De l'analyse d'opinions à la détection des problèmes d'interactions humain-machine : application à la gestion de la relation client

Irina Poltavchenko

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0030

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Irina POLTAVCHENKO
épouse MASLOWSKI

le 4 juin 2018

De l'analyse d'opinions à la détection des problèmes d'interactions humain-machine :

application à la gestion de la relation client

Directeur de thèse : **Catherine PELACHAUD**
Directeur de thèse : **Chloé CLAVEL**
Co-encadrement de la thèse : **Delphine LAGARDE**

Jury

Mme Sophie ROSSET, Directeur de Recherches, LIMSI, CNRS	Rapporteur
M. Frédéric LANDRAGIN, Directeur de Recherches, LATTICE, CNRS	Rapporteur
M. Kamel SMAÏLI, Professeur, LORIA, l'Université de Lorraine	Examinateur

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

Table des matières

Chapitre 1

Introduction générale

Sommaire

1.1 Problématique de recherche

La problématique de la détection des problèmes d'interaction dans les dialogues humain-machine revient à l'ordre du jour avec la digitalisation omniprésente des entreprises et surtout de la relation client. La définition d'un problème d'interaction varie en fonction du but final de leur détection et du niveau technique du système conversationnel. La définition la plus répandue et initiée par WALKER et collab. [2002] se focalise sur les dysfonctionnements du système, c'est à dire son incapacité à accomplir la tâche demandée par l'utilisateur.

La perception du déroulement de l'interaction par l'utilisateur n'est qu'une approche émergente. La problématique de l'analyse d'opinion en tant que telle est largement étudiée mais reste un sujet débattu. L'augmentation des données sur Internet nourrit les cas d'applications envisageables et imaginables. Toutefois, la plupart des analyses s'arrêtent au niveau de la détection de la polarité positive ou négative. Des analyses plus détaillées restent difficiles à effectuer. Une des problématiques liées à la détection de l'opinion de l'utilisateur est également la détection de la cible de l'opinion. Il est important, par exemple dans notre cas, de distinguer une opinion exprimée envers l'interaction avec le chatbot, de l'opinion exprimée envers les produits ou des services accompagnant ces produits. La majorité des analyses d'opinions, des émotions, de l'affect ou autres phénomènes reliés à l'opinion dans le domaine des systèmes conversationnels automatiques est réalisée pour les systèmes vocaux. Les interactions entre un chatbot et ses utilisateurs restent peu étudiées et encore moins pour le français.

Les interactions humain-agent peuvent être restreintes par le domaine de l'entreprise ou suivre une conversation libre. Suite à la popularité grandissante de "l'intelligence artificielle", et la démocratisation des outils de développement de chatbots, la satisfaction de l'utilisateur final revient au premier plan dans le monde industriel. Dans le monde académique les travaux continuent majoritairement à chercher les meilleures méthodes d'exploitation des données "objectives" (c'est-à-dire des données des logs) provenant du système du chatbot.

Par ailleurs, le contexte classique des analyses en TAL est un texte bien formé, contenant des mots aux normes dictionnairiques. Les travaux de recherche sur la détection des opinions et des émotions sur ce type de texte sont bien développés. Les spécificités des écrits de tchat sont souvent considérées comme du bruit, or elles peuvent également être porteuses d'informations sur l'état émotionnel de l'utilisateur. Une analyse détaillée est nécessaire pour décider de l'importance de chaque spécificité.

1.2 Contexte de réalisation des travaux de thèse

Cette thèse est réalisée dans le cadre d'un dispositif CIFRE reliant par une convention le laboratoire de Télécom ParisTech LTCI et le centre de recherche de l'entreprise EDF : EDF Lab Paris-Saclay. Ce partenariat nous permet de bénéficier des compétences théoriques de la communauté de l'agent conversationnel virtuel et des connaissances métier de l'application de l'entreprise.

Cette thèse est commanditée par le département commerce et en lien étroit avec l'équipe du marketing digital. Compte tenu du nombre d'appels téléphoniques croissant vers le centre de téléconseillers de l'entreprise, le déport des contacts vers le canal Internet est un besoin réel. Le conseiller virtuel doit répondre aux questions basiques des utilisateurs pour permettre aux conseillers humains de se concentrer sur les problèmes complexes. Les échanges avec les commanditaires nous ont permis d'avoir une vision

réelle de l'application de nos travaux ce qui est un aspect très stimulant de ce sujet de recherche. En effet, définir les types de phénomènes à détecter à partir du besoin métier permet d'assurer la suite applicative des présents travaux.

Cette thèse s'inscrit également dans la continuité des travaux réalisés à EDF R&D sur les appels téléphoniques et les réclamations clients. Ainsi, des travaux de recherche à EDF Lab ont déjà abordé le sujet de l'analyse des opinions et des sentiments avec l'utilisation de méthodes de traitement automatique des langues (TAL). LAVALLEY et collab. [2010] ont proposé une méthode d'extraction automatique de chaînes de mots relatifs aux opinions à partir d'un corpus étiqueté. En même temps, KUZNICK et collab. [2010] ont démontré l'efficacité de la modélisation des concepts métier dans la classification des enquêtes de satisfaction des clients. Ces travaux rejoignent l'idée de réaliser l'analyse de l'opinion et de sa cible. Le travail effectué par CAILLIAU et CAVET [2010], proposant une modélisation du déroulement d'une conversation téléphonique au sein d'un centre d'appel, est l'un des premiers pas vers la détection des problèmes d'interaction dans les conversations entre un client et un conseiller humain. Une chaîne de traitement performante a ensuite été développée pour la détection des opinions et de leurs cibles dans les conversations d'un centre d'appel par CLAVEL et collab. [2013].
Un outil a été également développé pour naviguer dans les questions posées à un agent virtuel SUIGNARD [2010]. Ces travaux ont servi de point de départ pour la recherche sur les problèmes d'interaction dans les dialogues de tchat humain-conseillère virtuelle.

1.3 Axes de recherche et positionnement

L'axe principal de notre recherche est l'étude des problèmes d'interaction humain-agent du point de vue de l'opinion de l'utilisateur. Par conséquent, les énoncés utilisateur dans le contexte du déroulement du dialogue doivent nous fournir les principaux indices d'un problème éventuel.

L'écrit des utilisateurs dans leurs échanges avec une conseillère virtuelle est produit librement et de manière spontanée. Nous choisissons d'utiliser des caractéristiques de "haut" niveau, par opposition à l'approche à base de n-grammes ou d'un sac de mots, pour détecter l'opinion et des phénomènes reliés aux opinions chez l'utilisateur. Ce positionnement implique une étude du langage utilisateur afin de le situer par rapport au langage des utilisateurs en ligne, en général.

Les spécificités scripturales peuvent témoigner d'une émotion de l'utilisateur. La manière dont le dialogue se poursuit peut révéler de manière implicite que l'utilisateur n'a pas obtenu la réponse souhaitée. Nous cherchons à prendre en compte tout comportement de l'utilisateur révélant l'existence d'un problème d'interaction à travers son texte.

1.4 Apports des travaux de thèse

Les travaux de cette thèse ont contribué à l'étude des problèmes d'interaction humain-machine de la manière suivante :

Analyse de corpus Nous caractérisons la production écrite spontanée de l'utilisateur en la comparant avec d'autres corpus humain-agent et avec les caractéristiques des écrits sur Internet.

Approche hybride Nous proposons un système DAPI (**D**étection **A**utomatique des **P**roblèmes d'**I**nteraction) pour la détection des problèmes d'interaction. L'architecture du système est hybride. Les règles linguistiques prennent en compte les spécificités du langage utilisateur, les répétitions et les reformulations de l'utilisateur et l'historique du dialogue. La détection des reformulations de l'utilisateur à base des distances linguistiques est complétée par la représentation sémantique des mots apprise par une méthode non-supervisée.

Typologie des problèmes d'interaction Nous proposons une typologie des problèmes d'interaction axée sur l'opinion de l'utilisateur.

Méthodologie de l'annotation manuelle Nous proposons un schéma d'annotation des problèmes d'interaction selon notre typologie et en forme de l'arbre de décision.

