Fusion Network de Zadeh et collab. (2017) (utilisant uniquement le texte et l'audio) sont loin des autres modèles pour la tâche envisagée. La taille de la base de données doit être un frein pour ce modèle neuronal possédant de nombreux paramètres (plus de 900 000 paramètres à optimiser contre moins de 30 000 pour le plus gros modèle HCRF). De plus, ce modèle est construit pour des données

TABLE 10.6: Scores de F1 et d'*Accurracy* (taux de bonnes réponses) avec les différentes représentations, modèles et segmentations

Modèle	Texte	Audio	F1+	F1-	F1	Acc
Dummy	∅	∅	79	0	40	66
LR-early	w2v+ling	AE	75	39	57	64
LR-early	w2v+ling	LLD	70	50	60	63
LR-late	w2v+ling	LLD	74	49	61	66
LSTM	w2v+ling	LLD	84	68	76	79
TFN (Zadeh et collab. (2017))	w2v	LLD	79	49	65	70
HCRF-1	∅	LLD	79	54	66	71
HCRF-4	w2v+ling	∅	**86**	**75**	**80**	**82**
HCRF-4	w2v+ling	LLD	**85**	**69**	**77**	**80**
HCRF-1	LSTM		**85**	**70**	**78**	**80**

multimodales utilisant la vidéo dans un contexte non interactionnel.

10.1.2.3 Analyse

10

Une analyse appronfondie du modèle ne permet pas de faire ressortir de caractéristiques éclairant la compréhension des faiblesses du modèle multimodal. Encore une fois, on se heurte à une trop grande dimension de la représentation du signal audio. On retrouve les mêmes erreurs que dans la partie précédente, et il serait pénible de les décrire une nouvelle fois. Un détail remarquable est le non fonctionnement de l'algorithme de Viterbi pour améliorer les résultats, technique permettant d'augmenter les performances pour le texte.

10.2 FUSION BIMODALE PRÉCOCE AVEC SÉLECTION DE DESCRIPTEURS

Nous avons pu voir dans la partie précédente que l'audio a de moins bons résultats que le texte, et que l'ajout de l'audio au texte impacte les performances des modèles de manière négative. Nous savons aussi que le vecteur de représentation du signal vocal a une taille deux fois plus grande que celui du signal textuel.

Dans cette section on étudie une sélection de descripteurs pour diminuer l'impact de l'audio en ne gardant que ce qui est essentiel à une bonne classification et avoir ainsi des structures plus simples à repérer dans l'espace des observations. Cette diminution de dimension se fait dans le but de permettre aux HCRF d'être efficaces avec une détection de structure simple au niveau des observations (pour rappel le HCRF

caractérise la compatibilité entre une observation et un état caché par un produit scalaire, donc utilisation d'un simple hyperplan).

Dans cette section, nous allons présenter les résultats obtenus lorsque nos modèles HCRF utilisaient l'ensemble de descripteurs audio décrits dans la Section 2.2 après une étape de sélection de descripteurs. Nous avons effectué deux types de sélections de descripteurs différentes : une avec un *Gradient Boosting* sur des arbres de décision (sous-section 10.2.1) et l'autre avec une pénalisation de type *Elastic Net* appliquée à des HCRF (sous-section 10.2.2).

10.2.1 Sélection de descripteurs à l'aide d'un Gradient Tree Boosting sur des tours de parole : Étude sur l'analyse d'opinion inter-locuteurs

Dans cette partie, nous présentons l'étude de nos modèles avec une sélection de descripteurs pour une tâche d'analyse d'opinion inter-locuteurs sur le corpus *SEMAINE-Léger*. Pour sélectionner les descripteurs, nous avons utilisé une technique de *Gradient Boosting* (Friedman, 2001) sur des arbres de décision. Ces arbres de décisions ont été entraînés pour la classification d'attitude de l'utilisateur sur des paires adjacentes. Une fois cela fait, nous avons sélectionné une partie des descripteurs les plus importants de l'audio, puis entrainé des HCRF utilisant notre représentation du signal verbal classique et une représentation de dimension plus faible du signal vocal.

10.2.1.1 Protocole

Pour la sélection de descripteurs, nous avons utilisé un algorithme de *Gradient Boosting* (Friedman, 2001) avec des arbres de décision. Nous avons intégré les descripteurs audio au niveau des tours de parole de l'agent, des tours de parole de l'utilisateur et de la paire adjacente entière. Nous avons aussi testé un modèle prenant en entrée la concaténation des intégrations des descripteurs sur les deux tours de parole de la PA.

Sur chacun des tours d'une validation croisée ayant la même partition que celle utilisée pour l'entrainement des modèles futurs, nous avons choisi l'intégration permettant d'avoir les meilleurs résultats (agent, utilisateur, agent et utilisateur concaténé, et PA entière). Sur ce modèle, nous avons choisi des représentations ayant les valeurs les plus importantes. Ces représentations sélectionnées ont ensuite été utilisés pour l'entraînement des LSTM et des HCRF. De cette manière, la sélection n'a pas eu lieu en utilisant des fichiers se trouvant sur l'ensemble de test des LSTM et des HCRF.

Afin de sélectionner les hyperparamètres des arbres de décision, nous avons pris le modèle ayant les meilleurs scores sur la validation croisée. Les hyperparamètres du *Gradient Tree Boosting* portent sur la profondeur des arbres de décision dans $\{3;4;5\}$.

TABLE 10.7: Scores de F1 et d'*Accuracy* (taux de bonnes réponses) des modèles HCRF avec les descripteurs acoustiques sélectionnés avec un *Gradient Boosting*

Segmentation Audio	Modalité	Sélection	Modèle	F1+	F1-	F1	Acc
word	Audio	Tous	HCRF-X	79	54	66	71
		10	HCRF-X	73	39	56	62
		20	HCRF-X	77	37	57	67
		30	HCRF-X	78	40	59	67
	Audio-Textuelle	Tous	HCRF-X	85	69	77	80
		10	HCRF-X	86	70	78	81
		20	HCRF-X	86	70	78	81
		30	HCRF-X	86	71	79	81

Des informations complémentaires sur la méthode d'apprentissage du *Gradient Tree Boosting* sont disponibles en Annexe G.

Afin de comparer les modèles utilisant une sélection de paramètres aux modèles textuels purs, nous avons choisi de prendre les 10, 20 et 30 représentations les plus importantes lors du premier apprentissage amenant la sélection. Concernant le protocole expérimental pour l'entraînement et la validation des LSTM et HCRF, nous avons utilisé le même que pour les expériences précédentes (voir sous-section 8.1.1).

10.2.1.2 Résultats

Les résultats des expériences sont visibles dans le tableau 10.7, les améliorations proposées ne permettent toujours pas d'atteindre des résultats meilleurs que les modèles textuels.

Pour un système unimodal audio : l'utilisation de peu de caractéristiques ne permet pas d'obtenir des résultats meilleurs qu'un système n'utilisant pas de sélection de features. Les dégradations de performances se font surtout sur les documents négatifs ($F1$ négatif passant de 54 à 39).

Pour un système multimodal : la sélection de paramètres permet effectivement d'améliorer les résultats du système. Cependant, il n'y a aucune amélioration globale, dans le sens où les résultats se rapprochent de ceux des systèmes unimodaux utilisant le texte sans jamais atteindre les mêmes performances.

10.2.1.3 Analyse

On peut voir dans les tableau 10.8 les descripteurs audio qui ont été sélectionnés à l'issue du *gradient boosting*, classés en fonction du nombre de fois où ils ont été choi-

TABLE 10.8: Représentations sélectionnées suite au *gradient boosting*

Nombre de sélections	Représentation
10	ΔMFCC_10_rangetot
9	ΔF_0_skew
8	ΔNAQ_median
7	ΔMFCC_8_skew
7	MFCC_8_mean
7	MFCC_6_lregcoef
7	MFCC_9_rangetot
6	F3_median
6	ΔJitter_skew
6	ΔF3_mean
6	ΔF1_mean
5	HNR_range25-75
5	ΔpeakSlope_quant90
5	MFCC_5_median

sis dans les 30 descripteurs les plus importants. On trouve le Δ (dérivée) de l'asymétrie de la fréquence fondamentale, cette asymétrie est liée à une intonation montante et à une voix stressée (Sigmund, 2013). Ici, ce n'est pas l'asymétrie de la fréquence fondamentale mais l'asymétrie de son Δ, c'est donc qu'on observe une asymétrie dans la distribution des pentes de la F_0. Ce qui équivaut à un fort changement de pente positivement ou bien négativement à un moment de l'intervalle étudié. On retrouve dans ces paramètres sélectionnés l'asymétrie de la dérivée du Jitter qui représente la variation de fréquence fondamentale d'un signal. Ceci est cohérent avec ce qu'on a obtenu précédemment. Le coefficient MFCC8 apparait plusieurs fois, ce coefficient correspond, d'après l'échelle Mel, aux fréquences autour de 625 Hz.

Concernant l'état neutre, on peut voir que dans le tableau 10.9 les poids des descripteurs audio sont très faibles et ont par conséquent peu d'impact sur l'utilisation de cet état. Les caractéristiques qui sont utilisées ont été choisies par rapport à leur pouvoir de discrimination d'une attitude positive par rapport à une attitude négative, il n'est pas surprenant qu'elles soient peu représentatives de l'état neutre.

Les résultats des modèles utilisant l'audio sont toujours moins bons que ceux qui utilisent uniquement la modalité textuelle. Même en ajoutant une représentation du signal vocal de faible dimension, les performances n'augmentent pas. La modalité acoustique n'apporte pas d'information complémentaire, ou du moins pas assez pour permettre d'obtenir une prédiction juste sur les fichiers qui sont problématiques pour le texte et ainsi augmenter les performances globales.

Le *Gradient Boosting* est effectué avec des arbres de décision sur des représen-

TABLE 10.9: Poids des représentations audio sélectionnées avec un *gradient boosting* pour 3 états différents qui au nouvel apprentissage ($\times 10^3$ pour plus de lisibilité)

(a) État **négatif**		(b) État **neutre**		(c) État **positif**	
Représentation	Poids	Représentation	Poids	Représentation	Poids
ΔQOQ_quant25	21,39	jitter_skew	2,50	Rd_quant25	30,74
ΔMFCC10_range10-90	15,46	ΔMFCC8_lregbias	1,49	shimmer_quant25	22,27
ΔMFCC8_lregbias	15,30	ΔpeakSlope_quant50	1,26	ΔLoudness_range10-90	15,66
HNR_quant90	11,25	Rd_quant25	1,26	ΔF2_std	11,17
jitter_skew	9,84	shimmer_quant25	1,25	MFCC7_std	9,69
ΔMDQ_range10-90	8,12	ΔF2_std	0,81	ΔF1_lregbias	4,09
ΔHNR_quant25	7,28	MFCC7_range25-75	0,77	ΔpeakSlope_quant50	4,08
MFCC9_range10-90	6,71	ΔQOQ_quant25	0,63	MFCC5_quant10	3,30
peakSlope_quant25	3,69	shimmer_quant90	0,51	ΔMFCC8_range25-75	2,13
shimmer_quant90	1,25	ΔF1_lregbias	0,35	shimmer_skew	0,00
shimmer_skew	-0,00	shimmer_skew	-0,00	peakSlope_quant25	-0,10
MFCC7_range25-75	-0,51	MFCC7_std	-0,10	MFCC7_range25-75	-0,26
ΔMFCC8_range25-75	-1,11	ΔMFCC10_range10-90	-0,33	shimmer_quant90	-1,76
MFCC5_quant10	-1,67	ΔHNR_quant25	-0,63	MFCC9_range10-90	-5,30
ΔF1_lregbias	-4,44	ΔLoudness_range10-90	-0,98	ΔHNR_quant25	-6,65
ΔpeakSlope_quant50	-5,34	ΔMFCC8_range25-75	-1,02	ΔMDQ_range10-90	-6,88
MFCC7_std	-9,59	ΔMDQ_range10-90	-1,24	HNR_quant90	-7,81
ΔF2_std	-11,98	MFCC9_range10-90	-1,42	jitter_skew	-12,35
ΔLoudness_range10-90	-14,68	MFCC5_quant10	-1,63	ΔMFCC10_range10-90	-15,13
shimmer_quant25	-23,52	HNR_quant90	-3,44	ΔMFCC8_lregbias	-16,79
Rd_quant25	-32,00	peakSlope_quant25	-3,59	ΔQOQ_quant25	-22,01

tations obtenues avec des descripteurs intégrés au niveau des tours de parole car le modèle n'est pas séquentiel. Or, le HCRF utilise des représentations provenant de descripteurs intégrés au niveau du mot. Ainsi, les représentations sélectionnées par le modèle de sélection de caractéristiques n'ont pas forcément le même comportement lors d'une intégration sur un intervalle de temporalité différente.

L'utilisation d'un RNN pour fusionner les descripteurs permet d'augmenter les résultats (Tableau 10.6), sans pour autant dépasser les modèles textuels. Nous avons remarqués que l'utilisation d'un LSTM simple comme c'est est le cas dans nos expériences n'est pas un choix pertinent. En effet, la couche cachée du LSTM se met à jour en fonction des nouvelles informations contenues dans les observations, et lorsque la séquence est finie, elle est utilisée pour prédire le label de la séquence entière. Le vecteur caché du LSTM contient de plus en plus d'information au fur et à mesure que le modèle avance le traitement de la séquence à analyser. Toutes les informations nécessaires à la tâche sont donc contenues dans le dernier vecteur caché du LSTM. L'utilisation de représentations n'est pas, de ce fait, un pré-traitement pertinent pour l'utilisation d'un HCRF. Une solution serait d'entrainer conjointement le RNN et le HCRF, ou d'utiliser un modèle bidirectionnel avec attention afin que chaque vecteur soit important.

10.2.2 Sélection de descripteurs à l'aide d'une régularisation Elastic net : Étude sur l'analyse d'opinion intra-locuteur

Dans cette étude, nous avons voulu étudier une sélection de descripteurs différente de la dernière méthode de type *Boosting*. En effet, nous avons pu noter des incohérences de protocole lors de l'usage de la technique de sélection de descripteurs utilisée dans l'étude précédente (voir Sous-sous-section 10.2.1.3). En effet, dans l'étude précédente les descripteurs audio ont été sélectionnés par rapport aux valeurs de leurs fonctionnelles sur des intervalles longs tels que des tours de parole alors qu'ils sont utilisés différemment par notre classifieur. Les classifieurs utilisaient les descripteurs intégrés sur des mots ou des UIP, le temps caractéristique étant différent, nous avons jugé que la méthode n'était pas la plus adaptée. Dans cette étude, nous avons choisi une approche où la segmentation est la même pour à la fois le modèle de sélection de paramètres et le classifieur. Ceci permet d'éviter de mettre de côté des caractéristiques qui sont importantes localement. Pour effectuer cette sélection, nous proposons un protocole de double validation croisée décrit en 10.2.2.1. Les résultats obtenus sont toujours moins bons que des systèmes utilisant uniquement le texte. Nous finissons par une analyse du système afin de voir les caractéristiques sélectionnées et quels sont leurs impacts sur chacun des états.

Résultat : Score d'un modèle appris avec un ensemble restreint de features
Création d'une partition balancée des documents de la base de données $\mathscr{P} = \{Test_i, i = 1..10\}$ (même partition que utilisée précédemment);
pour $i = 1...\ 10$ **faire**
  Séparation de la base de données en deux ensembles $Test_i$ et $Train_i = \mathscr{T} \setminus Test_i$;
  Création d'une partition $\mathscr{P}_i$ des documents de $Train_i$ telle que $\mathscr{P}_i = \{Test_{i,j}, j = 1..10\}$;
  pour $j = 1...\ 10$ **faire**
    Entraînement d'un HCRF ayant une régularisation de type *Elastic net* sur $Train_i \setminus Test_{i,j}$;
    Sélection des $FS_{i,j}$ descripteurs ayant des poids non nuls θ_o;
  fin
  $k_i = 10$;
  tant que $\left\{descripteur / \sum_j \mathbb{1}_{descripteur \in FS_{i,j}} \geq k_i\right\} \neq \varnothing$ **faire**
    $k_i = k_i - 1$;
  fin
  Sélection des $FS_i = \left\{\text{descripteur} / \sum_j \mathbb{1}_{\text{descripteur} \in FS_{i,j}} \geq k_i\right\}$ descripteurs étant non nuls sur le maximum k_i des 10 tours de j;
  Entraînement d'un HCRF sur $Train_i$ avec les descripteurs audio FS_i;
  Test de prédiction sur le document $Test_i$;
fin

Algorithme 0 : Algorithme de choix des descripteurs

10.2.2.1 Protocole

Pour faire une sélection de descripteurs à l'aide d'un modèle observant les descripteurs intégrés au niveau des mots, nous avons décidé d'utiliser des modèles HCRF avec des régularisations ℓ-1 et ℓ-2 combinées. Cette double régularisation se nomme régularisation de type *Elastic net*.

Afin d'obtenir un ensemble de descripteur réduit qui soit appris sur chaque tour de la validation croisée, nous avons effectué le protocole ci-dessous avec double validation croisée à 10 tours :

k_i est le nombre maximum tel que l'ensemble des descripteurs ayant un poids θ_o non nuls sur k_i tours ne soit pas vide, i.e. k_i est compris entre 1 et 10 et c'est le nombre maximum tel que $\{\text{descripteur} / \sum_j \mathbb{1}_{\text{descripteur} \in FS_{i,j}} \geq k_i\} \neq \varnothing$. Dit d'une autre manière, k_i est la valeur maximum telle que $\bigcup_{Ens \in B_{k_i}} \bigcap_{j \in Ens} F_{i,j} \neq \varnothing$ où B_{k_i} est l'ensemble des ensembles de 1 à 10 avec tirage simultané de k_i valeurs.

Ce protocole permet de sélectionner un ensemble de descripteurs d'une façon stable, en ne prenant que les plus importants. L'utilisation de la validation croisée à 10 tours permet de nous comparer avec les résultats des études précédentes car la partition de validation croisée à 10 tours reste la même.

Les expériences ont été faites uniquement avec une segmentation utilisant les pauses du locuteur de plus de 300 ms car c'est avec cette segmentation que l'on obtient les meilleurs résultats, à la fois pour le texte et pour l'audio.

10.2.2.2 Résultats

Les résultats des modèles utilisant les paramètres sélectionnés se trouvent dans le Tableau 10.11. Comme précédemment, on peut observer une baisse des performances par rapport au modèle textuel dès qu'on ajoute des descripteurs audio.

Pour un système unimodal audio : la sélection de caractéristiques fait chuter les performances. Encore une fois ces baisses de performances sont centrées sur les documents négatifs ($F1$ négatif passant de 62 à 55). Cependant, les performances augmentent sur les documents positifs ($F1$ positif passant de 79 à 82).

Pour un système multimodal : on observe les mêmes résultats que pour la partie précédente. La diminution des paramètres audio permet de faire rapprocher les résultats des modèles multimodaux de ceux des modèles textuels mais ne permet pas de les dépasser.

TABLE 10.11: Scores de F1 et d'*Accuracy* (taux de bonnes réponses) des modèles HCRF avec les descripteurs acoustiques sélectionnés avec un *Elastic Net*

Modalité	Segmentation du signal	Sélection	F1+	F1-	F1	Acc
Audio	IPU (300ms)	∅	79	62	74	73
		✓	82	55	73	74
Audio-Textuelle	IPU (300ms)	∅	82	65	76	76
		✓	83	68	78	78

10.2.2.3 Analyse

Pour chaque tour i de la validation croisée initiale, les valeurs du palier du nombre de tours k_i minimum contenant au moins une fois chaque caractéristique pour qu'elle soit sélectionnée (voir algorithme 0) sont reportés dans le tableau 10.12. A part les tours 5 et 6 où les nombres de paramètres sélectionnés sont de 26 et 74, on compte environ une soixantaine de paramètres sélectionnés.

TABLE 10.12: Palier k_i et nombre de représentations sélectionnées par tour *Elastic Net*

Tour i	Palier k_i	Nombre de paramètres
1	7	62
2	9	64
3	9	67
4	9	59
5	7	26
6	6	74
7	7	64
8	7	62
9	8	66
10	7	63

Dans le tableau 10.13, sont visibles les paramètres qui ont été sélectionnés sur chacun des tours de la validation croisée. Concernant les représentations cepstrales et prosodiques, on peut voir qu'il y a beaucoup de valeurs provenant de MFCC, et la moyenne de l'énergie (MFCC_0). Concernant la qualité de la voix, on trouve le coefficient de relaxation R_d qui est caractéristique d'une voix tendue ou relaxée et des valeurs représentant la variabilité du quotient quasi-ouvert, utile pour discriminer les voix tendues ou soufflées, donc des caractéristiques à valeurs élevées lorsqu'une voix passe de tendu à soufflé (ou l'inverse) au sein d'une même UIP.

Finalement, dans les tableaux 10.14, 10.16 et 10.18, le lecteur peut trouver les poids des caractéristiques les plus compatibles, incompatibles ou non corrélées avec les états négatif, neutre et positif.

Il faut garder en tête que les caractéristiques ont subi une normalisation préalable et qu'elles ont des valeurs positives ou négatives. Ainsi, un poids θ_o est à interpréter en fonction de cet aspect. Prenons deux exemples : les percentiles 75 et 90 du quotient-quasi-ouvert. Les poids d'observations de ces paramètres sont visibles dans le tableau 10.20.

Si les quantiles 75 et 90 sont au dessus de la moyenne, alors ces valeurs ne contribueront pas à aller dans l'état neutre. Cependant si ces valeurs sont en dessous de la moyenne du corpus, alors ces valeurs contribueront à aller dans l'état neutre. On a résumé ce qui se passe dans le tableau 10.21.

Un des inconvénients de notre système est son incapacité à prendre en compte des valeurs étant dans la moyenne et de supposer qu'une valeur basse sur une observation équivaut à l'opposé d'une valeur haute. Une solution pour contrer cette effet serait de représenter les caractéristiques audio de la même manière que nous avons intégré les valeurs provenant des lexiques de subjectivité. Ceci permettrait d'obtenir 3 valeurs distinctes caractéristiques du comportement de la variable et de rendre les

TABLE 10.13: Représentations sélectionnées suite à la sélection basé sur un *Elastic Net*

Nombre de sélections	Représentation
10	MFCC_10_quant10
10	ΔQOQ_std
10	MFCC_10_quant25
10	MFCC_0_quant75
10	MFCC_0_mean
10	MFCC_4_mean
10	MFCC_4_quant10
10	MFCC_4_quant25
10	MFCC_4_quant50
10	QOQ_quant90
10	MFCC_4_quant75
10	QOQ_range10-90
10	MFCC_4_quant90
10	QOQ_std
10	Rd_quant25
10	MFCC_6_quant10
10	Rd_mean
10	MFCC_6_quant25
10	ΔQOQ_quant10
10	MFCC_6_std

TABLE 10.14: Poids des représentations audio sélectionnées avec un *Elastic Net* après un nouvel apprentissage multimodal pour l'état **négatif** ($\times 10^3$ pour plus de lisibilité)

(a) Représentations incompatibles (θ_o très négatif)

Représentation	Poids
PSP_quant10	-44,04
H1H2_quant10	-41,72
ΔpeakSlope_quant10	-38,38
MFCC4_quant10	-37,99
MFCC4_quant25	-37,86
MFCC3_quant75	-37,04
HNR_mean	-37,00
MDQ_quant90	-34,10
MFCC7_quant90	-33,78
MDQ_quant10	-31,90
MFCC3_quant90	-29,47
MFCC5_quant90	-22,74
ΔRd_quant25	-18,83
MFCC3_quant50	-18,57
HNR_quant25	-18,36
ΔRd_quant75	-18,02

(b) Représentations ayant peu d'impact, ($|\theta_o|$ faibles)

Représentation	Poids
MFCC4_quant75	7,76
MFCC3_quant25	7,46
HNR_std	6,21
Δshimmer_lregcoef	6,20
MFCC9_quant50	5,94
ΔRd_quant10	5,91
MFCC5_mean	3,39
ΔPSP_quant75	1,98
Δshimmer_quant25	1,72
MFCC12_range10-90	1,11
ΔRd_range25-75	-1,06
ΔpeakSlope_range25-75	-1,20
ΔPSP_range25-75	-1,61
HNR_lregcoef	-3,87
MDQ_quant75	-5,07
HNR_quant75	-5,96

(c) Représentations compatibles, (θ_o très positifs)

Représentation	Poids
Δshimmer_quant50	46,30
ΔpeakSlope_quant90	42,14
ΔpeakSlope_range10-90	41,54
QOQ_quant90	40,93
Δshimmer_quant90	34,39
PSP_std	33,22
Δshimmer_skew	32,97
H1H2_quant25	29,54
Δshimmer_quant75	28,40
MFCC12_std	28,17
MFCC12_range25-75	28,14
MFCC5_quant50	27,12
HNR_quant10	24,68
ΔpeakSlope_std	20,12
QOQ_quant75	18,78
HNR_quant50	18,63

TABLE 10.16: Poids des représentations audio sélectionnées avec un *Elastic Net* après un nouvel apprentissage multimodal pour l'état **neutre** ($\times 10^3$ pour plus de lisibilité)

(a) Représentations incompatibles (θ_o très négatifs)

Représentation	Poids
Δshimmer_quant90	-52,47
Δshimmer_skew	-49,64
ΔpeakSlope_std	-46,65
ΔpeakSlope_range10-90	-46,31
ΔpeakSlope_quant90	-44,16
Δshimmer_quant75	-42,11
QOQ_quant75	-36,14
QOQ_quant90	-34,60
Δshimmer_quant50	-30,31
Δshimmer_lregcoef	-27,54
ΔpeakSlope_range25-75	-23,51
PSP_std	-22,41
ΔpeakSlope_quant75	-20,82
ΔRd_quant75	-19,21
HNR_quant25	-18,87
Δshimmer_quant25	-18,48

(b) Représentations ayant peu d'impact ($|\theta_o|$ faibles)

Représentation	Poids
MFCC7_quant75	7,73
MFCC3_quant90	6,44
ΔRd_quant25	6,33
H1H2_quant25	6,27
ΔPSP_quant75	5,87
MFCC12_range25-75	5,78
MFCC7_mean	3,34
HNR_quant10	0,03
ΔPSP_range25-75	-1,13
MFCC5_quant90	-2,78
HNR_quant50	-2,86
MFCC7_quant90	-3,44
H1H2_quant10	-4,11
MDQ_quant90	-4,24
MFCC5_quant75	-4,75
MFCC9_quant50	-6,01

(c) Représentations compatibles (θ_o très positifs)

Représentation	Poids
ΔpeakSlope_quant10	46,04
PSP_quant10	37,37
PSP_quant25	35,20
MDQ_quant50	28,72
MFCC4_quant10	28,14
MFCC4_quant50	27,98
MFCC4_mean	25,19
MFCC4_quant25	24,29
MFCC3_mean	22,63
MFCC4_quant75	22,47
MDQ_quant25	22,01
MFCC3_quant50	20,80
MDQ_quant10	19,16
MDQ_mean	17,49
MFCC12_range10-90	16,68
ΔRd_quant10	14,67

TABLE 10.18: Poids des représentations audio sélectionnées avec un *Elastic Net* après un nouvel apprentissage multimodal pour l'état **positif** ($\times 10^3$ pour plus de lisibilité)

(a) Représentations incompatibles (θ_o très négatif)

Représentation	Poids
MDQ_quant50	-42,57
MFCC12_std	-41,99
H1H2_quant25	-35,76
MFCC12_range25-75	-33,92
MFCC3_mean	-32,60
MFCC4_quant75	-30,23
HNR_quant10	-24,73
MFCC7_quant75	-22,84
ΔRd_quant10	-20,52
HNR_std	-20,31
PSP_quant25	-17,88
MFCC12_range10-90	-17,79
MFCC3_quant25	-17,19
MFCC5_quant50	-16,26
Δshimmer_quant50	-15,99
HNR_quant50	-15,75

(b) Représentations ayant peu d'impact ($|\theta_o|$ faibles)

Représentation	Poids
HNR_lregcoef	11,62
MFCC4_quant10	9,79
PSP_quant10	6,67
MFCC5_mean	5,66
ΔpeakSlope_range10-90	4,77
ΔPSP_range25-75	2,71
ΔpeakSlope_quant90	2,04
MFCC9_quant50	0,01
MDQ_mean	-0,05
MFCC3_quant50	-2,18
MDQ_quant75	-3,53
HNR_quant75	-4,32
QOQ_quant90	-6,30
MFCC5_quant75	-6,95
ΔpeakSlope_quant10	-7,70
ΔPSP_quant75	-7,82

(c) Représentations compatibles (θ_o très positif)

Représentation	Poids
HNR_mean	52,18
H1H2_quant10	45,82
MDQ_quant90	38,33
HNR_quant25	37,25
ΔRd_quant75	37,20
MFCC7_quant90	37,14
ΔpeakSlope_quant75	32,95
ΔpeakSlope_std	26,55
MFCC5_quant90	25,57
ΔpeakSlope_range25-75	24,70
ΔRd_quant90	23,29
MFCC3_quant90	23,06
MFCC3_quant75	22,54
ΔRd_range10-90	22,21
Δshimmer_lregcoef	21,36
Δshimmer_quant90	18,07

TABLE 10.20: Poids d'observation θ_o

Représentation	État **négatif**	État **neutre**	État **positif**
QOQ_quant_75	40, 93	-36,41	-6,30
QOQ_quant_90	18, 78	-34,60	17,39
MDQ_median	13, 86	28,70	-42,57

TABLE 10.21: Impact des percentiles 75 et 90 du quotient-quasi-ouvert sur les différents états

Représentation	Valeur par rapport à la moyenne	États compatibles
QOQ_quant_75	au-dessous	**neutre**
	égale	∅
	au-dessus	**positif**/**négatif**
QOQ_quant_90	au-dessous	**neutre**
	égale	∅
	au-dessus	**négatif**
MDQ_median	au-dessous	**positif**
	égale	∅
	au-dessus	**neutre**

poids indépendants.

Finalement, étant donné que nos systèmes forment toujours les mêmes configurations d'états, il serait possible de faire un post-traitement sur les états cachés des systèmes unimodaux afin de trouver une configuration optimale pour modéliser la chaine des états cachés. Cependant, ceci reviendrait à utiliser un MV-HCRF (Song et collab., 2012b,a) avec des entrainements séparés.

CONCLUSION

Contrairement à nos attentes, les architectures testées pour intégrer l'audio ne permettent pas de montrer l'apport de cette modalité dans notre système, les résultats obtenus restent toujours médiocres. Nous avons premièrement tenté une approche de force brute qui n'a eu aucun de succès. Les résultats de l'audio étant moins bons que ceux du texte et et le nombre de paramètres liés à l'audio étant bien plus important pour le texte, nous avons tenté de sélectionner les paramètres de l'audio pour en limiter l'impact. Deux types de sélection de paramètres ont été étudiées. Nous avons essayé une sélection avec un *gradient boosting* sur des arbres de décision et une sélection avec une pénalisation de type *elastic net* sur un HCRF. Les deux approches n'ont pas permis d'améliorer les résultats. Il semble que les HCRF ne soient pas des modèles efficaces pour ce type de descripteurs intégrés par des fonctionnelles statistiques sur des intervalles courts.

11

Conclusion et perspectives

11.1 APPORT DE NOTRE TRAVAIL

Les apports de cette thèse se situent sur plusieurs points, que l'on a pu segmenter en différentes questions de recherche. Pour répondre à ces questions, nous avons fourni des éléments de réponse qui sont synthétisés dans les différentes sous-sections ci-dessous.

Adaptation du domaine écrit au domaine oral de mots-vecteurs extérieurement appris

Cette partie du travail est liée à la recherche d'un ensemble de descripteurs adapté pour l'analyse des opinions dans la voix. Nous avons utilisé, en particulier pour le texte, des représentations distribuées apprises préalablement sur du texte écrit que nous avons utilisées sur des transcriptions de parole orale. Les représentations de mots distribuées permettent de représenter les mots par la sémantique qu'ils transportent, mais nécessitent de grandes quantités de données.

L'écrit et l'oral sont deux domaines différents du langage : on ne s'exprime pas de la même manière à l'oral et à l'écrit, et ainsi le signifié des mots change. L'utilisation des HCRF permet au modèle de s'affranchir de ces différences en détectant les représentations des mots qui ne sont pas adaptés pour diminuer leur impact sur la décision finale. Par exemple : 'hum', 'aah' , 'eeh', 'yeah', 'know'.

Nous avons aussi vu que les HCRF permettaient une deuxième adaptation de domaine en réduisant l'impact des mots spécifiques au type de discours. Ainsi, les mots liés au type de données (critique de film) qui n'avaient pas une représentation vectorielle cohérente avec le signifié du locuteur étaient eux aussi écartés. Par exemple : 'Justin', 'Mickael', 'Thanks'.

- *Q.R.1 => Quels sont les descripteurs adaptés?* : Nous avons montré que des représentations distribuées apprises sur un domaine contenant beaucoup de données disponibles comme le texte, pouvaient être utilisées pour modéliser des données d'un autre domaine comme l'oral.
- *Q.R.3 => Quel classifieur utiliser?* : Les états cachés des HCRF permettent cette adaptation de domaine en utilisant des états comme "poubelle" et ainsi éviter que ces représentations inappropriées aient un impact sur la décision finale.

Étude d'une segmentation automatique en pauses

L'étude de la segmentation est directement une des problématique soulevées en début de manuscrit. La segmentation est une étape cruciale car elle précède l'intégration des descripteurs qui a besoin de se faire sur des unités de temporalité cohérente

pour obtenir des représentations pertinentes. Nous avons choisi d'étudier une segmentation en fonction des pauses du locuteur car ces auto-interruptions permettent de segmenter automatiquement le discours en unités cohérentes à la compréhension des phénomènes vocaux et verbaux pour la reconnaissance des opinions.