1.5 Organisation du manuscrit

Ce document est organisé en trois parties. La première partie introduit le sujet de la thèse. Elle indique le positionnement de nos travaux par rapport aux travaux de l'état-de-l'art. La deuxième partie est consacrée à l'étude très importante du corpus et l'annotation du corpus de référence en problèmes d'interaction et en opinions et en phénomènes reliés aux opinions envers les produits et services. La troisième partie présente le développement du système d'annotation des problèmes d'interaction et les résultats obtenus. La quatrième et la dernière partie conclut notre travail en rappelant les points principaux à retenir. Elle donne également un aperçu des perspectives de notre recherche.

Partie I : État de l'art : détection des problèmes d'interaction Le chapitre 2 introduit des définitions de problèmes d'interaction, de l'opinion et des phénomènes reliés aux opinions. Il permet d'appréhender la grande diversité des typologies existantes. Les critères de choix d'une typologie de l'opinion et des phénomènes reliés aux opinions sont présentés en fonction du contexte industriel.

Le chapitre 3 donne une analyse des approches existantes dans la littérature pour la détection des problèmes d'interaction, des opinions et des phénomènes reliés aux opinions. Les indices utilisés pour la détection de ces phénomènes varient en fonction de l'approche choisie et des moyens d'analyse accessibles. De plus, l'utilité des prétraitements du texte, tel que l'annotation des catégories lexicales des mots et la correction d'orthographe, est discutée.

Partie II : Corpus, annotation et stratégie d'annotation Le chapitre 4 positionne notre corpus d'interactions écrites humain-agent par rapport aux corpus humain-humain et humain-agent décrits dans la littérature. Le corpus de l'étude est caractérisé selon son intérêt linguistique pour la tâche de détection des problèmes d'interaction. Les défis potentiels pour le traitement automatique du texte produit dans les conditions "in-the-wild" [1] sont également exposés.

Le chapitre 5 présente la taxonomie des problèmes d'interaction que nous proposons. Ensuite, il explique ce qui a guidé nos choix lors de la constitution du guide d'annotation.

1. terme utilisé par SCHULLER et collab. [2016] pour désigner les données collectés non pas dans les conditions d'un laboratoire mais dans un cadre applicatif réel

Partie III : Détection automatique des problèmes d'interaction Le chapitre 6 décrit le développement du système DAPI de détection des problèmes d'interaction. Les choix méthodologiques sont complétés par l'expérimentation.

Le chapitre 7 illustre les résultats obtenus lors de l'évaluation du système. Ils permettent de valider l'approche choisie. Les résultats de l'annotation automatique, ainsi que manuelle sont discutés. Les conclusions s'ouvrent sur les expérimentations supplémentaires décrites dans le chapitre 7.4.

Partie IV : Conclusion et perspectives En conclusion (chapitre 8), nous passons en revue les principaux points de cette thèse dont l'état-de-l'art, de notre approche, nos contributions et leurs perspectives.

Première partie

Partie 1 : État de l’art : détection des problèmes d’interaction

Résumé

Les études des problèmes d'interaction pour les chatbots sont encore rares. Dans cet état-de-l'art, nous considérons également les études pour les systèmes vocaux de dialogue humain-machine. La tendance principale dans ces travaux est de baser la définition des problèmes d'interaction sur le dysfonctionnement du système de l'agent. Nous rejoignons une tendance récente définissant un problème d'interaction du point de vue de l'utilisateur et d'annoter l'énoncé utilisateur. Les typologies existantes des problèmes d'interaction sont axées sur un évènement marquant, sur le niveau de compréhension mutuelle entre les participants du dialogue ou encore sur le principe de coopération. Nous proposerons une typologie axée sur l'opinion de l'utilisateur.

Les psychologues aussi bien que des linguistes développent des théories d'opinions et de phénomènes reliés aux opinions (OPRO) tels que les émotions, les attitudes, les appréciations, etc. Nous choisissons un modèle d'OPRO sans restreindre les phénomènes dont il est composé à une seule théorie ou un seul auteur. Un mélange de modèles catégoriels, dimensionnels et d'évaluation nous parait plus pertinent dans notre recherche. Les modèles catégoriels conviennent bien à l'analyse des textes car ils permettent d'attribuer une étiquette. Les modèles dimensionnels permettent d'évaluer l'intensité et la valence positive ou négative de l'émotion. La théorie d'évaluation permet l'identification de la source et de la cible des OPROs. Nous choisissons un modèle permettant de rechercher une OPRO, sa source et sa cible. Nous choisissons de détecter les opinions négatives des utilisateurs comme une classe générique et de détecter des émotions précises au sein de l'opinion lorsque cela est possible. Nous proposons également une liste de six émotions qui nous semblent pertinentes dans le cadre de notre recherche.

La répétition ou la reformulation, l'historique du dialogue et l'OPRO de l'utilisateur sont les indices les plus importants pour la détection des problèmes d'interaction. L'approche à base d'apprentissage automatique supervisé est la plus utilisée pour la détection des problèmes d'interaction et des OPRO, particulièrement pour l'anglais. Les méthodes hybrides sont encore peu répandues mais elles permettent de proposer des solutions pour des langues ayant peu de ressources linguistiques développées comme le français.
Nous choisissons l'approche hybride combinant une approche à base de règles et l'approche non-supervisée d'apprentissage des représentations des mots. Cette approche présente l'avantage de s'affranchir de l'annotation humaine d'un grand corpus de données requise pour les méthodes d'apprentissage supervisé. L'apprentissage de représentation permet de tirer parti, de manière non-supervisée, de la richesse du corpus collecté et cherche à modéliser ses spécificités langagières.

Chapitre 2

État de l'art. Définitions et typologies

Sommaire

Ce chapitre présente notre positionnement vis-à-vis des définitions et des typologies choisies pour la détection des problèmes d'interaction dans les interactions écrites humain-agent. Nous faisons ici un tour d'horizon des études des problèmes d'interaction. Compte tenu de la variété des terminologies et de la difficulté de distinguer parfois les opinions d'autres phénomènes tels que les émotions et les appréciations, nous avons choisi de rassembler l'ensemble de ces phénomènes autour du terme **opinions et phénomènes reliés aux opinions ou OPEM**, à partir de OPinion et EMotion, les phénomènes les plus représentés de cet ensemble. Le terme OPEM est synonymique au terme OPRO (opinions et des phénomènes reliés aux opinions).

2.1 Définition et typologies des problèmes d'interaction

La revue des définitions et des typologies des problèmes d'interactions dans les deux sous-sections suivantes nous permet de nous positionner sur la question de la détection des problèmes d'interaction. Il est à noter que nous considérons les études existantes pour les systèmes de dialogue humain-machine aussi bien vocaux que textuels. Les problèmes d'interaction ont essentiellement été abordés dans la communauté homme-machine pour les systèmes de dialogue vocal. Ce sujet est encore rarement abordé pour les chatbots.

2.1.1 Terminologie et définitions des problèmes d'interaction

LANGKILDE et collab. [1999] ont été parmi les premiers chercheurs à s'intéresser à la prédiction des dialogues problématiques. Ils ont défini le premier périmètre de la problématique : l'identification des raisons de l'échec du système à répondre aux demandes de l'utilisateur. Depuis, le courant principal des recherches sur cette thématique a été d'identifier les composants des systèmes vocaux à l'origine de l'apparition des problèmes d'interaction [GEORGILADAKIS et collab., 2016]. Les problèmes d'interactions sont alors souvent définis à travers des cas notables tels que :

— l'interruption du dialogue par l'utilisateur [LANGKILDE et collab., 1999], appelée également "breakdown" [MARTINOVSKY et TRAUM, 2006]

— l'intervention d'un agent humain [LANGKILDE et collab., 1999] ou le transfert de l'utilisateur vers un agent humain suite à sa demande [BEAVER et FREEMAN, 2016]

— l'échec du système à accomplir une tâche [LANGKILDE et collab., 1999]

WALKER et collab. [2002] donne une définition plus concise des dialogues problématiques : "Les dialogues dans lesquels le système (How May I Help You) n'a pas réussi à terminer la tâche donnée par l'utilisateur sont appelés problématiques". [1]

Par la suite le développement technique des systèmes dialoguant a permis d'affiner la définition d'un problème d'interaction et la prise en compte de la satisfaction client comme l'un des éléments constitutifs des problèmes d'interaction [HASTIE et collab., 2002]. Cette prise en compte s'appuie notamment sur la détection des émotions négatives de l'utilisateur dans ses échanges vocaux [LISCOMBE et collab., 2005; ANG et collab., 2002].

Un aspect différent de la prise en compte de l'utilisateur dans les échecs des communications avec un système vocal a été ensuite proposé par [MÖLLER et collab., 2007]. Les auteurs définissent les problèmes d'interaction comme le résultat d'un déséquilibre entre

1. "Dialogues in which the HMIHY system did not successfully complete the caller's task are referred to as PROBLEMATIC."

les modèles de l'utilisateur et du système utilisés lors de la conception du système et la vision du système par l'utilisateur ; autrement dit des "fausses idées" ("misconceptions"). Cette perception des problèmes d'interaction permet de se détacher des composants du système qui étaient au cœur des recherches précédentes. Ils proposent un modèle "réaliste" de l'utilisateur pour le futur développement des systèmes vocaux dialoguant.

MÖLLER et collab. [2007] présentent les aprioris de l'utilisateur sur le fonctionnement du système, par exemple la capacité du système à effectuer une tache qui ne fait pas partie de ses compétences. Tout en étant un aspect important à prendre en compte lors du développement d'un système conversationnel, un autre aspect de la vision du système par l'utilisateur peut être envisagé. Cet aspect est proposé par XIANG et collab. [2014] dans sa définition d'une situation problématique : "Les situations problématiques reflètent le fait que l'utilisateur humain n'est pas satisfait par les réponses proposées par un système conversationnel" [2]. Cette définition permet d'évaluer les performances du système du point de vue du résultat attendu par l'utilisateur et non du point de vue du processus de la prise en main. Les types génériques des problèmes d'interaction classés suivant l'angle du processus de la prise en main peuvent d'autant plus se traduire par des motifs différents d'actions utilisateur en fonction du système utilisé [MÖLLER et collab., 2007]. Nous sommes donc amenée à une revue des typologies existantes.

2.1.2 Typologies existantes

Les typologies des problèmes d'interaction développent les axes indiqués par les définitions des problèmes d'interaction. La typologie des problèmes d'interaction proposée par LANGKILDE et collab. [1999] : (1) "l'utilisateur raccroche le téléphone", (2) "l'agent humain reprend l'appel" et (3) "le système croit qu'il a bien réalisé la tâche, alors que la tâche n'a pas été accomplie avec succès". Cette typologie a été exploitée également par WALKER et collab. [2002].

HIRST et collab. [1994] ont modélisé des problèmes d'interaction entre les humains pour les appliquer ensuite aux machines. Ils définissent les types suivants de problèmes d'interaction :

- **incompréhension**, lorsque l'un des interlocuteurs n'est pas capable d'interpréter l'énoncé de l'autre ;
- **malentendu**, lorsque l'un des interlocuteurs ne se rend pas compte que son interprétation de l'énoncé de l'autre ne correspond pas au message prévu par l'émetteur. Les malentendus peuvent être à leur tour de deux types :
  - **un malentendu manqué**, lorsque les interlocuteurs ne remarquent pas l'existence d'un malentendu, et
  - **un malentendu remarqué**, lorsqu'au moins un des interlocuteurs se rend compte de l'existence d'un malentendu.
    Lorsque l'interlocuteur se rend compte de l'existence d'un malentendu, HIRST et collab. [1994] utilisent deux sous-classes :
    - les **malentendus auto-détectés** ("self-misunderstandings"), lorsqu'un malentendu est produit et détecté par le même participant d'un dialogue ;
    - les **malentendus détectés par autrui** ("other-misunderstandings"), lorsqu'un malentendu chez un participant est détecté par un autre participant.

2. "*Problematic situations* reflect that a human user is not satisfied with answers that a conversational system offers."

HIRST et collab. [1994] excluent les conceptions erronées de leur étude car elles font partie des connaissances préalables à la conversation. Afin de représenter tous les types de manque de compréhension menant aux problèmes d'interaction, ce qui correspond au deuxième angle, nous pouvons donc considérer la typologie des **conceptions erronées** de MÖLLER et collab. [2007] comme un complément à la typologie de HIRST et collab. [1994]. En fonction des conceptions erronées existant chez l'utilisateur, MÖLLER et collab. [2007] distinguent cinq niveau d'erreurs :

1. le but : l'utilisateur estime que le système peut accomplir une tâche dont il n'est pas capable ;
2. la tâche : l'utilisateur ne comprend pas comment atteindre son objectif d'interaction avec le système.
3. la formulation de la commande : l'utilisateur s'exprime avec un vocabulaire ou une grammaire incompréhensibles pour le système ;
4. le concept : la modélisation du système est basée sur une vision différente du "monde" de celle de l'utilisateur ;
5. la reconnaissance (de la voix ou du texte) : ce type d'erreurs n'est pas contrôlable par l'utilisateur.

Ils remarquent qu'il existe également d'autres erreurs qu'ils n'ont pas pu attribuer à un niveau spécifique.

La typologie de MÖLLER et collab. [2007] s'intéresse à l'absence de coopération de la part de l'utilisateur. DYBKJÆR et collab. [1996]; BERNSEN et collab. [1996] représentent les problèmes d'interaction comme des cas de violation des règles ou des "principes" d'un "dialogue coopératif". Pour identifier les principes de coopération pour un dialogue humain-machine, ils prennent, comme point de départ, les quatre maximes de GRICE et collab. [1975], proposées pour un dialogue entre deux humains :

1. maxime de qualité : contribuer à la conversation autant d'information que nécessaire ;
2. maxime de quantité : les contributions à la conversation doivent être pertinentes ;
3. maxime de relation : les contributions d'information doivent être fournies dans le bon ordre ;
4. maxime de manière : le partenaire de conversation doit préciser de quoi il a l'intention de parler.

Treize principes généraux découlent de ces quatre maximes :

1. La contribution doit être suffisamment informative ;
2. La contribution ne doit pas être surchargée en information ;
3. Ne pas dire de choses que vous croyez être fausses ;
4. Ne pas parler de choses que vous ne pouvez pas prouver ;
5. Être pertinent ;
6. Éviter l'imprécision ;
7. Éviter l'ambiguïté ;
8. Être bref ;
9. Avoir un discours ordonné ;
10. Communiquer aux partenaires de dialogue les éléments dont ils ont besoin pour être coopératif dans le dialogue ;

11. Prendre en compte les connaissances dont dispose le partenaire ;
12. Prendre en compte les attentes du partenaire vis-à-vis vos propres connaissances ;
13. Lancer l'initiative de clarification de l'information si l'interaction n'a pas abouti.

Les maximes ou principes correspondent à plusieurs des sept aspects du dialogue : le caractère informatif, la vérité et l'évidence, la pertinence, la manière de communication, l'asymétrie des partenaires, les connaissances préalables au dialogue et la "réparation" et la clarification du dialogue. Les auteurs ont complété ces principes génériques par onze principes spécifiques au dialogue humain-machine. Un de ces onze principes est de fournir une communication claire et compréhensible de ce que le système peut et ne peut pas faire. Ce principe correspond à l'aspect du dialogue "l'asymétrie des partenaires". Un autre principe est la différentiation, lorsque cela est possible, des besoins des utilisateurs novices et experts (pour permettre au dialogue de s'adapter à l'utilisateur). Ce principe fait partie de l'aspect "les connaissances préalables". Pour clarifier ou réparer un dialogue humain-machine, les auteurs proposent de s'appuyer sur un méta-langage, tel que des mots clés, permettant de réinitialiser la demande de l'utilisateur.

DYBKJÆR et collab. [1996] soulignent que sept de leurs principes, représentant trois aspects de dialogue, ne peuvent pas être réduits aux maximes de Grice. Les trois aspects sont les suivants : asymétrie entre les partenaires, les connaissances préalables et la métacommunication. Les deux premiers aspects ne sont pas présents dans un dialogue idéal entre deux humains et le dernier aspect est spécifique aux dialogues humain-machine.

En revanche, HIGASHINAKA et collab. [2015a]; HORII et collab. [2017] ont pris les maximes de Grice comme point de départ pour construire leur taxonomie d'erreurs qui peuvent mener à l'interruption ("breakdown") d'un dialogue avec un chatbot. Puisque l'interruption d'un dialogue survient souvent suite à la présence d'autres types de problèmes d'interaction, nous considérons que les erreurs amenant au "breakdown" sont des problèmes d'interaction et les catégories principales de leur taxonomie représentent la taxonomie des problèmes d'interaction. La taxonomie en question contient quatre catégories, en fonction du niveau de contexte où l'erreur se situe : le niveau de l'énoncé (du système), le niveau d'une paire adjacente, le contexte précédant la paire adjacente et l'environnement, c'est-à-dire lorsque la cause de l'erreur n'est pas liée au contexte local (par exemple, l'absence de connaissances sur le monde réel).