- *Q.R.2 => Quelle est la segmentation adaptée?* Une segmentation en fonction des pauses du locuteur permet dans certains cas d'améliorer les résultats en modélisant de manière pertinente des unités consistantes et faciles à extraire automatiquement. Cette segmentation est adaptée à la nature de la parole orale où la segmentation en phrase est difficile à mettre en œuvre.

Modélisation des dynamiques intra et inter-locuteurs

On a vu que les phénomènes d'opinion peuvent se réaliser de différentes manières, selon le contexte qui est à prendre en compte. Nous avons étudié l'opinion dans de longs monologues, modélisant le discours avec des états cachés, et dans des interactions de parole en utilisant le contexte d'une paire adjacente.

- *Q.R.1 => Quelles sont les représentations adaptées?* : Nous avons étudié l'utilisation de descripteurs interactionnels pour modéliser le contexte de la paire adjacente.
- *Q.R.2 => Quelle est la segmentation adaptée?* : Nous avons étudié plusieurs segmentations au niveau de la paire adjacente afin de mieux prendre en compte le contexte interactionnel.
- *Q.R.3 => Quel classifieur utiliser?* : Les états cachés des HCRF permettent de modéliser et de représenter la dynamique des opinions d'un locuteur à l'intérieur d'un tour de parole ou d'un discours, en imposant au système une structure entre les différents états cachés.

Annotation en opinion d'un corpus d'interactions humaines

Durant cette thèse, nous avons fait annoter 79 discussions de la base de données SEMAINE en opinion en tour de parole. De plus, nous avons aligné l'audio et le texte sur ces 79 sessions. Ce corpus va être mis à la disposition de la communauté scientifique. Nous prévoyons de l'ajouter à l'API MultimodalSDK [1] pour une utilisation simple de la part des utilisateurs.

1. https ://github.com/A2Zadeh/CMU-MultimodalSDK

Étude de l'utilisation jointe de l'audio et du texte pour de l'analyse d'opinion dans des interactions humaines orales

Finalement, nous avons étudié des modèles multimodaux utilisant signal verbal et vocal à la fois. Pour utiliser au maximum les bonnes performances de la modalité textuelle et augmenter quand même les résultats avec la modalité audio, nous avons décidé de diminuer l'impact de l'audio avec une sélection de descripteurs.

- *Q.R.1 => Quels sont les représentations adaptées ?* : Nous avons étudié l'utilisation de représentations acoustiques avec ou sans sélection de descripteurs afin de diminuer la dimension de l'espace des descripteurs.

11.2 PERSPECTIVES DE RECHERCHE

11.2.1 Les perspectives à court-terme

Apprentissage de représentations acoustiques externes

Un des aspects que nous avons remarqués par rapport aux représentations du signal audio est que ces représentations étaient trop brutes pour un modèle relativement simple comme un HCRF. Une solution appropriée pourrait être d'avoir une représentation du signal vocal de plus haut niveau. Ce type de représentation peut être obtenue avec l'utilisation d'un modèle d'apprentissage profond.

Nous avons essayé une approche non supervisée avec des auto-encodeurs appris sur le corpus SEMAINE entier selon la méthode de Freitag et collab. (2017). Cependant, ce type de stratégie relève surtout de la compression et n'a pas permis d'obtenir de meilleurs résultats. A contrario, l'utilisation d'un modèle plus complexe appris sur un corpus extérieur de vidéo comme le modèle profond de Abu-El-Haija et collab. (2016) appris sur le corpus Youtube-8M composé de plus de 500 000 heures de vidéos pourrait être une solution. Ce modèle permet, à la manière de word2vec, d'obtenir des représentations à partir d'un apprentissage fait sur un corpus extérieur de manière non supervisée.

Utilisation d'une modélisation plus complexe du langage

Dans les travaux présentés dans ce manuscrit nous avons utilisé des mots-vecteurs provenant de Mikolov et collab. (2013a). Depuis, de nombreuses avancées ont eu lieu en traitement du langage naturel et pour la création de représentation textuelle distribuée.

Une amélioration simple au modèle que l'on a pu proposer serait d'utiliser des mots vecteurs plus sophistiqués comme ELMo de Peters et collab. (2018b). Cette représentation permet d'obtenir des mots vecteurs à l'aide de la phrase entière, permettant de construire un vecteur du mot dépendant du contexte de son emploi.

Finalement, pour les modèles comme HCRF-3 qui ont besoin d'une représentation d'un tour de parole entier, nous avons pris la moyenne des mots-vecteurs sur le tour de parole du locuteur. Une amélioration serait d'utiliser un modèle permettant d'obtenir des représentations de chaines de caractères de tailles différentes (mots ou phrases), comme le BERT de Kenton et collab. (2018) ou le Google Universal Sentence Encoder de Cer et collab. (2018).

Multi-tâche sur les différents phénomènes affectifs

Le corpus SEMAINE comporte d'autres annotations liées à des phénomènes affectifs qu'il serait intéressant d'utiliser. Nous pensons notamment aux annotations continues en valence et en activations faites pour le challenge AVEC2011 (Schuller et collab., 2011b). L'utilisation de ces informations permettrait de construire un modèle multi-tâche utilisant les liens existants entre différentes tâches connexes pour apprendre de manière plus efficace en globalité (Collobert et Weston, 2008; Chen et collab., 2018). Il est même possible de prendre en compte la hiérarchisation des différents phénomènes étudiés dans le modèle, afin d'utiliser un modèle hiérarchique permettant de prédire d'abord les phénomènes bas-niveau (comme la valence), puis les phénomènes haut-niveau (comme les opinions), de la même manière que le HMTL de Sanh et collab. (2018).

11.2.2 Les perspectives à long-terme

Utilisation conjointe du modèle graphique avec un ANN

Nous avons vu dans l'étude de l'opinion dans le texte que les HCRF ont un pouvoir de regroupement (*clusterisation*) des observations à l'aide des états cachés. De plus, cette clusterisation permet de choisir quelle observation est importante ou non pour la décision finale. Ceci permet à une couche de HCRF d'agir comme une couche d'attention spéciale qui peut, en plus de détecter ce qui est important, structurer et modéliser la séquentialité des données. L'analyse de la séquence de clusters permettrait ensuite de labelliser finement aux niveau des observations à la manière des travaux de Angelidis et Lapata (2017).

Les HCRF que nous avons entrainés dans cette thèse ont été optimisés avec un L-BFGS, mais il est aussi possible d'utiliser une autre méthode comme une descente du

gradient stochastique et de les entraîner de manière conjointe avec un réseau de neurones. Ceci pourrait être un moyen d'effectuer une fusion des modalités avant l'entrée de la couche HCRF. Nous avons d'ailleurs observé une amélioration des performances des systèmes multimodaux sur le corpus SEMAINE-Léger, avec l'utilisation de représentations apprises par un LSTM.

Détection statistique de la cible de l'opinion

Une annotation plus précise du corpus permettrait de faire un système repérant les cibles associées aux opinions. Une des multiples versions codées de la plateforme d'annotation mise en place pour l'annotation du corpus SEMAINE permet de récupérer ce type d'information précise à la granularité du mot. Ceci permettrait d'utiliser un système de type ABSA (*Aspect-Based Sentiment Analysis*) (Liu, 2012; Brychcin et collab., 2014) sur de la parole orale et dans un contexte interactionnel.

Prise en compte d'un plus grand contexte dialogique

Une des limites de ce travail est la prise en compte d'un contexte dialogique restreint à une paire de tours de paroles. Cependant, dans une interaction entre deux individus le contexte général est important et étendre le contexte dialogique à la conversation entière est une amélioration qui serait notable pour notre système. Cependant au vu de la nature des HCRF, il faudrait un classifieur plus complexe ou du moins hiérarchique comme le modèle a état latent de Song et collab. (2013).

A

Plateforme d'annotation php

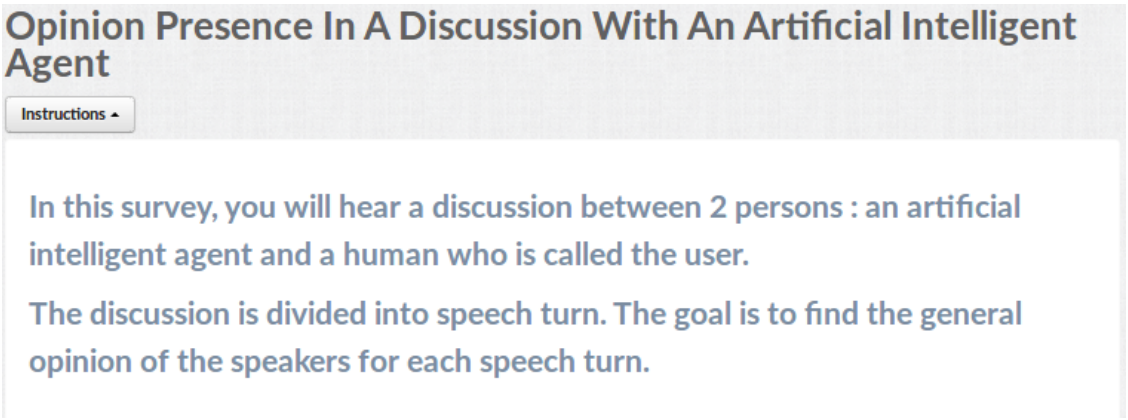


FIGURE A.1: Page de présentation de la plate-forme d'annotation, premier contact des annotateurs avec la tache (1/4)

Opinion :

For each speech turn you will have to annotate :

- the presence of at least one opinion,

- the type of the opinions (from very negative to very positive),

- if the speaker expressed several opinions with opposite types within a speech turn (e.g. positive and negative) within a speech turn, you will have to annotate the overall opinion type of the speech turn (the opinion type which is overcoming in the speech turn).

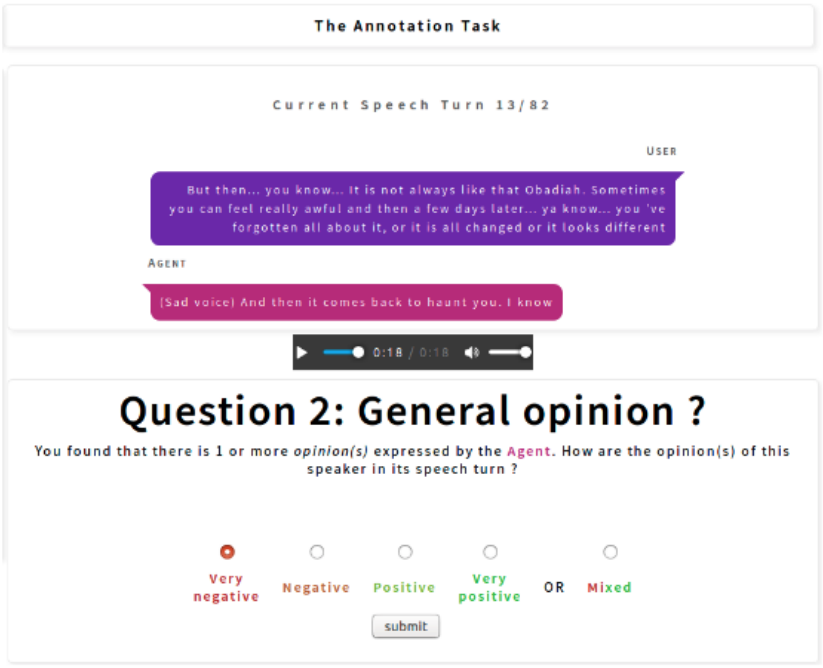


FIGURE A.2: Page de présentation de la plate-forme d'annotation, premier contact des annotateurs avec la tache (2/4)

Time : The annotation task lasts around 30mn (counting the preparation). Some discussions can contain more opinions than others. The discussion can last from 2mn to 10mn of speech.

FIGURE A.3: Page de présentation de la plate-forme d'annotation, premier contact des annotateurs avec la tache (3/4)

Payment : Each annotation of a discussion **will be paid regarding the size of the discussion to annotate**. For a session of 5mn speech to annotate, it takes around 30mn with the preparation and you'll get paid around 1.80$.

Hence, you will have to provide your Crowdflower ID in order to get paid regarding the duration of the session. We will check that the annotation task has been run seriously.

FIGURE A.4: Page de présentation de la plate-forme d'annotation, premier contact des annotateurs avec la tache (4/4)

INSTRUCTIONS

The main goal of this study is to label several sentences uttered by a person during an interaction with a virtual agent.

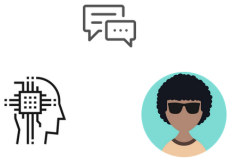


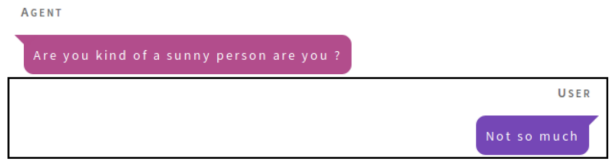
FIGURE A.5: Instructions de la plate-forme d'annotation (1/6)

Instructions

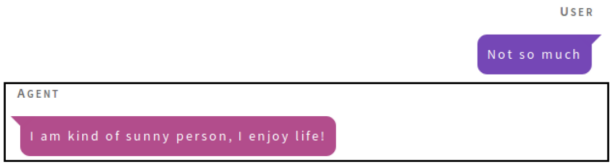
In this study, you will have to detect opinions expressed by the speakers. The conversation that you will annotate is between a virtual agent and a user.

Utterance Pairs

The conversation is displayed by pairs of speech turns (one speaker speaking and the other one answering) called *Utterance Pairs*, like the one you can see below :



Except at the beginning, You will always have 2 speech turns in front of you, and the window will slide one speech turn per speech turn.



You will use the simple context of an Utterance Pair in order to annotate opinions of the 2nd speaker. The annotation rules will be described later.

Previous Next

FIGURE A.6: Instructions de la plate-forme d'annotation (2/6)

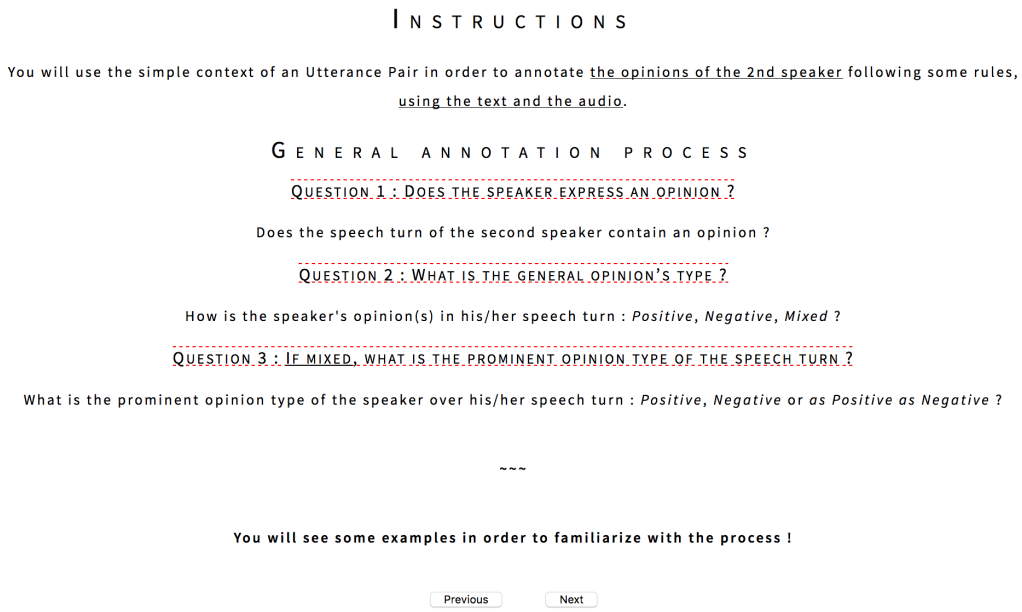


FIGURE A.7: Instructions de la plate-forme d'annotation (3/6)

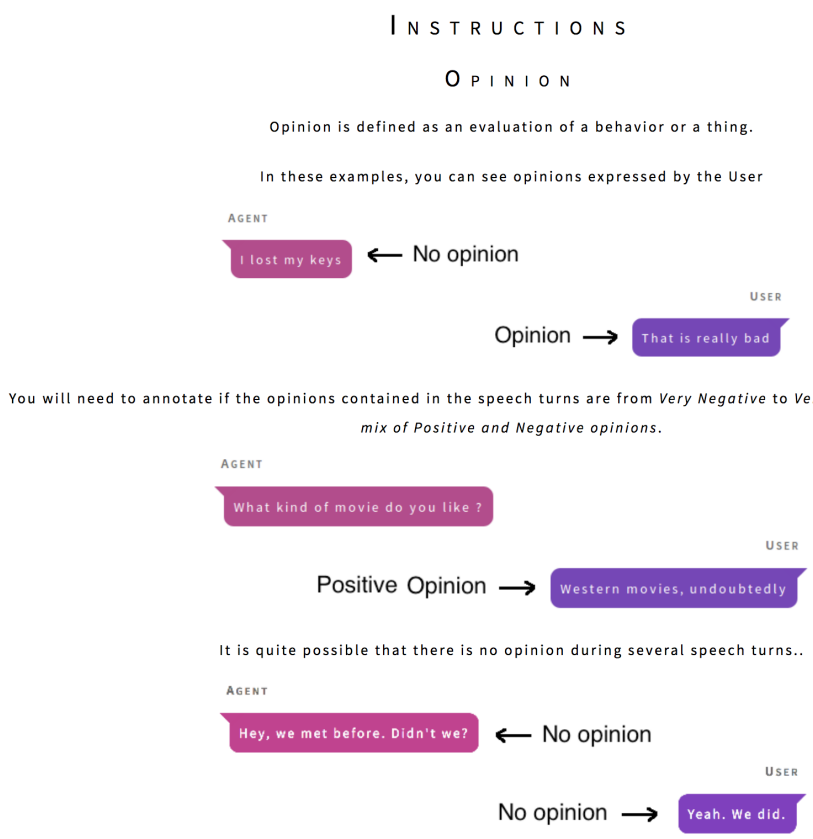


FIGURE A.8: Instructions de la plate-forme d'annotation (4/6)

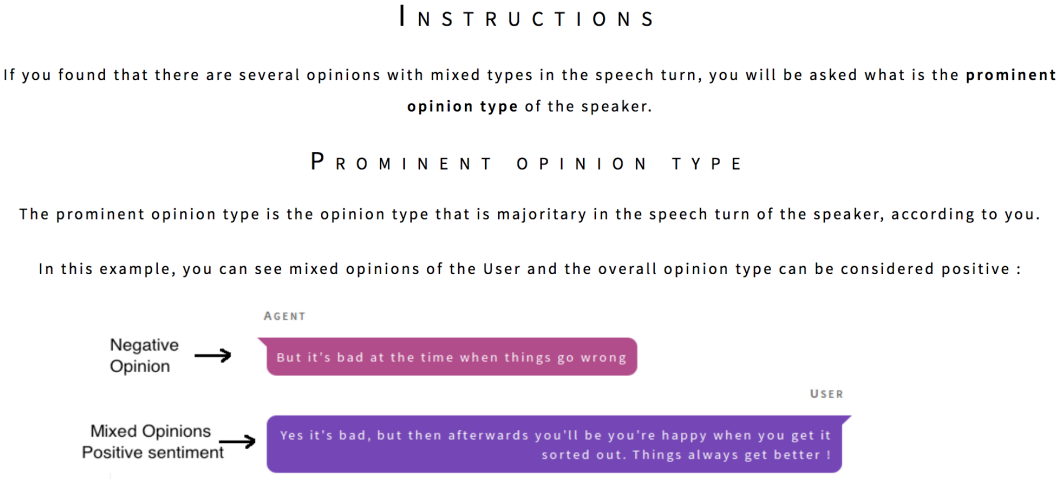


FIGURE A.9: Instructions de la plate-forme d'annotation (5/6)

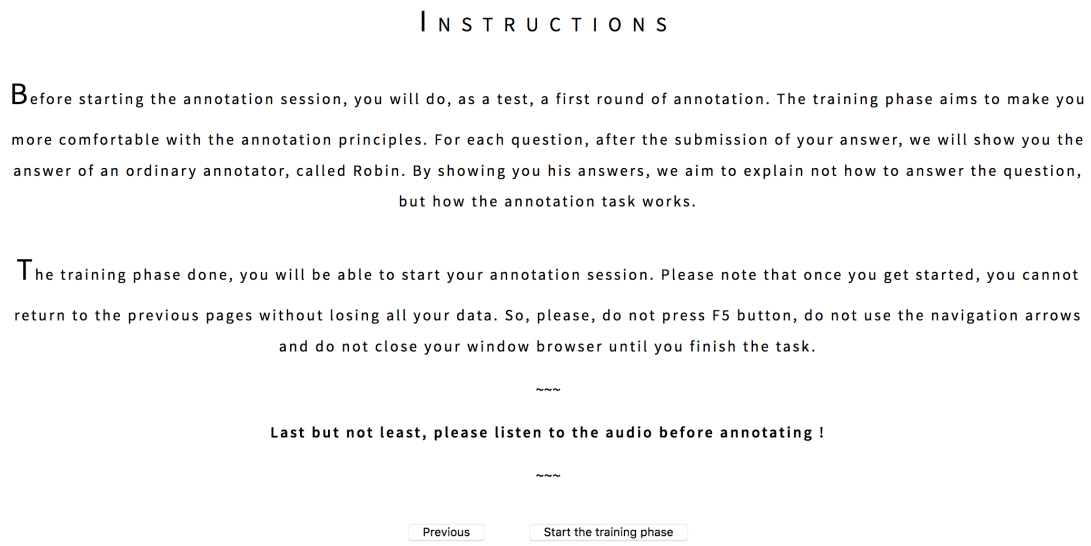


FIGURE A.10: Instructions de la plate-forme d'annotation (6/6)

B

Fusion des annotations de SEMAINE-Léger

B.1 Fusion des annotations de SEMAINE-Léger

Afin de créer une vérité terrain pour nos modèles d'apprentissage, nous avons dû choisir une manière d'agréger les étiquettes des différentes annotations et des différents annotateurs en une seule étiquette par PA.

Dans le sous-corpus **Semaine-D**, l'annotation a été faite par un seul individu. Pour les PA contenant différentes attitudes, si les valences étaient similaires nous avons étiqueté avec la valence commune, si les valences étaient différentes, nous avons écarté ces PA avec une étiquette spéciale "mixtes". Par exemple si l'utilisateur dit qu'il n'aime pas la plage, mais adore la montagne dans le même tour de parole, on est en présence d'une PA "mixte".

Pour le sous-corpus **Semaine-T**, nous avons choisi d'étiqueter "*attitude*" les PA où strictement plus de 2/3 des annotateurs ont trouvé une attitude (suivant le schéma en Figure B.1). Sur ces PA attitudes, si strictement moins de 2/3 des annotateurs ayant indiqué une attitude, étaient d'accord sur la valence dans ces PA d'attitude, nous écartions la PA avec une étiquette spéciale "mixte". Si les annotateurs étaient d'accord sur la valence, nous gardions celle-ci. Dans le cas ou plusieurs attitudes étaient présentes sur une PA, avec un même valence, nous avons étiqueté avec la valence globale. Afin d'augmenter la taille du corpus, avec l'aide d'une linguiste experte, nous avons décidé d'effecteur un arbitrage sur les données où des annotateurs avaient trouvé une attitude "mixte", en regardant quelle était l'attitude réelle contenue dans les PA en question. Ceci nous permit d'augmenter la taille du corpus final en obtenant plus de PA contenant des attitudes et les valences associées.

B.2 Limitations de SEMAINE-Léger

Typiquement, les transcriptions provenant de l'ASR, contenaient des erreurs glissées dans les index des tours de parole, provoquant un décalage par rapport à la réalité. Ces erreurs ont provoqué la création de tours de paroles factices dans le corpus

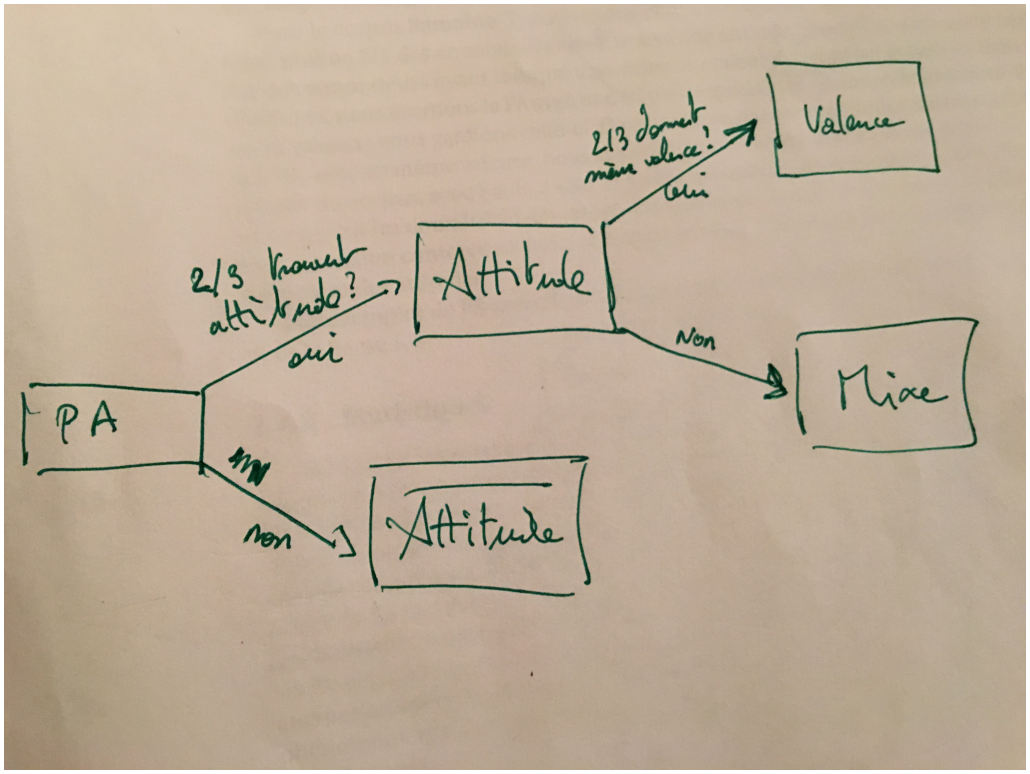


FIGURE B.1: Schéma utilisé pour l'agrégation des labels de SEMAINE-T

B

SEMAINE-T tels que décrit dans la figure B.2.

Pour éviter de reproduire un schéma similaire, nous avons choisi d'utiliser le matériel brut fournit avec le corpus SEMAINE afin d'obtenir des méta-données de qualité. Nous avons décidé de faire annoter la base de données depuis les transcriptions faites à la main, que nous avons vérifié automatiquement pour que le type de problème précédent n'ait pas lieu. De plus, en utilisant les transcriptions manuelles, nous pouvons utiliser toutes les autres informations fournies dans ces transcriptions comme la ponctuation et les annotations para-linguistiques.

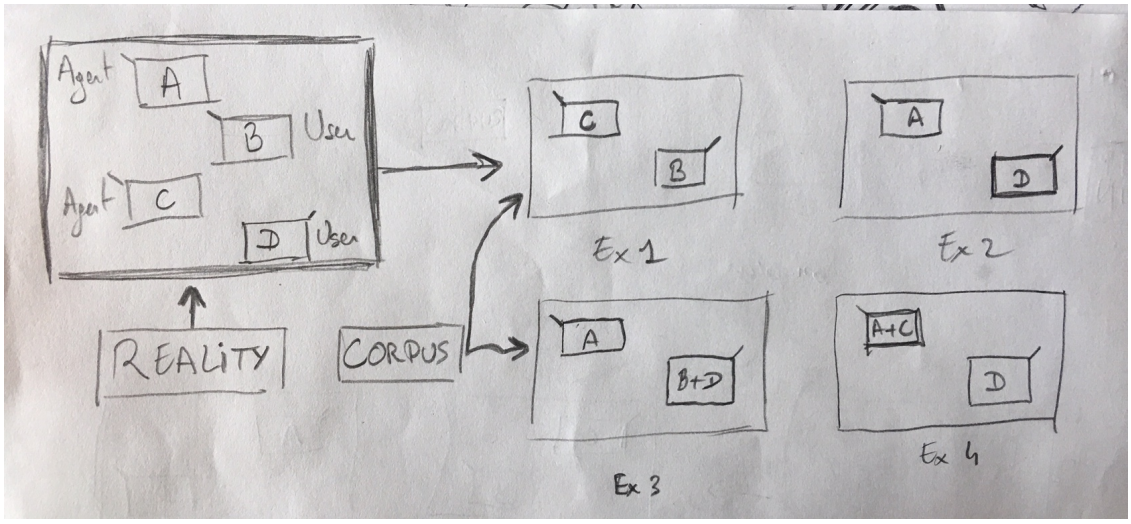


FIGURE B.2: Exemples d'erreurs apportées par la transcription automatique fournit avec SEMAINE

C

Obtention des timecodes des mots de SEMAINE-Opinions

Pour obtenir les codes temporels correspondant aux fichiers audio, nous nous sommes servis des codes temporels fournis avec les transcriptions permettant d'avoir le début et la fin d'un tour de parole. Cependant ces codes prenaient comme origine des temps le début de chaque session contenant 4 discussions et non le début de l'enregistrement audio, ainsi il était impossible de retrouver simplement le décalage . En faisant un premier passage avec un aligneur sur la discussion entière, nous avons obtenu de mauvais résultats en alignement, mais ceci a permis de retrouver le biais temporel pour chaque fichier audio.

Une fois ce biais trouvé, nous avons découpé le fichier audio en tour de parole et utilisé l'aligneur une deuxième fois de manière plus précise pour repérer les temps de début et fin de chaque mot.

D

Flux glottal

L'estimation du flux glottal est le procédé d'estimation du conduit vocal et des composantes du flux glottal depuis un signal de parole ; décrit dans Fant et collab. (1985); Fant (1995) comme le modèle Liljencrants-Fant (LF). Plus précisément, le modèle LF est un modèle à 5 paramètres du flux glottal différentié composé de 2 phases : les phases d'ouverture et de fermeture.

La phase d'ouverture est une sinusoïde croissante couplée à un coefficient exponentiel :

$$U'_g(t) = E_0 e^{\alpha t} \sin(\omega_g t), \qquad t_0 \leq t \leq t_e \tag{D.1}$$

où t_0 est le point de départ du cycle glottal, t_e est le point d'excitation maximum, $\omega_g = \pi / T_p = \frac{\pi}{t_p - t_0}$. Sachant le temps $t_e - t_0$durant lequel le flux glottal croit, α et E_0 peuvent être facilement trouvés.

La phase de fermeture est décrit comme suit dans l'équation D.2 :

$$U'g(t) = -\frac{EE}{\varepsilon T_a}(e^{-\varepsilon(t-te)} - e^{-\varepsilon Tb}), \qquad t_e < t < t_c \tag{D.2}$$

où $t_c = 1/f_0$ est l'inverse du pitch, $T_b = t_c - t_e$ la longueur du cycle glottal, et où ϵ est résolu de manière itérative suivant T_a. La méthode de Newton-Raphson est utilisé pour trouver ϵ, α et E_0 à l'aide de l'algorithme présenté par Gobl et Ní Chasaide (2003).

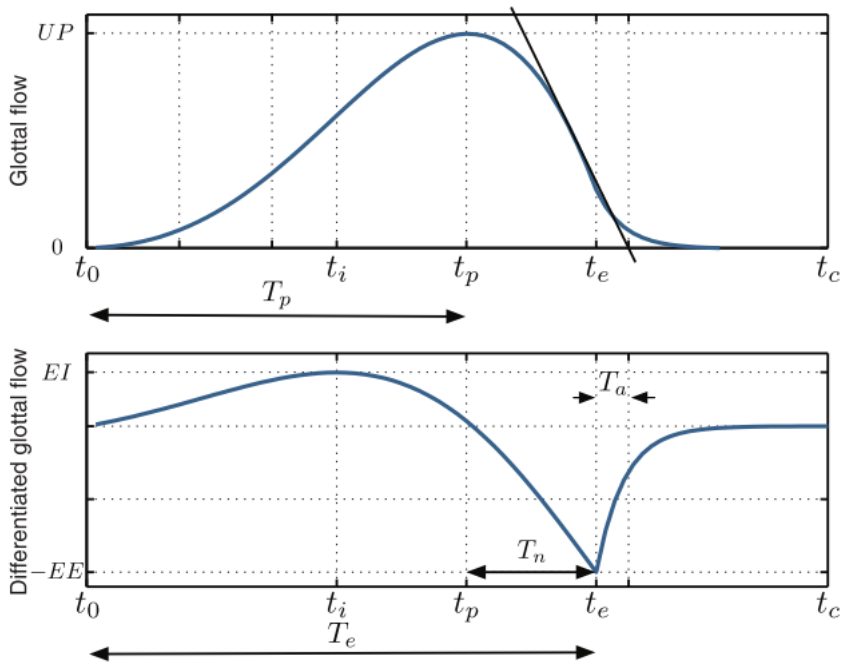


FIGURE D.1: Cycle complet du flux glottal (haut) et ses dérivées (bas) comme décrit dans Scherer et collab. (2013)

Entrainement : Le modèle Skip-gram et échantillonnage négatif

Sachant une phrase de mots $w_1, w_2, ..., w_T$ et une fenêtre de contexte de taille c, l'objectif du *Skip-gram* est de maximiser la log probabilité :

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in[\![-c;c]\!]\setminus\{0\}}\log p(w_{t+j}|w_t) \tag{E.1}$$

Où le mot w_t est représenté à l'aide du vecteur d'entrée v_{w_t} et du vecteur de sortie v'_{w_t} qui sont les projection des vecteurs BoW de w_t dans un espace de faible dimension.