Selon HIGASHINAKA et collab. [2015a], les erreurs au niveau de l'énoncé proviennent le plus souvent des problèmes de génération des énoncés par le système. Les problèmes au niveaux d'une paire adjacente sont liés aux maximes de Grice : excès ou manque d'information, incompréhension, manque de pertinence, l'absence de clarté de lien entre l'énoncé de l'utilisateur et la réponse du système et une mauvaise compréhension. Le contexte précédant la paire adjacente peut révéler des erreurs de communication qui correspondent également aux maximes de Grice, par exemple, la manque de nouvelle information ou son excès, contradictions, les réponses impertinentes de la part du système, l'absence de lien dans la réponse du système vis-àvis les paires adjacentes précédente (example des auteurs : l'utilisateur parle du style de surf et le système parle du style des vêtements), le système n'a pas détecté le changement de thème chez l'utilisateur. Les erreurs au niveau de l'environnement sont les cas lorsque le système fait des confirmations non-fondées (ex. "tous le poisson est contaminé", hors ce n'est pas forcement le cas), ses affirmations manque de bon sens ("c'est bien d'être malade"), ou encore lorsque les énoncés du système manque de politesse.

La catégorisation des problèmes d'interaction dans les dialogues avec un chatbot par XIANG et collab. [2014] propose de différencier deux cas :

— le problème d'interaction est lié au sentiment d'insatisfaction de l'utilisateur vis à vis de l'interaction ;

— le problème d'interaction est lié à l'intention de l'utilisateur (une action de l'utilisateur telle qu'une question ou une demande).

Ils définissent quatre catégories d'intention : changement de thème, reformulation, suite (ex. une question plus détaillée que la question précédente), clarification. Les catégories de sentiments sont les suivantes : les salutations (qui peuvent être exprimées de façon intime, par exemple, "Bonjour mon cœur"), les commentaires positifs ou négatifs sur la réponse, les exclamations ou déclarations dont la cible n'est pas le chatbot, les expressions grossières, les ordres ou les énoncés différents des cas énumérés ci-dessus et contenant de la ponctuation. Ce nouvel axe place la satisfaction de l'utilisateur en avant et la prend en compte dans sa communication naturelle.

2.1.3 La portée du problème à détecter : la granularité

Les travaux portant sur la détection des problèmes d'interaction dans les dialogues humain-machine définissent la portée du problème à détecter en fonction des objectifs (par example, l'objectif peut être d'améliorer une des module du système) et des moyens techniques disponibles (de nos jours, de nouveaux algorithmes et des techniques d'analyse de données sont disponibles, ce qui n'était pas le cas il y a quelques années). WALKER et collab. [2002]; HASTIE et collab. [2002] se positionnent au niveau du dialogue car ils considèrent qu'il n'est pas possible de juger si le dialogue a échoué tant qu'il n'est pas fini. Pour détecter des problèmes d'interaction dans un dialogue en cours, VAN DEN BOSCH et collab. [2001] se concentrent sur l'énoncé de l'utilisateur soit pour prédire que cet énoncé serait la source du problème, soit pour trouver dans l'énoncé des indices d'un problème créé par l'énoncé précédent. XIANG et collab. [2014] analysent également l'énoncé de l'utilisateur pour comprendre si la réponse du système était problématique.

SCHMITT et collab. [2011] proposent d'évaluer la qualité de l'interaction au niveau d'une paire adjacente en soulignant le haut niveau de subjectivité de l'utilisateur. Afin d'identifier le module du système à améliorer, CHAI et collab. [2006]; MEENA et collab. [2015] proposent de détecter des énoncés problématiques du système lorsque le système a des difficultés à effectuer une compréhension correcte du langage parlé ou écrit.

2.1.4 Notre positionnement : le point de vue de l'utilisateur

Suite à cette revue des typologies et définitions, le travail de XIANG et collab. [2014] nous semble particulièrement pertinent pour notre contexte. En effet, les travaux évoqués dans la section 2.1.2, s'appuient sur l'analyse des événements, survenant lors de l'interaction et provoqués par le système. Nous nous positionnons du point de vue de l'utilisateur. Prenons la répétition de l'utilisateur d'un tour de parole à l'autre en tant qu'exemple. Elle peut être considérée comme un indice d'un malentendu, selon le deuxième angle. En revanche, nous la considèrerons comme un indice potentiel de l'insatisfaction de l'utilisateur, en accord avec le travail de XIANG et collab. [2014]. Notre positionnement est étroitement lié à la nature des dialogues entre les utilisateurs et la conseillère virtuelle. Un chatbot sur un site web de l'entreprise constitue un service et par conséquent, les utilisateurs s'attendent à un service de qualité et donc à ce que le chatbot comprenne leurs demandes.

En accord avec la définition de XIANG et collab. [2014] des situations problématiques, nous définissons les problèmes d'interaction comme le reflet de l'insatisfaction de l'uti-

lisateur de son interaction avec un chatbot. Puisque notre objectif est de détecter si l'utilisateur est satisfait par l'interaction afin de pouvoir ensuite modifier le comportement de l'agent, nous annotons l'énoncé de l'utilisateur quand il témoigne qu'un problème a eu lieu. Nous établissons également notre propre typologie des problèmes d'interaction entièrement basée sur la détection des signes de la probable insatisfaction de l'utilisateur lors de son interaction avec l'agent conversationnel, décrit dans la Section 5.1 page 68.

2.2 Définition et typologies des opinions et des phénomènes reliés aux opinions

Les auteurs dans la littérature utilisent une grande variété de termes tels que l'opinion, l'attitude, le sentiment, l'émotion, l'affect, l'humeur, l'appréciation, la position sociale ("social stance"), qui décrivent des phénomènes psychologiques avec des spécificités propres à chaque terme mais aussi des caractéristiques communes. En général, les définitions les plus développées se trouvent dans les travaux du domaine de la psychologie. Selon l'analyse présentée par CLAVEL et CALLEJAS [2016] des travaux issus de communautés de l'opinion mining et de l'agent conversationnel, la communauté de l'opinion mining emploie ces termes de façon moins stricte et ne précise pas toujours leur définition. La communauté de l'agent conversationnel, en revanche, reprend les définitions et les modèles définis par la psychologie. Notre travail combine une analyse linguistique des dialogues afin de détecter l'opinion de l'utilisateur et le fait que l'un des interlocuteurs soit un agent conversationnel. Il se situe donc au croisement des communautés de l'opinion mining et des agents conversationnels. Dans ce qui suit, nous donnons des définitions, développées aussi bien par des linguistes que par des psychologues, sur lesquelles s'appuie notre recherche. Ensuite nous présentons brièvement les typologies principales des émotions.

2.2.1 Définitions des opinions et des phénomènes reliés aux opinions

Notre étude se plaçant à cheval entre les deux communautés de l'opinion mining et les agents conversationnels, puisque nous analysons des textes produits dans les interactions humain-agent, nous choisissons de bien définir les phénomènes que nous cherchons à identifier sans pour autant restreindre le choix de nos définitions à une seule théorie ou un auteur. Ce choix s'explique par le fait que certaines définitions psychologiques sont moins applicables aux textes que d'autres.

Puisque nous cherchons à détecter des problèmes d'interaction à travers le point de vue de l'utilisateur, nous présentons ici les définitions qui nous semblent les plus pertinentes avec notre but. Pour le domaine de la fouille de texte, MUNEZERO et collab. [2014] définissent l'**opinion** comme une interprétation personnelle de l'information pouvant inclure ou non une émotion. A part l'opinion, MUNEZERO et collab. [2014] distinguent également des affects, des émotions et des sentiments. Les affects étant une réaction inconsciente, ils n'en trouvent pas d'expression dans le texte, selon MUNEZERO et collab. [2014]. Les sentiments sont des phénomènes plus longs dans la durée par rapport aux émotions. Les exemples des sentiments données dans l'article sont l'amour romantique ou l'amour des parents envers leurs enfants, l'amitié ou encore la peine. En ce qui concerne l'émotion, MUNEZERO et collab. [2014] la définissent comme un phénomène complexe et portant l'empreinte de la culture, se référant à l'étude des émotions à travers les cultures de WIERZBICKA [1999b]. Afin de donner une idée plus claire de la nature du phénomène

de l'émotion, nous complétons la définition de MUNEZERO et collab. [2014] par celle de SCHERER [2005] qui dit que **l'émotion** est un épisode de changements dans l'organisme en réponse à l'évaluation d'un événement stimulant externe ou interne, comme dans le cas de la peur, la honte ou la fierté. SCHERER [2005] distingue également l'émotion des phénomènes tels que les préférences, l'attitude, le positionnement interpersonnel et l'humeur. Un des éléments constituant l'émotion pour MUNEZERO et collab. [2014] et SCHERER [2005], sont les "*feelings*". Nous conservons ici ce terme anglais pour faire la différence avec le terme en anglais "sentiment" puisque tous les deux peuvent se traduire par "sentiment" en français.

Malgré la divergence des termes utilisés, nous considérons que la théorie du langage de l'évaluation de MARTIN et WHITE [2005] peut-être fusionnée avec les définitions précédentes de l'opinion et de l'émotion. MARTIN et WHITE [2005] définissent des "feelings" comme des "attitudes" composées des trois dimensions : l'affect, le jugement et l'appréciation. Leur définition de l'affect est l'émotion et une réaction à un comportement. Leurs définitions du jugement et de l'appréciation sont des éléments de l'évaluation et, par conséquent, correspondent à l'opinion. L'avantage de cette théorie linguistique est son modèle de l'opinion que nous présenterons dans la section 2.2.2.

2.2.2 Les typologies des opinions et des phénomènes reliés aux opinions

Suite au choix des définitions de MUNEZERO et collab. [2014] et MARTIN et WHITE [2005], présentées dans la section précédente et où l'opinion et l'émotion sont étroitement liés, nous nous concentrons ici sur les modèles des opinions et des émotions. Ces modèles peuvent être rassemblés en trois grands groupes : les modèles catégoriels, dimensionnels et les modèles établis sur la théorie de l'évaluation.

Les modèles catégoriels et dimensionnels

Les modèles catégoriels et dimensionnels ont été développés pour la psychologie. Nous les décrivons ensemble, car les modèles dimensionnels prennent leurs racines dans les modèles catégoriels. Les deux types de modèles ne tiennent pas compte des aspects cognitifs, d'où leur principale différence avec la théorie de l'évaluation.

Puisque nous cherchons à détecter des problèmes d'interaction, nous sommes intéressée par la détection des émotions négatives et également par leur intensité. RUSSELL [1980] a proposé un modèle bi-dimensionnel d'excitation ("arousal") et de valence (voir la figure 2.1 page 19).

La valence correspond à un axe de plaisir/détresse permettant décrire les émotions comme positives ou négative. L'excitation appelée ensuite par OSGOOD et collab. [1975] "activation", permet de caractériser l'intensité de l'émotion exprimée.

Afin d'aller au niveau plus fin des phénomènes détectés, nous nous intéressons à l'un des *modèles catégoriels* les plus connus, le modèle discret de EKMAN et collab. [1972]. Il a proposé une liste de six émotions universelles détectées dans des expressions faciales : colère, peur, tristesse, joie, dégoût, surprise. L'idée d'attribution d'étiquettes à un phénomène détecté convient bien à l'analyse de texte. En revanche, les émotions telles que la peur, correspondant plutôt aux émotions primaires, sont trop fortes pour le contexte de tchat avec un agent virtuel.

Le modèle multidimensionnel de PLUTCHIK [1980] permet de représenter des "demi-teintes" d'émotions. Ce modèle est aussi appelé "la roue de Plutchik". (voir la figure 2.2 page 19)

FIGURE 2.1 – Un modèle circulaire de l'affecte de RUSSELL [1980]

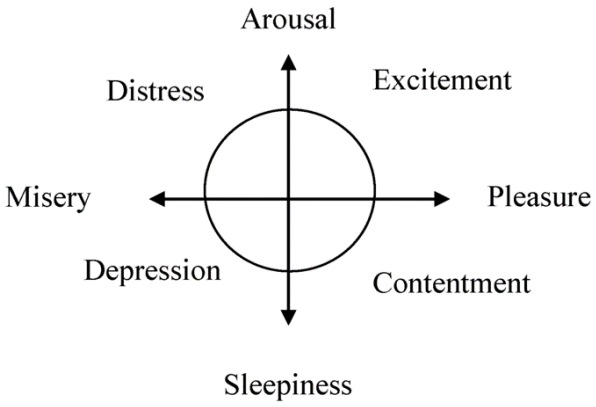
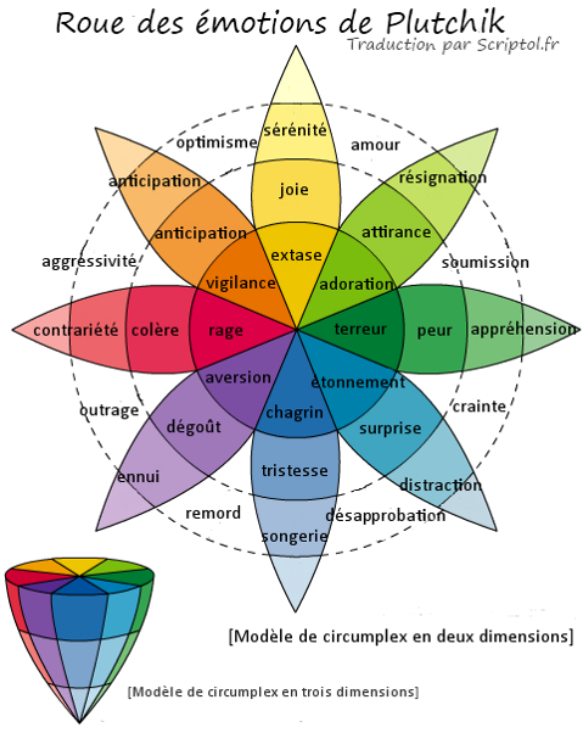


FIGURE 2.2 – La roue des émotions de Plutchik [PLUTCHIK, 1980]



Dans les travaux qui suivent, les listes des émotions principales s'allongent [COWIE et CORNELIUS, 2003], ce qui complexifie la tâche du choix des termes d'émotions. Le tableau récapitulatif (voir le tableau 2.3 page 20) proposé par COWIE et CORNELIUS [2003] des catégories d'émotions de LAZARUS [1999], EKMAN [1999], HAVILAND et LEWIS [1993] et de BANSE et SCHERER [1996], d'états affectifs de BUCK [1999] et d'états liés aux émotions (comme "confiant, ennuyé") de COWIE et CORNELIUS [2003], montre des catégories les plus représentatives de "la vie émotionnelle", selon les auteurs. Ces listes d'émotions peuvent servir de base lors du choix des émotions pertinente pour la détection d'émotions des utilisateurs dans le texte des dialogues.

Le travail de la psychologue canadienne LARIVEY [2002] est intéressant pour nous car elle lie les émotions simples positives avec la satisfaction et les émotions simples néga-

FIGURE 2.3 – La liste des catégories principales des émotions et des états reliés aux émotions de COWIE et CORNELIUS [2003]

Recent lists of key emotions

Lazarus (1999a)	Ekman (1999)	Buck (1999)	Lewis and Haviland (1993)	Banse and Scherer (1996)	Cowie et al. (1999b)
Anger	Anger	Anger	Anger/hostility	Rage/hot anger Irritation/cold anger	Angry
Fright	Fear	Fear	Fear	Fear/terror	Afraid
Sadness	Sadness/distress	Sadness	Sadness	Sadness/dejection Grief/desperation	Sad
Anxiety		Anxiety	Anxiety	Worry/anxiety	Worried
Happiness	Sensory pleasure	Happiness	Happiness	Happiness Elation (joy)	Happy
	Amusement Satisfaction Contentment		Humour		Amused Pleased Content
		Interested Curious Surprised			Interested
	Excitement				Excited
		Bored		Boredom/indifference	Bored Relaxed
		Burnt out			
Disgust	Disgust Contempt	Disgust Scorn	Disgust	Disgust Contempt/scorn	
Pride	Pride	Pride Arrogance	Pride		
Jealousy Envy		Jealousy Envy			
Shame Guilt	Shame Guilt Embarrassment	Shame Guilt	Shame Guilt Embarrassment	Shame/guilt	
					Disappointed
Relief Hope	Relief				
					Confident
Gratitude Love			Love		Loving Affectionate
Compassion		Pity Moral rapture Moral indignation			
Aesthetic					

tives avec l'insatisfaction. De plus, WIERZBICKA [1999a] a démontré que la perception des émotions peut varier en fonction de la langue qui nous fournit les moyens d'expression verbale. C'est pourquoi nous arrêtons nos choix sur les définitions de termes d'émotions de LARIVEY [2002] car ils sont donnés en français. Les définitions que nous citons dans la section 2.2.4, page 23, sont également disponibles sur le site web des "psychologues humanistes" [3].