La formulation classique du *Skip-gram* définit les probabilités $p(w_O|w_I)$ avec une fonction *softmax* comme décrit dans l'équation E.2,

$$p(w_O|w_I) = \frac{\exp({v'}_{w_O}^T v_{w_I})}{\sum_{w=1}^{W}\exp({v'}_{w}^T v_{w_I})} \tag{E.2}$$

Où v_w et v'_w sont les vecteur d'entrées et de sorties du mot w dans le modèle *Skip-gram*. Un aperçu est visible en Figure E.1.

Des méthodes plus efficace qu'un calcul brut sont utilisées afin d'alléger l'apprentissage. Calculer $\nabla \log p(w_O|w_I)$ est proportionnel à la taille du vocabulaire W qui est généralement entre 10^5 et 10^7 pour ce genre d'entrainement. Ainsi, l'échantillonnage négatif, le *softmax* hiérarchique et le sous-échantillonnage des mots fréquents permettent de diminuer grandement le temps d'apprentissage.

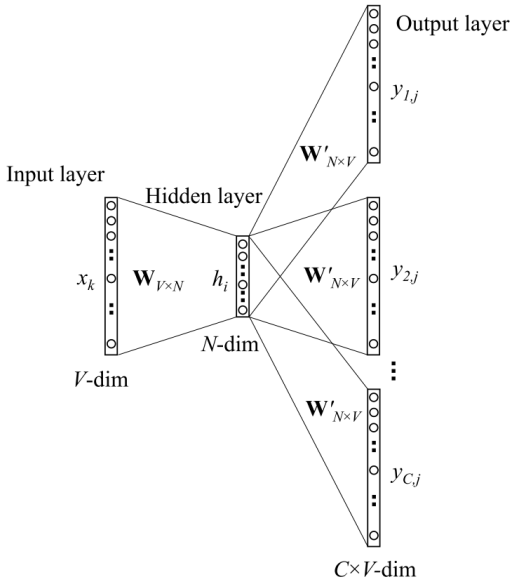


FIGURE E.1: Apprentissage par *Skip-gram*, Figure de Rong (2014)

F

Métriques

Les performances des systèmes sont calculées via différentes métriques : les scores F1 de chaque classe (voir Éq F.1d), l'exactitude (*Accuracy*) (voir Éq F.1c), et le F1 global (voir Éq F.1e).

$$Precision = \frac{VP}{VP + FP} \tag{F.1a}$$

$$Rappel = \frac{VP}{VP + FN} \tag{F.1b}$$

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \tag{F.1c}$$

$$\frac{F1_i}{2} = \left(\frac{1}{Precision_i} + \frac{1}{Rappel_i} \right)^{-1} = \frac{2 \times Precision_i \times Rappel_i}{Precision_i + Rappel_i} \tag{F.1d}$$

$$F1 = \sum_i \alpha_i F1_i \tag{F.1e}$$

Où VP représentent le nombre de vrais positifs, VN le nombre de vrais négatifs, FP le nombre de faux positifs et FN le nombre de faux négatifs. α_i peut représenter la proportion des fichiers de la classe i, ou alors être une constante égale à l'inverse du nombre de classe.

Nous calculons les scores $F1$ de chaque classe pour deux raisons. D'abord car ce sont les moyennes harmoniques des précisions et des rappels, et que nous jugeons important d'avoir une métrique reflétant autant la précision que le rappel. La précision sur une classe retranscrit la capacité du système à ne pas se tromper quand il prédit la dite classe. Si le système se trompe souvent en prédisant cette classe. Le rappel sur une classe retranscrit la capacité du système à ne pas manquer un échantillon de la dite classe. Ensuite, cela permet de bien séparer les performances des systèmes sur les différentes classes. Le taux de reconnaissance, qui est le pourcentage de vrais prédictions sur l'ensemble des prédictions, est aussi calculée afin d'observer les performances du système sur l'ensemble de la base de données, qui n'est pas équilibré. Finalement le

F1 global afin d'avoir une métrique globale autre que le taux de reconnaissance qui est faussée car le corpus n'est pas totalement balancé.

F

G

Gradient Tree Boosting

L'idée principale est là encore d'agréger plusieurs classifieurs ensembles mais en les créant manière itérative. Ces classifieurs faibles sont généralement des fonctions simples et paramétrées, dans notre cas des arbres de décision dont chaque paramètre est le critère de séparation des branches de l'arbre. Le classifieur final est une pondération (par un vecteur w) de ces classifieurs faibles. Une approche pour construire ce classifieur est de :

I. prendre une pondération quelconque (poids w_i) de classifieurs faibles (paramètres a_i) et former son classifieur global ;
II. calculer l'erreur induite par le classifieur global, et chercher le classifieur faible qui s'approche le plus de cette erreur ;
III. retrancher le classifieur faible au classifieur global tout en optimisant son poids par rapport à une fonction de perte ;
IV. répéter le procédé de manière itérative.

Le classifieur du *gradient boosting* est donc au final paramétré par les poids de pondération des différents classifieurs faibles, ainsi que par les paramètres des fonctions utilisées.

Bibliographie

Abric, J. 2008, *Psychologie de la communication : théories et méthodes.* 21, 26

Abu-El-Haija, S., N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan et S. Vijayanarasimhan. 2016, «YouTube-8M : A Large-Scale Video Classification Benchmark», . 47, 216

Airas, M. et P. Alku. 2007, «Comparison of multiple voice source parameters in different phonation types», dans *Interspeech.* 137

Albrecht, K. 2006, «Social Intelligence», *Journal of Leadership Excellence.* 22

Alku, P. 1992, «Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering», *Speech Communication*, vol. 11, nº 2-3, doi :10.1016/0167-6393(92)90005-R, p. 109–118, ISSN 01676393. 49

Alku, P., T. Bäckström et E. Vilkman. 2002, «Normalized amplitude quotient for parametrization of the glottal flow.», *The Journal of the Acoustical Society of America*, vol. 112, nº 2, doi :10.1121/1.1490365, p. 701–710, ISSN 00014966. 136

Alku, P., H. Strik et E. Vilkman. 1997, «Parabolic spectral parameter - A new method for quantification of the glottal flow», *Speech Communication*, vol. 22, nº 1, doi :10.1016/S0167-6393(97)00020-4, p. 67–79, ISSN 01676393. 134, 135

Amiriparian, S., N. Cummins, S. Ottl et M. Gerczuk. 2017, «Sentiment Analysis Using Image-based Deep Spectrum Features», dans *ACIIW*, ISBN 9781538606803, p. 26–29. 47

Angelidis, S. et M. Lapata. 2017, «Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis», *TACL.* 62, 217

Arias, P., P. Belin et J. J. Aucouturier. 2018, «Auditory smiles trigger unconscious facial imitation», doi :10.1016/j.cub.2018.05.084. 23

Atrey, P. K., M. A. Hossain, A. El Saddik et M. S. Kankanhalli. 2010, «Multimodal fusion for multimedia analysis : A survey», *Multimedia Systems*, vol. 16, nº 6, doi :10.1007/s00530-010-0182-0, p. 345–379, ISSN 09424962. 70

Baccianella, S., A. Esuli et F. Sebastiani. 2010, «SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining», *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, vol. 0, nº January, doi :citeulike-article-id:9238846, p. 2200–2204. 42, 120, 124

Baker, C. F. C., C. J. Fillmore et J. B. Lowe. 1998, «The Berkeley FrameNet Project», dans *Proceedings of the 36th annual meeting on Association for Computational Linguistics* -, vol. 1, p. 86, doi :10.3115/980845.980860. 67

Barras, C., E. Geoffrois, Z. Wu et M. Liberman. 2001, «Transcriber : Development and use of a tool for assisting speech corpora production», *Speech Communication*, vol. 33, nº 1-2, doi :10.1016/S0167-6393(00)00067-4, p. 5–22, ISSN 01676393. 81

Barriere, V. 2017, «Hybrid Models for Opinion Analysis in Speech Interactions», dans *ICMI*, ISBN 9781450355438, p. 647–651. 34

Barriere, V., C. Clavel et S. Essid. 2017, «Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields», dans *INTERSPEECH*. 34

Barriere, V., C. Clavel et S. Essid. 2018a, «Attitude Classification in Adjacency Pairs of a Human-Agent Interaction with Hidden Conditional Random Fields», dans *ICASSP*. 34

Barriere, V., C. Clavel et S. Essid. 2018b, «Classification d'attitude dans des paires adjacentes à l'aide de champs aléatoires conditionnels cachés», dans *WACAI*. 34

Ben Youssef, A., M. Chollet, H. Jones, N. Sabouret, C. Pelachaud et M. Ochs. 2015, «Towards a Socially Adaptive Virtual Agent Vizart3D : Visual articulatory feedback for Speech Therapy View project EmotionML View project Towards a Socially Adaptive Virtual Agent», dans *IVA*, p. 1–14, doi :10.1007/978-3-319-21996-7. 31

Benamara, F., M. Taboada et Y. Mathieu. 2016, «Evaluative Language Beyond Bags of Words : Linguistic Insights and Computational Applications», *Computational Linguistics*, doi :10.1162/COLI_a_00278, ISSN 0891-2017. 174

Bengio, Y., R. Ducharme, P. Vincent et C. Janvin. 2003, «A Neural Probabilistic Language Model», *The Journal of Machine Learning Research*, vol. 3, doi :10.1162/153244303322533223, p. 1137–1155, ISSN 15324435. 126

Biel, J.-I. et D. Gatica-Perez. 2013, «The youtube lens : Crowdsourced personality impressions and audiovisual analysis of vlogs», *IEEE Transactions on Multimedia*, vol. 15, nº 1, doi :10.1109/TMM.2012.2225032, p. 41–55, ISSN 15209210. 78

Bilakhia, S., S. Petridis, A. Nijholt et M. Pantic. 2015, «The MAHNOB Mimicry Database : A database of naturalistic human interactions», *Pattern Recognition Letters*, vol. 66, doi :10.1016/j.patrec.2015.03.005, p. 52–61, ISSN 01678655. 61, 79, 82, 112

Bisot, V., R. Serizel, S. Essid et G. Richard. 2016, «Acoustic scene classification with matrix factorization for unsupervised feature learning», dans *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISBN 978-1-4799-9988-0, ISSN 15206149, p. 6445–6449, doi :10.1109/ICASSP.2016.7472918. 140

Bousmalis, K., L.-P. Morency et M. Pantic. 2011, «Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition», dans *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, ISBN 9781424491407, doi :10.1109/FG.2011.5771341. 59, 60, 146

Bousmalis, K., S. Zafeiriou, L.-P. Morency et M. Pantic. 2013, «Infinite hidden conditional random fields for human behavior analysis.», *IEEE transactions on neural networks and learning systems*, vol. 24, nº 1, doi :10.1109/TNNLS.2012.2224882, p. 170–7, ISSN 2162-2388. 59

Bozkurt, B., T. Dutoit, B. Doval et C. D'Alessandro. 2004, «Improved differential phase spectrum processing for formant tracking», dans *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH) - ICSLP*, p. 1–4. 131

Bradley, M. et P. J. Lang. 1994, «Measuring Emotion : The Self-Assessment Semantic Differential Manikin and the», *Journal of Behavior Therapy and Experimental Psychiatry*, doi :10.1016/0005-7916(94)90063-9, ISSN 00057916. 42

Bradley, M. M. et P. J. Lang. 2010, «Affective Norms for English Words(ANEW) : Affective ratings of words and instruction manual», . 42

Bradley, M. M. et P. P. J. Lang. 1999, «Affective Norms for English Words (ANEW) : Instruction Manual and Affective Ratings», *Psychology*, vol. Technical, nº C-1, doi : 10.1109/MIC.2008.114, p. 0, ISSN 10897801. 42

Breck, E., Y. Choi et C. Cardie. 2007, «Identifying expressions of opinion in context», *IJCAI International Joint Conference on Artificial Intelligence*, doi :10.1016/j.jad.2005.02.015, p. 2683–2688, ISSN 10450823. 31, 41, 42, 43, 66

Brilman, M. et S. Scherer. 2015, «A Multimodal Predictive Model of Successful Debaters or How I Learned to Sway Votes», *Proceedings of the 23rd ACM international conference on Multimedia*, doi :10.1145/2733373.2806245, p. 149–158. 52

Brychcin, T., M. Konkol et J. Steinberger. 2014, «UWB : Machine Learning Approach to Aspect-Based Sentiment Analysis», *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, , nº SemEval, doi :10.3115/v1/S14-2145, p. 817–822, ISSN 1744-8352. 218

Busso, C., M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee et S. S. Narayanan. 2008, «IEMOCAP : Interactive emotional dyadic motion capture database», *Language Resources and Evaluation*, vol. 42, nº 4, doi :10.1007/s10579-008-9076-6, p. 335–359, ISSN 1574020X. 52, 80, 82

Busso, C., S. Parthasarathy, A. Burmania, M. Abdelwahab, N. Sadoughi et E. M. Provost. 2017, «MSP-IMPROV : An Acted Corpus of Dyadic Interactions to Study Emotion Perception», *IEEE Transactions on Affective Computing*, vol. 8, nº 1, doi :10.1109/TAFFC.2016.2515617, p. 67–80, ISSN 19493045. 76, 80, 81, 82, 84

Cafaro, A., N. Glas et C. Pelachaud. 2016, «The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions», *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, doi :10.1210/jc.2009-2701, p. 911–920, ISSN 1945-7197. 31

Cafaro, A., J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André et M. Valstar. 2017, «The NoXi database : multimodal recordings of mediated novice-expert interactions», dans *ICMI*. 9, 79, 82

Callejas, Z., B. Ravenet, M. Ochs et C. Pelachaud. 2014, «A computational model of social attitudes for a virtual recruiter», dans *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, vol. 1, ISBN 9781634391313, ISSN 03694232, p. 93–100. 107

Cambria, E. et A. Hussain. 2015, «Sentic Computing», *Cognitive Computation*, vol. 7, nº 2, doi :10.1007/s12559-015-9325-0, p. 183–185, ISSN 18669964. 68

Cambria, E., D. Olsher et R. Dheeraj. 2014, «SenticNet 3 :A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis», *Proceeding of Twenty-Eighth AAAI Conference on Artificial Intelligence*, p. 1515–1521. 9, 68

Cambria, E., S. Poria, D. Hazarika et K. Kwok. «SenticNet 5 : Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings», *2018*. 69

Campbell, N. et P. Mokhtari. 2003, «Voice Quality : the 4th Prosodic Dimension», *15th ICPhS*, p. 2417–2420. 136

Campione, E. et J. Véronis. 2002, «A large-scale multilingual study of pause duration», dans *Speech Prosody 2002. Proceedings of the1st International Conference on Speech Prosody*, p. 199–202, doi :10.1.1.12.561. 25, 48, 145

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope et R. Kurzweil. 2018, «Universal Sentence Encoder», . 44, 45, 217

Chelliah, M. et S. Sarkar. 2017, «Product Recommendations Enhanced with Reviews», *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*, , nº August, doi :10.1145/3109859.3109936, p. 398–399. 77

Chen, J., X. Qiu, P. Liu et X. Huang. 2018, «Meta Multi-Task Learning for Sequence Modeling», *Aaai*. 217

Chen, M., S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh et L.-p. Morency. 2017, «Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning», dans *ICMI 2017*, ISBN 9781450355438, p. 163–171. 72, 145, 186, 193

Choi, Y., E. Breck et C. Cardie. 2006, «Joint extraction of entities and relations for opinion recognition», *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, doi :10.3115/1610075.1610136, p. 431–439. 42, 66, 67, 123

Choi, Y., C. Cardie, E. Riloff et S. Patwardhan. 2005, «Identifying sources of opinions with conditional random fields and extraction patterns», *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing HLT 05*, , nº 2003, doi :10.3115/1220575.1220620, p. 355–362. 66, 123

Chollet, F. 2015, «Keras», . 163, 174

Chung, J., C. Gulcehre, K. Cho et Y. Bengio. 2015, «Gated feedback recurrent neural networks», *Proceedings of the 32nd International Conference on Machine Learning, {ICML} 2015*, vol. 37, doi :10.1145/2661829.2661935, p. 2067—-2075, ISSN 18792782. 64

Clavel, C. 2007, *Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales*, thèse de doctorat, Telecom-Paris. 46, 49, 118

Clavel, C. et Z. Callejas. 2016, «Sentiment Analysis : From Opinion Mining to Human-Agent Interaction», doi :10.1109/TAFFC.2015.2444846. 28, 54

Cohen, J. 1960, «A Coefficient of Agreement for Nominal Scales», *Educational and Psychological Measurement*, doi :10.1177/001316446002000104, ISSN 15523888. 108

Collobert, R. et J. Weston. 2008, «A unified architecture for natural language processing : Deep neural networks with multitask learning», *Proceedings of the 25th international conference on Machine learning*, doi :10.1145/1390156.1390177, p. 160–167, ISSN 07224028. 63, 126, 217

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu et P. Kuksa. 2011, «Natural Language Processing (Almost) from Scratch», *Journal of Machine Learning Research*, vol. 12, doi :10.1145/2347736.2347755, p. 2493–2537, ISSN 1532-4435. 45, 127

Conneau, A., D. Kiela, H. Schwenk, L. Barrault et A. Bordes. 2017, «Supervised Learning of Universal Sentence Representations from Natural Language Inference Data», *arXiv [cs.CL]*, doi :10.1.1.156.2685. 44, 45