La théorie de l'évaluation

Une des problématiques du sujet de notre recherche est de distinguer les OPEM (OPinion + EMotions) de l'utilisateur envers l'interaction des OPEM exprimées par rapport aux produits, par exemple. ORTONY et collab. [1990] ont proposé une théorie, appelée dans la littérature OCC (selon les initiales des auteurs : Ortony, Clore, Collins), basée sur la psychologie cognitive dont le but est de créer un modèle d'émotion utilisable pour la création d'une intelligence artificielle. La théorie est adoptée par la communauté humain-agent

3. Les "psychologues humanistes" "considèrent chaque personne comme unique" http://www.redpsy.com/guide/

qui l'intègre dans les agents artificiels [BARTNECK, 2002; GRATCH et MARSELLA, 2005], car elle permet de définir une réaction émotionnelle de l'agent en fonction de l'objet, de l'événement ou de l'action l'élicitant. Elle est également utilisée pour les études des interactions humain-machine [CONATI et ZHOU, 2002] et pour la détection de l'affect dans le texte [SHAIKH et collab., 2009].

BARTNECK et collab. [2017]; BARTNECK [2002] ont pointé les limites du modèle OCC dans le cadre du développement des agents virtuels. Le modèle ne contient pas d'information nécessaire pour lier l'état émotionnel d'un personnage virtuel à son comportement. Il lui manque des fonctions de l'historique, de l'interaction des émotions et de la création de personnalité. Selon BARTNECK [2002], l'absence de ces éléments rend artificiel le comportement des personnages virtuels. Dans leurs recherches de meilleures solutions, les chercheurs tentent d'hybrider plusieurs modèles. LIU et PAN [2005], par exemple, utilisent, lors de la création d'un agent virtuel, le modèle OCC et le modèle de Plutchik simultanément.

Il existe des modèles d'évaluation alternatifs [FRIJDA, 1986; LAZARUS, 1966; SCHERER et collab., 2001]. Le *modèle d'évaluation* ("appraisal") de SCHERER et collab. [2001], développé dans le cadre de l'"affective computing" (pour des agents virtuels et robots dotés d'émotions), permet de modèliser, entre autres aspects, la source des émotions. Klaus Scherer a fait partie du réseau d'excellence européenne HUMAINE (Human-Machine Interaction Network on Emotion), dont le but était le développement des systèmes orientés émotions, capables de modéliser ou d'influencer les émotions humaines.

BARRETT [2017] observe que les théories d'évaluation sont trop hétérogènes et propose de se limiter aux théories constructionistes et descriptives de l'évaluation, plus faciles à modéliser dans les systèmes informatiques et plus proches de la réalité humaine, car elle préssuposent que la variabilité est une norme. Dans sa proposition, les phénomènes d'émotion peuvent être décrits de façon continue par la valence, l'excitation, et des dimensions de l'évaluation. Il nous semble difficile d'adapter une annotation continue des OPEM dans le texte.

Le système des attitudes proposé par MARTIN et WHITE [2005] est adapté à l'analyse du texte. Le modèle propose de distinguer une source qui exprime une évaluation sur une cible (situation, produit ou personne). Les attitudes peuvent être mesurées en termes de polarité (positive/négative) et de graduation (il est possible d'amplifier ou de diminuer l'ampleur de l'attitude). Les attitudes peuvent être exprimées de façon directe ou indirecte. WIEBE et collab. [2005] distingue également non seulement la source et la cible des opinions et des émotions mais aussi des expressions directes et implicites des opinions et des émotions. Nous reprenons cette approche pour proposer les types explicites et implicites des problèmes d'interaction et pour modéliser la cible d'OPEM, exprimée implicitement ou explicitement.

Exemple 1 *Les éléments d'une relation*
(...)
User : *tu comprend rien*
Agent : *Je fais de mon mieux et je vais faire en sorte de m'améliorer le plus vite possible. ;Je suis ici pour répondre à vos questions concernant EDF.*
User : *t un ordi*
Agent : *Veuillez m'excuser, je n'ai pas compris ce que vous venez de dire.Je vous invite à reformuler afin que je puisse vous fournir une réponse.*
User : *t virtuelle tu peux pas m aider*
Agent : *Je fais de mon mieux et je vais faire en sorte de m'améliorer le plus vite possible. ;Je*

suis ici pour répondre à vos questions concernant EDF.
(...)

Dans l'exemple 1, dans l'énoncé utilisateur " t virtuelle tu peux pas m aider", "virtuelle ... peux pas m aider" est l'expression de l'opinion négative vers une cible explicite (l'agent) et "m" est la source (l'utilisateur).

Le choix d'un modèle concret dépend d'un contexte d'interaction (dans notre cas c'est une interaction écrite humain-agent) et des objectifs de l'analyse. Dans notre cas, il faut prendre en compte cadre applicatif qui est commercial. Dans la sous-section suivante, nous présenterons les typologies qui sont traditionnellement utilisées dans le cadre commercial d'application.

2.2.3 Les typologies utilisées dans le contexte industriel

Dans le domaine commercial et celui du marketing, les études du comportement des consommateurs se basent sur les théories d'OPEM existantes, en les adaptant en fonction du type d'étude et du champ d'application. Certaines typologies sont purement théoriques, d'autres sont obtenues à la suite de plusieurs expériences.

Typologie théorique des émotions des consommateurs. DERBAIX et PHAM [1989] proposent une typologie des réactions affectives compatible avec un outil mesurant des réactions affectives en *marketing*. Leur but est de couvrir avec les types proposés tous les champs applicatifs du domaine du marketing (la publicité, l'analyse des avis des consommateurs, l'image de marque). Cette typologie est créée à partir de nombreux travaux existants dans le domaine de la psychologie, tels que KEMPER [1987], PLUTCHIK [1980] et des travaux sur l'analyse du comportement des consommateurs tels que HOLBROOK [1986], PIETERS et VAN RAAIJ [1988] et GARDNER [1985]. La typologie en question comprend sept types de réactions affectives, organisés en fonction de l'implication du cognitif dans l'affectif. En adoptant les définitions de IZARD et BUECHLER [1979] et BUCK [1984] de l'émotion, les auteurs classifient l'émotion comme l'un des types de réaction affective les plus spontanés et incontrôlables. Ils mettent l'accent sur l'émotion de surprise puisqu'elle intéresserait beaucoup les publicitaires. Selon les auteurs, les réactions cognitives telles que l'attitude et l'appréciation d'une cible spécifique intéresseraient plutôt les spécialistes de l'analyse des avis client sur un produit ou un service. En complétant la typologie par une proposition de méthodologie d'étude pour chaque type d'affect, les auteurs tentent de relancer les études des réactions affectives dans le marketing.

Typologies appliquées des émotions des consommateurs. Certaines typologies ne conviennent pas à l'analyse de la relation client et même dans le domaine du marketing, le choix des émotions peut varier en fonction des produits et des services proposés aux clients, par exemple, selon MACHLEIT et EROGLU [2000], les échelles d'émotions proposées par IZARD [1977] et les types d'émotions proposés par PLUTCHIK [1980] conviennent mieux pour analyser les émotions de la clientèle des magasins physiques.
LAROS et STEENKAMP [2005] ont proposé un modèle hiérarchique des émotions des consommateurs afin de concilier des types d'émotions de différents courants de recherche sur le sujet. Leur hiérarchie consiste en trois niveaux de généricité : le niveau le plus abstrait est la polarité. Les niveaux suivant sont les émotions "basiques" : satisfaction, bonheur, amour et fierté pour les émotions positives et tristesse, peur, colère et honte pour les émotions négatives. Le niveau le plus détaillé est le niveau des émotions spécifiques qui

contient 42 types d'émotions basés sur RICHINS [1997] (eux-mêmes basés sur des listes d'émotions des consommateurs des études précédentes et des études expérimentales de l'auteur). Le but de l'étude étant d'analyser les émotions des consommateurs dans les conditions d'achat des produits d'alimentation les plus répandus et facilement disponibles, deux émotions de base "l'amour" et "la fierté" ont été exclues de leur étude empirique. L'émotion de l'amour ne correspondait pas au critère de généricité, car elle est caractéristique du contexte très particulier des achats sentimentaux. L'émotion de "la fierté" apparaissant dans les conditions de comparaison du sujet vis-à-vis d'une autre personne, ne correspondait pas non plus au contexte de l'étude. Leur étude empirique a montré que l'utilisation des émotions "basiques" offre de meilleurs résultats que l'utilisation unique du niveau le plus abstrait. La détection de ce type d'émotions permet une meilleure compréhension des émotions des consommateurs. Néanmoins LAROS et STEENKAMP [2005] soulignent que le choix des émotions "basiques" ou des émotions spécifiques doit s'effectuer selon le cadre de la recherche.

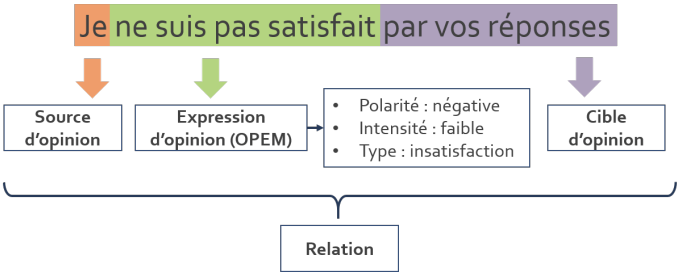
L'étude des émotions des consommateurs des magasins physiques est liée à la perception de l'atmosphère intérieure du magasin. Dans le cas des services dématérialisés, il est plus accessible d'analyser les réactions des consommateurs sur la qualité des produits qu'analyser d'autres facteurs influençant l'utilisateur. Ainsi, PANG et LEE [2008] proposent d'utiliser l'analyse de texte pour parcourir l'Internet en quête d'avis positifs, négatifs et mitigés sur les caractéristiques et la qualité des produits. En même temps, dans le cas des services proposés soit directement chez le client, soit sur le site web de l'entreprise, l'interaction est un des éléments influant sur la perception de la marque, de l'entreprise et du service par le client.
Dans notre cas, l'interaction est très spécifique. C'est une interaction écrite sous forme de tchat entre l'utilisateur du site web de l'entreprise et la conseillère virtuelle. Lors de l'interaction, l'utilisateur peut exprimer des OPEM aussi bien envers les produits et services qu'envers l'agent virtuel. L'utilisateur peut également exprimer des OPEM qui ne sont liés à aucun de ces deux éléments, mais au contexte de sa vie personnelle. Il est donc important de pouvoir les distinguer lors de développement du système de la détection automatique des problèmes d'interaction.

2.2.4 Notre positionnement

A l'instar des chercheurs dans les domaines applicatifs, dont nous avons cité les travaux dans la Section 2.2.3, nous avons étudié les définitions et les typologies des OPEM existant dans les travaux théoriques et nous avons également choisi les types correspondant le mieux à l'analyse des avis des utilisateurs dans les dialogues humain - agent, et qui conviennent également à l'analyse des OPEM dont la cible est les produits ou les services. Cette orientation élargie est motivée par les besoins de l'entreprise. Dans notre recherche, nous nous appuyons sur les définitions d'opinion de MARTIN et WHITE [2005] et MUNEZERO et collab. [2014] qui nous permettent de détecter des opinions comme une classe générique et de détecter des émotions précises au sein de l'opinion lorsque cela est possible.
En ce qui concerne le modèle computationnel des OPEM, nous nous appuyons sur le modèle de MARTIN et WHITE [2005]. Nous distinguerons pour tous les OPEM la **source**, la **cible**, la **polarité** : positive ou négative, et l' **intensité** : faible ou élevée. La relation et ses éléments sont illustrés sur la figure 2.4 page 24. Nous avons exclu la polarité neutre car dans le cas de notre corpus et suivant notre hypothèse, il est peu probable qu'un utilisateur exprime un avis sur un produit, service ou interaction sans porter d'appréciation

personnelle.
Dans le cas des OPEM contenant une émotion, nous préciserons le **type d'émotion**. Pour

FIGURE 2.4 – Illustration d'une relation selon [MARTIN et WHITE, 2005]



pouvoir attribuer une étiquette correspondant à une émotion dans le texte, nous avons recensé les émotions "basiques" les plus fréquentes en psychologie dans la littérature, en nous appuyant surtout sur le tableau transversal d'émotion de COWIE et CORNELIUS [2003], ainsi que quelques émotions mixtes qui sont intéressantes à détecter dans une interaction, par exemple, l'impatience ou le mépris [LARIVEY, 2002]. En se basant sur la première analyse des exemples de dialogues de notre corpus, nous avons ensuite sélectionné 9 types d'émotions positives et 10 types d'émotions négatives qui seraient les plus pertinentes à détecter dans notre corpus. Lors de la confrontation de cette typologie d'émotions avec la sphère applicative du marketing, ainsi qu'en tenant compte du besoin de termes les moins ambigus possibles pour l'annotation manuelle du corpus d'évaluation, cette typologie a été réduite et légèrement modifiée.
Comme nous l'avons déjà mentionné, de nombreuses recherches applicatives se concentrent sur la détection des OPEMs négatives afin de pouvoir les anticiper. Notre liste finale des types d'émotions comporte ainsi six émotions négatives : **la colère, l'inquiétude, l'insatisfaction, la surprise négative, la tristesse et le désarroi**.

Selon LARIVEY [2002], toutes les émotions négatives simples indiquent l'*insatisfaction*. Nous considérons alors que tous les cas où nous pouvons identifier une opinion négative sans pouvoir identifier la colère, l'inquiétude, la surprise négative, la tristesse ou le désarroi, reflètent alors l'insatisfaction de l'utilisateur.

L'**insatisfaction** est une émotion inverse à la satisfaction. Nous trouvons chez LARIVEY [2002] la définition suivante de la satisfaction : "la satisfaction indique que le besoin est comblé". La définition marketing est plus développée : "la satisfaction/insatisfaction est un sentiment de plaisir ou de déplaisir qui naît de la comparaison entre des attentes préalables et une expérience de consommation." [LENDREVIE et LÉVY, 2014]

La colère représente alors une forte insatisfaction. LARIVEY [2002] définit la colère comme "une émotion simple qui traduit l'insatisfaction. Elle est vécue à l'égard de ce qu'on identifie, à tort ou à raison, comme étant "responsable" de notre frustration. On éprouve donc de la colère envers "l'obstacle" à notre satisfaction." [4]

Certaines émotions proposées par les chercheurs contenaient des types d'émotions synonymiques. Nous avons donc choisi celles correspondant le mieux à notre domaine. C'est le cas de l'**inquiétude** que nous avons préféré à l'anxiété, où le second terme reflète l'état psychologique d'une personne ou un trouble psychologique. Nous avons aussi

4. http ://www.redpsy.com/guide/colere.html

considéré que la présence de l'inquiétude est plus probable dans notre corpus, par rapport à la peur, une émotion "primaire" (voir la discussion sur les émotions primaires chez WIERZBICKA [1999a]) et probablement trop forte dans notre contexte. LARIVEY [2002] définit l'inquiétude comme une pseudo-émotion et "une opération intellectuelle qui consiste à imaginer et extrapoler, à partir du présent, une situation plus ou moins désagréable et à se faire du souci à son sujet."

Il faut noter également que l'émotion de surprise n'a pas de polarité chez [PLUTCHIK, 1980] ou [LARIVEY, 2002]. Nous avons ainsi choisi de différencier la surprise négative et positive puisque nous avons émis l'hypothèse que dans le cas d'une interaction écrite dans le cadre de la relation client, il ne peut pas exister d'OPEM de polarité neutre. Comme nous avons statué plus haut, nous nous concentrons que sur les émotions négatives. Nous ne donnons donc pas ici de définition de surprise positive. Nous définissons la surprise négative dans le cadre de notre travail. La **surprise négative** est une réaction émotionnelle de l'utilisateur suite à un changement inattendu alors qu'il aurait préféré, soit rester à l'état antérieur, soit avoir été prévenu du changement.
La **tristesse** est aussi liée au changement. Selon LARIVEY [2002], c'est un manque de nature affective. Dans le cadre de notre corpus cela concerne des changements sans gravité pour l'utilisateur mais qu'il ne soutient pas. L'utilisateur peut ainsi regretter la perte d'un service auquel il était habitué et être déçu de l'aspect irrévocable de ce changement.