Cronbach, L. J. 1951, «Coefficient alpha and the internal structure of tests», *Psychometrika*, doi :10.1007/BF02310555, ISSN 00333123. 109

Dai, P., U. Iurgel et G. Rigoll. 2003, «A Novel Feature Combination Approach for Spoken Document Classification with Support Vector Machines», *Multimedia Information Retrieval Workshop*, p. 1–5. 120

D'Alessandro, C. 2002, *Analyse, synthèse et codage de la parole*, hermes éd.. 132

Dang, T., V. Sethu et E. Ambikairajah. 2018, «Dynamic Multi-Rater Gaussian Mixture Regression Incorporating Temporal Dependencies of Emotion Uncertainty Using Kalman Filters», dans *ICASSP*, ISBN 9781538646588, p. 4929–4933. 111

Degottex, G., J. Kane, T. Drugman, T. Raitio et S. Scherer. 2014, «COVAREP - A collaborative voice analysis repository for speech technologies», dans *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, ISBN 9781479928927, ISSN 15206149, p. 960–964, doi :10.1109/ICASSP.2014.6853739. 51, 52, 140, 145

Degottex, G., A. Roebel et X. Rodet. 2011, «Function of phase-distortion for glottal model estimation», dans *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, ISBN 9781457705397, ISSN 15206149, p. 4608–4611, doi :10.1109/ICASSP.2011.5947381. 138

Devault, D., R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo et L.-p. Morency. 2014, «SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support», *2014 International Conference on Autonomous Agents and Multi-Agent Systems. International*

Foundation for Autonomous Agents and Multiagent Systems, doi :10.1016/j.imavis.2005.08.005, ISSN 1367-4803, 1460-2059. 32

Douglas-Cowie, E., R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir et K. Karpouzis. 2007, «The HUMAINE Database : Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data», *Affective Computing and Intelligent Interaction*, doi : 10.1007/978-3-540-74889-2_43, p. 488–500, ISSN 0302-9743. 58

Drugman, T. et A. Alwan. 2011, «Joint robust voicing detection and pitch estimation based on residual harmonics», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISBN 19909772 (ISSN), ISSN 19909772, p. 1973–1976. 133

Drugman, T. et T. Dutoit. 2009, «Glottal closure and opening instant detection from speech signals», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISSN 19909772, p. 2891–2894. 49

Dubuisson Duplessis, G. 2014, *Modèle de comportement communicatif conventionnel pour un agent en interaction avec des humains : Approche par jeux de dialogue*, thèse de doctorat. 25, 26

Eyben, F. 2010, «openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor Categories and Subject Descriptors», *Delta*, doi :10.1145/1873951.1874246, p. 1459–1462. 51

Eyben, F., K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan et K. Truong. 2015, «The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing», *IEEE Transactions on Affective Computing*, vol. X, nº X, doi :10.1109/TAFFC.2015.2457417, p. 1–1, ISSN 1949-3045. 46, 51, 55, 134, 140, 145

Eyben, F., F. Weninger, F. Groß, B. Schuller, F. Gross et B. Schuller. 2013, «Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor», dans *Proceedings of the 21st ACM International Conference on Multimedia (MM 2013)*, May, ISBN 9781450324045, p. 835–838, doi :10.1145/2502081.2502224. 51, 140, 145

Eyben, F., M. Wöllmer et B. Schuller. 2009, «OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit», dans *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, ISBN 9781424447992, doi :10.1109/ACII.2009.5349350. 51

Fant, G. 1970, *Acoustic Theory of Speech Production : With Calculations Based on X-Ray Studies of Russian Articulations*, ISBN 9027916004, doi :10.2307/304731. 46

Fant, G. 1981, «The source filter concept in voice production», cahier de recherche. 49

Fant, G. 1995, «The LF-model revisited. Transformations and frequency domain analysis», *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech.*, vol. 2-3, p. 121–156, ISSN 11045787. 47, 49, 130, 137, 233

Fant, G., J. Liljencrants et Q. Lin. 1985, «A four-parameter model of glottal flow», *Stlqpsr*, vol. 4, doi :10.1016/0167-6393(89)90001-0, p. 1–13, ISSN 01676393. 49, 233

Fellbaum, C. 1998, *WordNet : An Electronic Lexical Database*, vol. 71, ISBN 026206197X, 423 p., doi :10.1139/h11-025. 42, 124

Fleiss, J. L. 1971, «Measuring nominal scale agreement among many raters», *Psychological Bulletin*, doi :10.1037/h0031619, ISSN 00332909. 109

Freeman, V., J. Chan, G. Levow et R. Wright. 2014, «ATAROS Technical Report 1 : Corpus collection and initial task validation», *Depts.Washington.Edu*. 82

Freitag, M., S. Amiriparian, S. Pugachevskiy, N. Cummins et B. Schuller. 2017, «auDeep : Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks», . 46, 47, 52, 118, 141, 195, 216

Friedman, J. H. 2001, «Greedy function approximation : A gradient boosting machine», *Annals of Statistics*, doi :DOI10.1214/aos/1013203451, ISSN 00905364. 199

Garcia, A. 2017, «Annotation guide Movie review corpus», . 78

Ghosh, D., A. R. Fabbri et S. Muresan. 2017, «The Role of Conversation Context for Sarcasm Detection in Online Interactions», , nº August, p. 186–196. 55, 174

Ghosh, D., A. R. Fabbri et S. Muresan. 2018, «Sarcasm Analysis using Conversation Context», , nº Haverkate 1990, doi :arXiv:1808.07531v1, ISSN 0891-2017. 26

Ghosh, S., E. Laksana, L.-P. Morency et S. Scherer. 2016a, «Learning Representations of Affect from Speech», *Iclr 2016*, , nº 2, p. 1–8. 52

Ghosh, S., E. Laksana, L.-P. Morency et S. Scherer. 2016b, «Representation Learning for Speech Emotion Recognition», doi :10.21437/Interspeech.2016-692, p. 3603–3607, ISSN 19909772. 53

Gobl, C. et A. Ní Chasaide. 2003, «The role of voice quality in communicating emotion, mood and attitude», *Speech Communication*, vol. 40, nº 1-2, doi :10.1016/S0167-6393(02)00082-1, p. 189–212, ISSN 01676393. 49, 233

Goleman, D. 2006, «The Socially Intelligent Leader.», *Educational Leadership*, doi : Article, ISSN 0013-1784. 22

Hacki, T. 1989, «Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie», *Folia Phoniatrica et Logopaedica*. 136

Hall, J. et W. H. Watson. 1970, «The Effects of a Normative Intervention on Group Decision-Making Performance», *Human Relations*, doi : 10.1177/001872677002300404, ISSN 1741282x. 80

Hallgren, K. A. 2012, «Computing Inter-Rater Reliability for Observational Data : An Overview and Tutorial», *Tutorials in Quantitative Methods for Psychology*, vol. 8, nº 1, doi :10.20982/tqmp.08.1.p023, p. 23–34, ISSN 1913-4126. 109

Harris, Z. S. 1954, «Distributional Structure», *Word*, vol. 10, nº 2-3, doi :10.1007/978-94-009-8467-7_1, p. 146–162, ISSN 0043-7956. 120

Hazarika, D., E. Cambria et R. Zimmermann. 2018a, «Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos», dans *Naacl*. 32, 54, 64

Hazarika, D., S. Poria, R. Mihalcea, E. Cambria et R. Zimmermann. 2018b, «ICON : Interactive Conversational Memory Network for Multimodal Emotion Detection», dans *EMNLP*, p. 2594–2604. 32

He, X., Q. H. Tran, W. Havard, L. Besacier, I. Zukerman et G. Haffari. 2018, «Exploring Textual and Speech information in Dialogue Act Classification with Speaker Domain Adaptation», . 54

Heider, F. 1958, «The Psychology of Interpersonal Relations.», *American Sociological Review*, doi :10.2307/2089062, ISSN 00031224. 31

Hemamou, L., G. Felhi, V. Vandenbussche, J.-c. Martin et C. Clavel. 2019, «HireNet : a Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews», dans *AAAI*. 24

Henderson, M. 2015, «Machine Learning for Dialog State Tracking : A Review», *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*. 54

Hermansky, H. 1990, «Perceptual linear predictive (PLP) analysis of speech», *The Journal of the Acoustical Society of America*, doi :10.1121/1.399423, ISSN 0001-4966. 134

Hinton, G. E., J. L. McClelland et D. E. Rumelhart. 1986, «Distributed representations», *Parallel Distributed Processing*, doi :10.1146/annurev-psych-120710-100344, p. 77–109, ISSN 1534-7362. 43, 127

Hochreiter, S. et J. Schmidhuber. 1997, «LONG SHORT-TERM MEMORY», *Neural Computation*, vol. 9, nº 8, doi :10.1162/neco.1997.9.8.1735, p. 1735–1780, ISSN 0899-7667. 162, 163, 172, 174

Howard, J. et S. Ruder. 2018, «Fine-tuned Language Models for Text Classification», . 44, 62, 63

Hu, M. et B. Liu. 2004, «Mining and summarizing customer reviews», *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, vol. 04, doi :10.1145/1014052.1014073, p. 168, ISSN 1581138889. 67

Hughes, M. A. et D. E. Garrett. 1990, «Intercoder Reliability Estimation Approaches in Marketing : A Generalizability Theory Framework for Quantitative Data», *Journal of Marketing Research*, doi :10.2307/3172845, ISSN 00222437. 109

Hutto, C. J. et E. Gilbert. 2014, «Vader : A parsimonious rule-based model for sentiment analysis of social media text», *Eighth International AAAI Conference on Weblogs and* . . . , p. 216–225. 43, 65, 67

Irsoy, O. et C. Cardie. 2013, «Bidirectional Recursive Neural Networks for Token-Level Labeling with Structure», *Advances in neural information processing systems*, p. 1–9. 62

Irsoy, O. et C. Cardie. 2014, «Opinion mining with deep recurrent neural networks», dans *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 720–728. 45, 62

Johansson, R. et M. Alessandro. 2010, «Reranking models in fine-grained opinion analysis», . . . *of the 23rd International Conference on* . . . , , nº August, p. 519–527. 41

Kane, J. et C. Gobl. 2011, «Identifying regions of non-modal phonation using features of the wavelet transform», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISBN 19909772 (ISSN), ISSN 19909772, p. 177–180. 138

Kane, J. et C. Gobl. 2013a, «Evaluation of glottal closure instant detection in a range of voice qualities», *Speech Communication*, vol. 55, nº 2, doi :10.1016/j.specom.2012.08.011, p. 295–314, ISSN 01676393. 139

Kane, J. et C. Gobl. 2013b, «Wavelet maxima dispersion for breathy to tense voice discrimination», *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, nº 6, doi :10.1109/TASL.2013.2245653, p. 1170–1179, ISSN 15587916. 139

Kennedy, A. et D. Inkpen. 2006, «Sentiment classification of movie reviews using contextual valence shifters», dans *Computational Intelligence*, vol. 22, ISBN 1467-8640, ISSN 08247935, p. 110–125, doi :10.1111/j.1467-8640.2006.00277.x. 65, 123

Kenton, M.-w. C., L. Kristina et J. Devlin. 2018, «BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding», . 217

Khosla, S., N. Chhaya et K. Chawla. 2017, «Aff2Vec : Affect–Enriched Distributional Word Representations», . 45

Kim, M. et V. Pavlovic. 2010, «Structured output ordinal regression for dynamic facial emotion intensity prediction», dans *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, ISBN 364215557X, ISSN 03029743, doi :10.1007/978-3-642-15558-1_47. 61

Kingma, D. et J. Ba. 2014, «Adam : A Method for Stochastic Optimization», *International Conference on Learning Representations*, doi :http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503, p. 1–13, ISSN 09252312. 164, 174

Kiritchenko, S., X. Zhu et S. M. Mohammad. 2014, «Sentiment analysis of short informal texts», *Journal of Artificial Intelligence Research*, vol. 50, doi :10.1613/jair.4272, p. 723–762, ISSN 10769757. 67

Kiros, R., Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun et S. Fidler. 2015, «Skip-Thought Vectors», *ArxiV*, , nº 786, doi :10.1017/CBO9781107415324.004, p. 1–11, ISSN 1098-6596. 44, 45, 118

Krippendorff, K. 2004, *Content Analysis : An Introduction to Its Methodology*, ISBN 0761915451, doi :10.2307/2288384. 109

Krippendorff, K. 2011, «Computing Krippendorff ' s Alpha-Reliability», . 109

Krippendorff, K. 2013, «Content Analysis : An Introduction to Its Methodology», dans *Content Analysis : An Introduction to Its Methodology*, ISBN 9781412983150, doi :10.1007/s13398-014-0173-7.2. 97, 109, 111

Kumar, A., D. Kawahara et S. Kurohashi. 2018, «Knowledge-enriched Two-layered Attention Network for Sentiment Analysis», *NAACL*. 68

Lai, C., J. Carletta et S. Renals. 2013, «Modelling Participant Affect in Meetings with Turn-Taking Features», *Proceedings of WASSS 2013, Grenoble, France*. 54

Langlet, C. 2018, *Analyse des Sentiments dans les Conversations Humain-Agent : Vers un Modèle des Goûts de l'Utilisateur*, thèse de doctorat. 10, 30, 87, 94

Langlet, C. et C. Clavel. 2014, «Modelling user's attitudinal reactions to the agent utterances : focus on the verbal content», dans *5th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals (ES3 2014).*, Reykjavik, Iceland. 85, 93, 95

Langlet, C. et C. Clavel. 2015a, «Adapting sentiment analysis to face-to-face human-agent interactions : From the detection to the evaluation issues», *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, doi : 10.1109/ACII.2015.7344545, p. 14–20. 94

Langlet, C. et C. Clavel. 2015b, «Improving social relationships in face-to-face human-agent interactions : when the agent wants to know user's likes and dislikes», dans *ACL-IJCNLP 2015*, p. 1064–1073. 43, 55, 83, 85, 93, 94, 95

Langlet, C. et C. Clavel. 2016, «Grounding the detection of the user's likes and dislikes on the topic structure of human-agent interactions», *Knowledge-Based Systems*, vol. 106, doi :10.1016/j.knosys.2016.05.038, p. 116–124, ISSN 09507051. 9, 31, 66, 87, 95, 104

Langlet, C., G. D. Duplessis et C. Clavel. 2017, «A web-based platform for annotating sentiment-related phenomena in human-agent conversations», dans *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, ISBN 9783319674001, ISSN 16113349, doi :10.1007/978-3-319-67401-8_30. 102

Le, Q. et T. Mikolov. 2014, «Distributed Representations of Sentences and Documents», *International Conference on Machine Learning - ICML 2014*, vol. 32, p. 1188–1196. 45

Lewis, M., D. Yarats, Y. N. Dauphin, D. Parikh et D. Batra. 2017, «Deal or No Deal? End-to-End Learning for Negotiation Dialogues», dans *EMNLP*, doi :arXivpreprintarXiv: 1706.05125. 21

Liang, P. P., Z. Liu, A. Zadeh et L.-P. Morency. 2018, «Multimodal Language Analysis with Recurrent Multistage Fusion», dans *EMNLP*. 24

Likert, R. 1932, «A technique for the measurement of attitudes», doi :2731047. 77

Liu, B. 2012, «Sentiment Analysis and Opinion Mining», *Synthesis Lectures on Human Language Technologies*, vol. 5, nº 1, doi :10.2200/S00416ED1V01Y201204HLT016, p. 1–167, ISSN 1947-4040. 218

Lombard, M., J. Snyder-Duch et C. C. Bracken. 2002, «Content Analysis in Mass Communication : Assessment and Reporting of Intercoder Reliability», doi :10.1093/hcr/28.4.587. 108, 109

Lotfian, R. et C. Busso. 2017, «Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings», *IEEE Transactions on Affective Computing*, vol. XX, nº X, doi :10.1109/TAFFC.2017.2736999, p. 1–14, ISSN 19493045. 80, 82

Ma, Y., H. Peng et E. Cambria. 2018, «Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM», dans *AAAI Conference on Artificial Intelligence (AAAI-18)*. 68

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng et C. Potts. 2011, «Learning Word Vectors for Sentiment Analysis», *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, doi : 978-1-932432-87-9, p. 142–150. 71

Mairesse, F., J. Polifroni et G. Di Fabbrizio. 2012, «Can prosody inform sentiment analysis ? Experiments on short spoken reviews», *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, doi :10.1109/ICASSP.2012.6289066, p. 5093–5096, ISSN 15206149. 30

Majumder, N., S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh et E. Cambria. 2018, «DialogueRNN : An Attentive RNN for Emotion Detection in Conversations», . 32

Mandelbrot, B. 1957, «Etude de la loi d'Estoup et de Zipf : fréquences des mots dans le discours», *Logique, langage et théorie de l'information*. 121

Marchi, E., A. Batliner, B. Schuller, S. Fridenzon, S. Tal et O. Golan. 2012, «Speech, emotion, age, language, task, and typicality : Trying to disentangle performance and feature relevance», dans *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, ISBN 9780769548487, p. 961–968, doi : 10.1109/SocialCom-PASSAT.2012.97. 49

Martin, J. R. et P. R. R. White. 2003, «The Language of Evaluation : The Appraisal Framework», *Lecture Notes in Computer Science*, doi :10.1057/9780230511910, p. 256, ISSN 00426806. 29, 93, 94

Martin, O., I. Kotsia, B. Macq et I. Pitas. 2006, «The eNTERFACE'05 audio-visual emotion database», *International Conference on Data Engineering Workshops*, , nº 1, doi : 10.1109/ICDEW.2006.145, p. 8–15. 94

Mayer, J. D., P. Salovey et D. R. Caruso. 2008, «Emotional Intelligence : New Ability or Eclectic Traits?», *American Psychologist*, vol. 63, nº 6, doi :10.1037/0003-066X.63.6.503, p. 503–517, ISSN 0003066X. 22

Mccallum, A., D. Freitag et F. Pereira. 2000, «Maximum Entropy Markov Models for Information Extraction and Segmentation», *Icml*, p. 591–598. 59

Mccann, B., N. S. Keskar, C. Xiong et R. Socher. 2018, «The Natural Language Decathlon : Multitask Learning as Question Answering», Nips. 62, 63

McCowan, I., J. Carletta et W. Kraaij. 2005, «The AMI meeting corpus», *Proceedings Methods and Techniques in Behavioral Research*, doi :10.1016/S0271-5309(99)00018-X, p. 137–140, ISSN 0037-1998. 54, 79

McKeown, G., M. Valstar, R. Cowie, M. Pantic et M. Schröder. 2012, «The SEMAINE database : Annotated multimodal records of emotionally colored conversations between a person and a limited agent», *IEEE Transactions on Affective Computing*, vol. 3, nº 1, doi :10.1109/T-AFFC.2011.20, p. 5–17, ISSN 19493045. 9, 33, 58, 83, 85, 92, 93, 95, 96, 98

McKeown, G., M. F. Valstar, R. Cowie et M. Pantic. 2010, «The semaine corpus of emotionally coloured character interactions», *2010 IEEE International Conference on Multimedia and Expo, ICME 2010*, doi :10.1109/ICME.2010.5583006, p. 1079–1084, ISSN 1945-7871. 80, 82, 83

Mehrabian, A. 1971, *Silent messages*, ISBN 0534000592, viii, 152 p. p.. 30

Mikolov, T., G. Corrado, K. Chen et J. Dean. 2013a, «Efficient Estimation of Word Representations in Vector Space», *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, doi :10.1162/153244303322533223, p. 1–12, ISSN 15324435. 41, 119, 126, 163, 216

Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch et A. Joulin. 2017, «Advances in Pre-Training Distributed Word Representations», , nº 1. 44

Mikolov, T., I. Sutskever, K. Chen, G. Corrado et J. Dean. 2013b, «Distributed Representations of Words and Phrases and their Compositionality», dans *Proc. NIPS*, ISBN 2150-8097, ISSN 10495258, p. 1–9, doi :10.1162/jmlr.2003.3.4-5.951. 173

Mikolov, T., W.-t. Yih et G. Zweig. 2013c, «Linguistic regularities in continuous space word representations», *Proceedings of NAACL-HLT*, , nº June, p. 746–751. 45

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross et K. J. Miller. 1990, «Introduction to wordnet : An on-line lexical database», *International Journal of Lexicography*, vol. 3, nº 4, doi :10.1093/ijl/3.4.235, p. 235–244, ISSN 09503846. 66

Mitchell, M., J. Aguilar, T. Wilson et B. V. Durme. 2013, «Open Domain Targeted Sentiment», *EMNLP*, , nº October, p. 1643–1654. 61, 67

Mohammad, S. M. 2018, «Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words», *Proceedings of the 56th Annual Meeting ofthe Association for Computational Linguistics*, p. 1–11. 42

Mordatch, I. et P. Abbeel. 2017, «Emergence of Grounded Compositional Language in Multi-Agent Populations», . 21

Morency, L.-P. 2007, «Hidden-state Conditional Random Field (HCRF) Library», . 164, 175

Morency, L.-p., R. Mihalcea et P. Doshi. 2011, «Towards Multimodal Sentiment Analysis : Harvesting Opinions from the Web», *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI-11)*, doi :10.1145/2070481.2070509, p. 169–176. 9, 42, 59, 68, 77, 123, 146

Morency, L.-P., L.-P. Morency, A. Quattoni, A. Quattoni, T. Darrell et T. Darrell. 2007, «Latent-Dynamic Discrimiative Models for Continuous Gesture Recognition», *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 58

Munezero, M., C. S. Montero, E. Sutinen et J. Pajunen. 2014, «Are they different? affect, feeling, emotion, sentiment, and opinion detection in text», *IEEE Transactions on Affective Computing*, vol. 5, nº 2, doi :10.1109/TAFFC.2014.2317187, p. 101–111, ISSN 19493045. 9, 28, 29, 30

Musto, C., G. Semeraro et M. Polignano. 2014, «A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts», dans *DART 2014 8th Internation Workshop on Information Filtering and Retrieval*, ISSN 16130073. 42

Narayanan, S. et A. Potamianos. 2002, «Creating conversational interfaces for children», *IEEE Transactions on Speech and Audio Processing*, vol. 10, nº 2, doi :10.1109/89.985544, p. 65–78, ISSN 10636676. 80

Nielsen, F. Å. 2011, «A new ANEW : Evaluation of a word list for sentiment analysis in microblogs», dans *CEUR Workshop Proceedings*, vol. 718, ISSN 16130073, p. 93–98. 42

Nojavanasghari, B., D. Gopinath, J. Koushik, T. Baltrušaitis et L.-P. Morency. 2016, «Deep Multimodal Fusion for Persuasiveness Prediction», dans *ICMI 2016 - Proceedings of the 2016 ACM International Conference on Multimodal Interaction*, ISBN 9781450345569, p. 1–5, doi :10.1145/2993148.2993176. 70, 71

Ochshorn, R. M. et M. Hawkins. 2017, «Gentle forced aligner», . 101

Palau, R. M. et M.-F. Moens. 2009, «Argumentation Mining : The Detection, Classification and Structure of Arguments in Text», dans *ICAIL*, p. 98–107. 25

Paleari, M. et B. Huet. 2008, «Toward emotion indexing of multimedia excerpts», dans *2008 International Workshop on Content-Based Multimedia Indexing, CBMI 2008, Conference Proceedings*, ISBN 9781424420445, p. 425–432, doi :10.1109/CBMI.2008.4564978. 71

Pang, B. et L. Lee. 2004, «A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts», *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, doi :10.3115/1218955.1218990, p. 271, ISSN 1554-0669. 31, 40, 77

Park, S., J. Gratch et L.-P. Morency. 2012, «I already know your answer : Using nonverbal behaviors to predict immediate outcomes in a dyadic negotiation», *... of the 14th ACM international conference ...*, doi :10.1145/2388676.2388682, p. 19–22. 82

Park, S., S. Scherer, J. Gratch, P. J. Carnevale et L.-P. Morency. 2015, «I can already guess your answer : Predicting respondent reactions during dyadic negotiation», *IEEE Transactions on Affective Computing*, vol. 6, nº 2, doi :10.1109/TAFFC.2015.2396079, p. 86–96, ISSN 19493045. 52

Park, S., H. S. Shim, M. Chatterjee, K. Sagae et L.-P. Morency. 2014, «Computational Analysis of Persuasiveness in Social Multimedia : A Novel Dataset and Multimodal Prediction Approach», *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, doi :10.1145/2663204.2663260, p. 50–57. 52, 78

Paul, R., A. Augustyn, A. Klin et F. R. Volkmar. 2005, «Perception and production of prosody by speakers with autism spectrum disorders», *Journal of Autism and Developmental Disorders*, doi :10.1007/s10803-004-1999-1, ISSN 01623257. 26, 48

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot et É. Duchesnay. 2012, «Scikit-learn : Machine Learning in Python», *Journal of Machine Learning Research*, vol. 12, doi :10.1007/s13398-014-0173-7.2, p. 2825–2830, ISSN 15324435. 163, 174

Pellegrini, T. et V. Barriere. 2015, «Time-continuous estimation of emotion in music with recurrent neural networks», dans *CEUR Workshop Proceedings*, vol. 1436, ISSN 16130073. 51

Pennington, J., R. Socher et C. D. Manning. 2014, «GloVe : Global Vectors for Word Representation», *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, doi :10.3115/v1/D14-1162, p. 1532–1543, ISSN 10495258. 45, 127

Pentland, A. 2007, «Social Signal Processing [Exploratory DSP]», *IEEE Signal Processing Magazine*, doi :10.1109/MSP.2007.4286569, ISSN 1053-5888. 21

Perez-Rosas, V., R. Mihalcea et L.-P. Morency. 2013a, «Multimodal sentiment analysis of spanish online videos», *IEEE Intelligent Systems*, vol. 28, nº 3, doi :10.1109/MIS.2013.9, p. 38–45, ISSN 15411672. 78

Perez-Rosas, V., R. Mihalcea et L.-P. Morency. 2013b, «Utterance-Level Multimodal Sentiment Analysis», *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 973–982. 71, 164

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee et L. Zettlemoyer. 2018a, «Deep contextualized word representations», dans *Naacl*. 44, 45

Peters, M. E., M. Neumann, L. Zettlemoyer et W.-t. Yih. 2018b, «Dissecting Contextual Word Embeddings : Architecture and Representation», doi :arXiv:1808.08949v2, p. 1499–1509. 217

Polanyi, L. et A. Zaenen. 2006, «Contextual valence shifters», *Computing attitude and affect in text : Theory and Applications*, vol. 20, nº 1, doi :10.1007/1-4020-4102-0_1, p. 1–10, ISSN 4020-4026. 65, 154

Poria, S., E. Cambria, R. Bajpai et A. Hussain. 2017a, «A Review of Affective Computing : From Unimodal Analysis to Multimodal Fusion», , p. 1–34. 9, 24, 88

Poria, S., E. Cambria et A. Gelbukh. 2015, «Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis», dans *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, ISBN 9781941643327, p. 2539–2544. 45, 70, 71, 123, 127

Poria, S., E. Cambria et A. Gelbukh. 2016, «Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network», *Knowledge-Based Systems*, vol. 108, doi :10.1016/j.knosys.2016.06.009, p. 42–49, ISSN 09507051. 43, 68

Poria, S., E. Cambria, D. Hazarika, C. Science et N. Mazumder. 2017b, «Context-Dependent Sentiment Analysis in User-Generated Videos», dans *ACL 2017*. 72, 183

Porter, M. 1980, «An algorithm for suffix stripping», *Program : electronic library and information systems*, vol. 14, nº 3, doi :10.1108/eb046814, p. 130–137, ISSN 0033-0337. 163

Quattoni, A., S. Wang, L.-P. Morency, M. Collins et T. Darrell. 2007, «Hidden-state conditional random fields.», *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, nº 10, doi :10.1109/TPAMI.2007.1124, p. 1848–1853, ISSN 0162-8828. 59, 74, 148, 149

Rabiner, L. et B. Juang. 1986, «An Introduction to Hidden Markov Models», *IEEE ASSP MAGAZINE*, , nº January, doi :10.1002/0471250953.bia03as18, ISSN 1934-340X. 59

Radford, A. et T. Salimans. 2018, «Improving Language Understanding by Generative Pre-Training», , p. 1–12. 63

Rajagopalan, S. S., L.-P. Morency, T. Baltrušaitis et R. Goecke. 2016, «Extending long short-term memory for multi-view structured learning», dans *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9911 LNCS, ISBN 9783319464770, ISSN 16113349, p. 338–353, doi :10.1007/978-3-319-46478-7_21. 72

Rakicevic, N., O. Rudovic, S. Petridis et M. Pantic. 2017, «Multi-modal Neural Conditional Ordinal Random Fields for agreement level estimation», *Proceedings - International Conference on Pattern Recognition*, doi :10.1109/ICPR.2016.7899967, p. 2228–2233, ISSN 10514651. 60

Ramirez, G. A., T. Baltrušaitis et L.-P. Morency. 2011, «Modeling latent discriminative dynamic of multi-dimensional affective signals», *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6975 LNCS, nº PART 2, doi :10.1007/978-3-642-24571-8_51, p. 396–406, ISSN 03029743. 58

Ringeval, F. 2011, *Ancrages et modèles dynamiques de la prosodie : application à la reconnaissance des émotions actées et spontanées*, thèse de doctorat. 22, 25

Ringeval, F., A. Sonderegger, J. Sauer et D. Lalanne. 2013, «Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions», dans *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, ISBN 9781467355452, doi :10.1109/FG.2013.6553805. 80

Rizzo, A. A., S. Scherer, D. Devault, J. Gratch, R. Artstein, A. Hartholt, C et L.-p. Morency. 2014, «Detection and computational analysis of psychological signals using a virtual human interviewing agent», *Proceedings of the 10th Intl Conf. Disability, Virtual Reality & Associated Technologies*, doi :10.1111/j.1472-765X.2004.01606.x, p. 2–4, ISSN 19395914. 32

Röbel, A. et X. Rodet. 2005, «Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation», *DAFx-05*, doi :10.1.1.66.1759, p. 1–6. 132

Rong, X. 2014, «word2vec Parameter Learning Explained», *arXiv :1411.2738*, p. 1–19. 11, 236

Sahay, S., S. H. Kumar, R. Xia, J. Huang et L. Nachman. 2018, «Multimodal Relational Tensor Network for Sentiment and Emotion Classification», . 24, 45, 67, 72, 145

Sanh, V., T. Wolf, S. Ruder, H. Court et H. Row. 2018, «A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks», . 217

Scherer, K. R. 2003, «Vocal communication of emotion : A review of research paradigms», *Speech Communication*, vol. 40, nº 1-2, doi :10.1016/S0167-6393(02)00084-5, p. 227–256, ISSN 01676393. 52

Scherer, K. R. 2005, «What are emotions? and how can they be measured?», *Social Science Information*, vol. 44, nº 4, doi :10.1177/0539018405058216, p. 695–729, ISSN 05390184. 29, 130

Scherer, S., Z. Hammal, Y. Yang, L.-P. Morency et J. F. Cohn. 2014, «Dyadic Behavior Analysis in Depression Severity Assessment Interviews», dans *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, ISBN 9781450328852, p. 112–119, doi :10.1145/2663204.2663238. 52, 136

Scherer, S., J. Kane, C. Gobl et F. Schwenker. 2013, «Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification», *Computer Speech and Language*, vol. 27, nº 1, doi :10.1016/j.csl.2012.06.001, p. 263–287, ISSN 08852308. 10, 11, 49, 137, 139, 234

Schuller, B. 2013, «Intelligent Audio Analysis», *Signals and Communication Technology*. 135

Schuller, B., A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous et V. Aharonson. 2007, «The relevance of feature type for the automatic classification of emotional user states : Low level descriptors and functionals», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2, ISBN 9781605603162, ISSN 19909772, p. 881–884. 49

Schuller, B. et G. Rigoll. 2009, «Recognising interest in conversational speech - Comparing bag of frames and supra-segmental features», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISSN 19909772, p. 1999–2002. 49

Schuller, B., J. Schenk, G. Rigoll et T. Knaup. 2009a, «"The Godfather" vs. "Chaos" : Comparing linguistic analysis based on on-line knowledge sources and bags-of-N-grams for movie review valence estimation», *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, doi :10.1109/ICDAR.2009.194, p. 858–862, ISSN 15205363. 41, 119, 164

Schuller, B., S. Steidl et A. Batliner. 2009b, «The INTERSPEECH 2009 emotion challenge», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISBN 978-1-61567-692-7, ISSN 19909772, p. 312–315. 13, 50, 51, 118

Schuller, B., S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho et K. Evanini. 2016, «The INTERSPEECH 2016 Computational Paralinguistics Challenge : Deception, Sincerity & Native Language», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 51

Schuller, B., S. Steidl, A. Batliner, F. Schiel et J. Krajewski. 2011a, «The INTERSPEECH 2011 speaker state challenge», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISBN 19909772 (ISSN), ISSN 19909772, p. 3201–3204. 51

Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente et S. Kim. 2013, «The INTERSPEECH 2013 computational paralinguistics challenge : Social signals, conflict, emotion, autism», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISSN 19909772, p. 148–152. 51

Schuller, B., M. Valstar, F. Eyben, G. McKeown, R. Cowie et M. Pantic. 2011b, «AVEC 2011 - The first international audio/visual emotion challenge», dans *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, ISBN 9783642245701, ISSN 03029743, doi :10.1007/978-3-642-24571-8_53. 59, 115, 217

Schuster, S. et C. D. Manning. 2016, «Enhanced English Universal Dependencies : An Improved Representation for Natural Language Understanding Tasks», *Proceedings of LREC 2016*, p. 2371–2378. 99, 123, 163, 173

Scott, W. A. 1955, «Reliability of Content Analysis : The Case of Nominal Scale Coding», *Public Opinion Quarterly*, doi :10.1086/266577, ISSN 0033362X. 108

Shannon, C. E. 1948, «A Mathematical Theory of Communication», *Bell System Technical Journal*, vol. 27, nº July 1928, p. 623–656. 21, 22

Shin, B., T. Lee et J. D. Choi. 2016, «Lexicon Integrated CNN Models with Attention for Sentiment Analysis», doi :10.18653/V1/W17-5220, p. 149–158. 67

Shouse, E. 2005, «Feeling, emotion, affect», *Journal of Media and Culture (M/C)*, doi : 10.1177/1476127007077559. 28

Shriberg, E., A. Stolcke, D. Hakkani-Tür et G. Tür. 2000, «Prosody-based automatic segmentation of speech into sentences and topics», *Speech Communication*, doi : 10.1016/S0167-6393(00)00028-5, ISSN 01676393. 25, 48

Sigmund, M. 2013, «Statistical analysis of fundamental frequency based features in speech under stress», *Information Technology and Control*, vol. 42, nº 3, doi :10.5755/j01.itc.42.3.3895, p. 286–291, ISSN 1392124X. 201

Socher, R., A. Perelygin et J. Wu. 2013, «Recursive deep models for semantic compositionality over a sentiment treebank», *EMNLP-2013 : Conference on Empirical Methods in Natural Language Processing*, doi :10.1371/journal.pone.0073791, p. 1631–1642, ISSN 1932-6203. 9, 62, 63

Soleymani, M., D. Garcia, B. Jou, B. Schuller, S. F. Chang et M. Pantic. 2017, «A survey of multimodal sentiment analysis», *Image and Vision Computing*, vol. 65, doi :10.1016/j.imavis.2017.08.003, p. 3–14, ISSN 02628856. 30

Somasundaran, S., J. Wiebe et J. Ruppenhofer. 2008, «Discourse Level Opinion Interpretation», *Proceedings of the 22nd International Conference on Computational Linguistics*, , nº August, doi :10.3115/1599081.1599182, p. 801–808. 54

Song, Y., L.-P. Morency et R. Davis. 2012a, «Multi-view latent variable discriminative models for action recognition», *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi :10.1109/CVPR.2012.6247918, p. 2120–2127, ISSN 10636919. 9, 60, 211

Song, Y., L.-P. Morency et R. Davis. 2012b, «Multimodal Human Behavior Analysis : Learning Correlation and Interaction Across Modalities», *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12*, doi :10.1145/2388676.2388684, p. 27–30, ISSN 9781450314671. 60, 152, 211

Song, Y., L.-p. Morency et R. Davis. 2013, «Action Recognition by Hierarchical Sequence Summarization», dans *IEEE Conference on Computer Vision and Pattern Recognition*, doi :10.1109/CVPR.2013.457. 218

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever et R. Salakhutdinov. 2014, «Dropout : A Simple Way to Prevent Neural Networks from Overfitting», *Journal of Machine Learning Research*, vol. 15, doi :10.1214/12-AOS1000, p. 1929–1958, ISSN 15337928. 163, 174

Strapparava, C. et A. Valitutti. 2004, «WordNet-Affect : an affective extension of WordNet», *Proceedings of the 4th International Conference on Language Resources and Evaluation*, doi :10.1.1.122.4281, p. 1083–1086. 42

Sutskever, I., O. Vinyals et Q. V. Le. 2014, «Sequence to Sequence Learning with Neural Networks», *Nips*, doi :10.1007/s10107-014-0839-0, p. 9, ISSN 09205691. 142

Taboada, M., J. Brooke, M. Tofiloski, K. Voll et M. Stede. 2011, «Lexicon-Based Methods for Sentiment Analysis», *Computational Linguistics*, vol. 37, nº 2, doi :10.1162/COLI_a_00049, p. 267–307, ISSN 0891-2017. 40, 63, 65, 120, 123, 125

Täckström, O. et R. McDonald. 2011, «Discovering Fine-Grained Sentiment with Latent Variable Structured Prediction Models», dans *European Conference on Information Retrieval*, ISBN 978-3-642-20160-8, ISSN 16113349, p. 368–374, doi :10.1007/978-3-642-20161-5_37. 59, 153

Tahon, M., G. Degottex et L. Devillers. 2012, «Usual voice quality features and glottal features for emotional valence detection», *Proceedings of Speech ...*, p. 2–5. 137

Tahon, M. et L. Devillers. 2016, «Towards a Small Set of Robust Acoustic Features for Emotion Recognition : Challenges», *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 24, nº 1, p. 16–28. 46

Tang, D., F. Wei, N. Yang, M. Zhou, T. Liu et B. Qin. 2014, «Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification», *Acl*, doi :10.3115/1220575.1220648, p. 1555–1565, ISSN 03029743. 45

Ten Bosch, L., N. Oostdijk et L. Boves. 2005, «On temporal aspects of turn taking in conversational dialogues», doi :10.1016/j.specom.2005.05.009. 25

Titze, I. R. et J. Sundberg. 1992, «Vocal intensity in speakers and singers», *Journal of the Acoustical Society of America*, vol. 91, nº 5, doi :10.1121/1.402929, p. 2936–2946, ISSN 0001-4966. 137

Tokuda, K., T. Kobayashi, T. Masuko et S. Imai. 1994, «MEL-GENERALIZED CEPSTRAL ANALYSIS - A UNIFIED APPROACH TO SPEECH SPECTRAL ESTIMATION», dans *In Proc. of ICSLP*, p. 1043–1046. 133

Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller et S. Zafeiriou. 2016, «Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network», dans *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISBN 978-1-4799-9988-0, ISSN 15206149, p. 5200–5204, doi :10.1109/ICASSP.2016.7472669. 9, 47, 53, 140

Turney, P. D. 2002, «Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews», *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, , nº July, doi :10.3115/1073083.1073153, p. 417–424, ISSN 0738467X. 77

Tzirakis, P., G. Trigeorgis, M. A. Nicolaou, B. Schuller et S. Zafeiriou. 2017, «End-to-End Multimodal Emotion Recognition using Deep Neural Networks», vol. 14, nº 8, p. 1–9. 46, 47

Varni, G., I. Hupont, C. Clavel et M. Chetouani. 2017, «Computational Study of Primitive Emotional Contagion in Dyadic Interactions», *IEEE Transactions on Affective Computing*, vol. 14, nº 8, doi :10.1109/TAFFC.2017.2778154, p. 1–1, ISSN 1949-3045. 31, 112

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser et I. Polosukhin. 2017, «Attention Is All You Need», *arXiv :1706.03762 [cs]*. 63

Venek, V., S. Scherer, L.-P. Morency, A. S. Rizzo et J. Pestian. 2014, «Adolescent suicidal risk assessment in clinician-patient interaction : A study of verbal and acoustic behaviors», dans *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, ISBN 9781479971299, p. 277–282, doi :10.1109/SLT.2014.7078587. 52

Vinciarelli, A., M. Pantic et H. Bourlard. 2009, «Social signal processing : Survey of an emerging domain», *Image and Vision Computing*, vol. 27, nº 12, doi :10.1016/j.imavis.2008.11.007, p. 1743–1759, ISSN 02628856. 21

Vinciarelli, A., M. Pantic, H. Bourlard et A. Pentland. 2008, «Social Signal Processing : State-of-the-Art and Future Perspectives of an Emerging Domain», *Proceedings of the 16th ACM international conference on Multimedia*, doi :10.1145/1459359.1459573, p. 1061–1070, ISSN 02628856. 22

Vinciarelli, A., M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico et M. Schroeder. 2011, «Bridging the Gap Between Social Animal and Unsocial Machine : A Survey of Social Signal Processing», *IEEE Transactions on Affective Computing (to appear)*, vol. 3, nº 1, doi :10.1109/T-AFFC.2011.27.1949-3045/12/, p. 1–20, ISSN 1949-3045. 23

Wang, W., S. J. Pan, D. Dahlmeier et X. Xiao. 2016, «Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis», *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*. 67

Warren, P. 1996, «Prosody and Parsing : An Introduction», *Language and Cognitive Processes*, doi :10.1080/016909696387204, ISSN 0169-0965. 25, 48

Warriner, A. B., V. Kuperman et M. Brysbaert. 2013, «Norms of valence, arousal, and dominance for 13,915 English lemmas», *Behavior Research Methods*, vol. 45 VN - r, nº 4, doi :10.3758/s13428-012-0314-x, p. 1191–1207, ISSN 1554-3528. 40, 42, 45, 120, 124

Weninger, F., F. Eyben, B. Schuller, M. Mortillaro et K. R. Scherer. 2013, «On the acoustics of emotion in audio : What speech, music, and sound have in common», *Frontiers in Psychology*, vol. 4, nº MAY, doi :10.3389/fpsyg.2013.00292, p. 1–12, ISSN 16641078. 51

Wiebe, J., T. Wilson et C. Cardie. 2005, «Annotating expressions of opinions and emotions in language», *Language Resources and Evaluation*, vol. 39, nº 2-3, doi :10.1007/s10579-005-7880-9, p. 165–210, ISSN 1574020X. 62, 65

Wierzbicka, A. 1999, *Emotions Across Languages and Cultures : Diversity and Unversals*. 29

Wilson, T., J. Wiebe et P. Hoffman. 2005, «Recognizing contextual polarity in phrase level sentiment analysis», *ACL*, vol. 7, nº 5, doi :10.3115/1220575.1220619, p. 12–21, ISSN 0891-2017. 40, 42, 65, 67

Wöllmer, M., F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie et R. Cowie. 2008, «Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies», dans *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISSN 19909772, p. 597–600. 58

Wöllmer, M., M. Kaiser, F. Eyben, B. Schuller et G. Rigoll. 2013a, «LSTM-modeling of continuous emotions in an audiovisual affect recognition framework», *Image and Vision Computing*, vol. 31, nº 2, doi :10.1016/j.imavis.2012.03.001, p. 153–163, ISSN 02628856. 10, 85, 86, 111, 162

Wöllmer, M., F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae et L.-P. Morency. 2013b, «YouTube Movie Reviews : Sentiment Analysis in an Audio-Visual Context», *IEEE Intelligent Systems*, vol. 28, nº 3, doi :10.1109/MIS.2013.34, p. 46–53, ISSN 1541-1672. 71, 78, 81, 154, 163, 172, 188, 189

Xu, C., D. Tao et C. Xu. 2013, «A Survey on Multi-view Learning», doi :10.1145/1553374.1553391, p. 1–59, ISSN 1605585165. 186

Yang, B. et C. Cardie. 2012, «Extracting opinion expressions with semi-markov conditional random fields», dans *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, p. 1335–1345. 66

Yang, B. et C. Cardie. 2013, «Joint Inference for Fine-grained Opinion Extraction», *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,*, p. 1640–1649. 41, 43, 66, 67

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola et E. Hovy. 2016, «Hierarchical Attention Networks for Document Classification», *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, doi :10.18653/v1/N16-1174, p. 1480–1489. 62, 163, 173

Yildirim, S., S. Narayanan et A. Potamianos. 2011, «Detecting emotional state of a child in a conversational computer game», *Computer Speech and Language*, vol. 25, nº 1, doi :10.1016/j.csl.2009.12.004, p. 29–44, ISSN 08852308. 54

Zadeh, A., M. Chen, S. Poria, E. Cambria et L.-P. Morency. 2017, «Tensor Fusion Network for Multimodal Sentiment Analysis», dans *EMNLP*. 45, 52, 72, 73, 109, 111, 112, 187, 195, 197, 198

Zadeh, A., P. P. Liang, N. Mazumder, S. Poria, E. Cambria et L.-P. Morency. 2018a, «Memory Fusion Network for Multi-view Sequential Learning», dans *AAAI*. 52, 64, 73

Zadeh, A., P. P. Liang, S. Poria, E. Cambria et L.-P. Morency. 2018b, «Multimodal Language Analysis in the Wild : CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph», *Proceedings of ACL*, p. 2236–2246. 52, 64, 78, 92, 154

Zadeh, A., P. P. Liang, S. Poria, P. Vij, E. Cambria et L.-p. Morency. 2018c, «Multi-attention Recurrent Network for Human Communication Comprehension», dans *AAAI*. 45, 52, 64, 72, 193

Zadeh, A., L.-P. Morency, P. Pu Liang, S. Poria, E. Cambria et S. Scherer. 2018d, «First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)», dans *Workshop on Human Multimodal Language (Challenge-HML) - ACL*, ISBN 9781948087469. 32, 78

Zadeh, A., R. Zellers, E. Pincus et L.-p. Morency. 2016, «MOSI : Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos», . 78, 107, 108

Zhang, M., Y. Zhang et D. T. Vo. 2015, «Neural Networks for Open Domain Targeted Sentiment», *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, , nº September, doi :10.18653/v1/D15-1073, p. 612–621. 61

Zhou, Y., S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-p. Morency et J. Cassell. 2013, «Multimodal Prediction of Psychological Disorders : Learning Verbal and Nonverbal Commonalities in Adjacency Pairs», *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, p. 160–169. 59, 60

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Analyse d'opinion dans les interactions orales

Mots clés : Informatique Affective, Traitement du Langage Naturel, Analyse d'opinion, Champs aléatoires conditionels cachés

Résumé : La reconnaissance des opinions d'un locuteur dans une interaction orale est une étape cruciale pour améliorer la communication entre un humain et un agent virtuel. Dans cette thèse, nous nous situons dans une problématique de traitement automatique de la parole (TAP) sur les phénomènes d'opinions dans des interactions orales spontanées naturelles. L'analyse d'opinion est une tâche peu souvent abordée en TAP qui se concentrait jusqu'à peu sur les émotions à l'aide du contenu vocal et non verbal. De plus, la plupart des systèmes récents existants n'utilisent pas le contexte interactionnel afin d'analyser les opinions du locuteur. Dans cette thèse, nous nous penchons sur ces sujets. Nous nous situons dans le cadre de la détection automatique en utilisant des modèles d'apprentissage statistique. Après une étude sur la modélisation de la dynamique de l'opinion par un modèle à états latents à l'intérieur d'un monologue, nous étudions la manière d'intégrer le contexte interactionnel dialogique, et enfin d'intégrer l'audio au texte avec différents types de fusion. Nous avons travaillé sur une base de données de Vlogs au niveau d'un sentiment global, puis sur une base de données d'interactions dyadiques multimodales composée de conversations ouvertes, au niveau du tour de parole et de la paire de tours de parole. Pour finir, nous avons fait annoté une base de données en opinion car les base de données existantes n'étaient pas satisfaisantes vis-à-vis de la tâche abordée, et ne permettaient pas une comparaison claire avec d'autres systèmes à l'état de l'art. A l'aube du changement important porté par l'avènement des méthodes neuronales, nous étudions différents types de représentations: les anciennes représentations construites à la main, rigides mais précises, et les nouvelles représentations apprises de manière statistique, générales et sémantiques. Nous étudions différentes segmentations permettant de prendre en compte le caractère asynchrone de la multi-modalité. Dernièrement, nous utilisons un modèle d'apprentissage à états latents qui peut s'adapter à une base de données de taille restreinte, pour la tâche atypique qu'est l'analyse d'opinion, et nous montrons qu'il permet à la fois une adaptation des descripteurs du domaine écrit au domaine oral, et servir de couche d'attention via son pouvoir de clusterisation. La fusion multimodale complexe n'étant pas bien gérée par le classifieur utilisé, et l'audio étant moins impactant sur l'opinion que le texte, nous étudions différentes méthodes de sélection de paramètres pour résoudre ce problème.

Title : Opinion analysis in speech interactions

Keywords : Affective Computing, Natural Language Processing, Opinion Analysis, Hidden Conditional Random Fields

Abstract : Recognizing a speaker's opinions in an oral interaction is a crucial step in improving communication between a human and a virtual agent. In this thesis, we find ourselves in a problematic of automatic speech processing (SP) on opinion phenomena in natural spontaneous oral interactions. Opinion analysis is a task that is not often addressed in SP that focused until recently on emotions using voice and non-verbal content. In addition, most existing systems do not use the interactional context to analyze the speaker's opinions. In this thesis, we focus on these topics. We are in the context of automatic detection using statistical learning models. After a study on the modeling of the dynamics of the opinion by a latent model within a monologue, we study how to integrate the dialogical interactional context, and finally to integrate both audio and text with different types of fusion. We worked on a database of Vlogs at the level of the document, then on a database of multimodal dyadic interactions composed of open conversations, at the levels of speech turns and pair of speech turns. Finally, we annotated a database in opinion because the existing databases were not satisfying with respect to the task addressed, and did not allow a clear comparison with other systems in the state of the art. At the dawn of the important change brought about by the advent of neuronal methods, we study different types of representations: the old representations built by hand, rigid but precise, and new representations learned in a statistical, general and containing semantics. We study different segmentations to take into account the asynchronous nature of multi-modality. Finally, we use a latent state learning model that can adapt to a small database, for the atypical task of opinion analysis, and we show that it allows both an adaptation of the descriptors from the written domain to the oral domain, and serves as an attention layer via its clustering power. Complex multimodal fusion is not well managed by the classifier used, and audio being less impacting on opinion than text, we study different methods of parameter selection to solve this problem.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France