Le cas du terme "désarroi" est un peu particulier. Nous avons choisi de l'utiliser alors qu'il ne fait pas partie des listes d'émotions dans la littérature étudiée afin de faciliter la compréhension du type d'émotion et également de le rendre plus pertinent pour notre corpus. Le **désarroi** est alors proposé comme une émotion remplaçant l'embarras [EKMAN, 1999; LEWIS et collab., 2010] et la misère [RUSSELL, 1980], étant plus intense que l'inquiétude, à la frontière avec le désespoir [BANSE et SCHERER, 1996; LARIVEY, 2002]. Nous proposons ce terme couplé avec le terme "désemparé", ce qui doit permettre à l'annotateur de mieux différencier les types d'émotions proposés afin de garder les annotations cohérentes. Cette émotion traduit les cas où l'utilisateur ne sait plus comment faire pour arriver à ses fins et demande de l'aide à l'agent virtuel.

Les types des émotions choisies sont illustrés par des exemples dans le Chapitre 5 page 67.

TYPE D'ÉMOTION	EXEMPLES
Colère	« 1 honte cette taxe CTA »
Inquiétude	« bonjour, je viens de payer ma facture par CB au téléphone, mais je n'ai toujours pas reçu de mail de confirmation de paiement. »
Insatisfaction	*Utilisateur* : « je n'arrive pas à me connecter » *Agent* : « Bonjour Madame N ; Si vous avez rencontré un problème au moment de la connexion, il peut s'agir d'un problème technique. Nous faisons le maximum pour vous garantir le minimum de désagréments. » *Utilisateur* : « ce serait bien agréable de le savoir car voilà deux fois que me l'on change mon mot de passe. Pas très satisfaite »
Surprise négative	« je suis surpris de l'augmentation de mes kw »
Tristesse	« c'est dommage que vous arrêtez le tarif T, je l'aimais bien »
Désarroi	« j'ai perdu mon mot passe mais en plus il me dit que mon identifiant et faux je ne comprends pas ...j'ai bloqué mon compte. aider moi merci d'avance. » « je suis une mère isolée avec deux enfants. Mon électricité est coupée, quoi faire ? » « quoi faire après avoir reçu une relance injustifiée de facture impayée ? »

TABLEAU 2.1 – Types de l'OPEM négative (termes émotionnels). Le tableau tiré du guide d'annotation qui prévoit l'annotation de l'OPEM vers deux types de cibles : interaction et produits/services

2.3 Conclusion

Dans ce chapitre nous avons présenté les définitions et les typologies des problèmes d'interaction, des opinions et des phénomènes reliés aux opinions. Nous avons vu que les définitions des problèmes d'interaction présentées dans la littérature sont souvent données en fonction du contexte applicatif. Nous nous arrêtons donc sur la définition permettant d'avoir le point de vue de l'utilisateur sur l'interaction et non le point de vue du système. Nous avons vu également que les typologies existantes des problèmes d'interaction se basent surtout soit sur la présence d'un problème évident (un breakdown/une intervention d'un agent humain), soit sur l'absence de compréhension ou de coopération. L'approche de la détection des problèmes d'interaction du point de vue de l'opinion de l'utilisateur dans les interactions humain-agent est récente et peu étudiée pour le texte. Elle n'a pas encore été étudiée pour le français. La distinction des attitudes directes et indirectes de MARTIN et WHITE [2005] et des expressions explicites et implicites des OPEMs chez WIEBE et collab. [2005] nous inspirent pour la typologie des problèmes d'interaction décrit dans la Section 5.1 page 68.

La littérature étant très riche en travaux sur la théorie des émotions, nous avons présenté dans ce chapitre les courants les plus significatifs de la théorie sur les opinions et les phénomènes reliés aux opinions : les modèles dimensionnels, les modèles catégoriels, et les modèles reposant sur la théorie de l'évaluation. Nous avons revu plus en détail les définitions et les modèles de l'opinion et de l'émotion.
Nous avons démontré que suivant le domaine d'application les choix de modèles varient

en fonction du but de l'analyse des émotions et du contexte. Les conditions de l'analyse des émotions influent également dans le cadre de la relation client (par exemple, présence ou absence d'un espace commercial physique ou virtuel) et dans le cadre de l'interaction : face-à-face, vocale ou écrite.
Dans la dernière section nous avons présenté la typologie des OPEM, que nous avons choisie pour l'analyse des interactions écrites pour la relation client : sa cible, sa source, sa polarité, son intensité et son type.

Chapitre 3

État de l'art. Méthodes de détection des opinions et des problèmes d'interaction

Sommaire

Nous décrivons dans ce chapitre les méthodes utilisées pour la détection des problèmes d'interaction aussi bien que pour la détection des opinions et des phénomènes reliés aux opinions. Cet état-de-l'art des méthodes nous permet de justifier nos choix pris lors du développement du système de détection des problèmes d'interaction. Le panorama des approches existantes nous permettra également d'exposer nos idées pour les développements futurs du système.

3.1 Méthodes de détection des problèmes d'interaction

Lorsque la question de la détection de problèmes d'interaction (PI) se pose, il est important de connaître les indices signalant la présence d'un phénomène. Ces indices peuvent être soit manipulés par des règles linguistiques, soit servir de descripteurs en entrée des systèmes d'apprentissage. Nous détaillons ensuite les travaux existants dans le domaine.

3.1.1 Indices utilisés pour la détection des problèmes d'interaction dans les dialogues humain-machine

Nous parlerons ici des indices utilisés dans la littérature pour la détection des problèmes d'interaction dans les dialogues humain - agent conversationnel animé (ACA). Nous présenterons ces indices d'abord de manière générale et ensuite nous donnerons d'avantage de détails sur les indices dans le texte.

La présentation générale des indices

Lorsque l'ACA est capable de "voir" l'utilisateur, les *indices visuels* peuvent renforcer la détection des problèmes d'interaction basée sur d'autres indices. Cela peut être la détection du départ de l'utilisateur, de ses gestes ou de ses expressions faciales [TURK, 1996; BARKHUYSEN et collab., 2005].

Dans les systèmes ou les centres d'appels humain-humain où seule l'interface vocale est disponible, les indices utilisés peuvent être séparés en six groupes : *les retours de l'utilisateur via un questionnaire* [HONE et GRAHAM, 2000; HARTIKAINEN et collab., 2004; HASTIE et collab., 2002], les indices liés au *signal audio* (prosodie [HIRSCHBERG et collab., 1999; GEORGILADAKIS et collab., 2016]), *les caractéristiques techniques de dialogue* telles que les séries temporelles caractérisant le déroulement d'un dialogue (la longueur d'un tour de parole [SUIGNARD et collab., 2012; KRAHMER et collab., 1999], la durée d'un dialogue [LANGKILDE et collab., 1999; HASTIE et collab., 2002]) ou d'autres informations quantitatives (nombre de tours de parole [LANGKILDE et collab., 1999; HASTIE et collab., 2002]), les données concernant le bon fonctionnement du système de l'agent (*logs* [WALKER et collab., 2002]), l'historique du dialogue [VAN DEN BOSCH et collab., 2001] et l'analyse de la parole transcrite [WALKER et collab., 2002; KRAHMER et collab., 1999]. Ces indices peuvent être obtenus de façon automatique ou annotés manuellement [LANGKILDE et collab., 1999]. Les deux derniers types d'indices sont le point de convergence entre les indices utilisés pour les systèmes vocaux et les chatbots où le seul moyen de communication disponible est l'écrit. Les indices contenus dans le texte sont des indices linguistiques. Dans la littérature, les indices linguistiques se séparent en :

— mesures matématiques, appliquées aux chaînes de caractères et applées indices lexicaux [GEORGILADAKIS et collab., 2016], et

— indices sémantiques qui ont pour but de modéliser le sens des mots dans le texte.

Nous allons donc détailler davantage les *indices lexicaux*, *sémantiques* et la prise en compte de l'*historique* du dialogue. Il est possible également d'analyser le texte en terme de présence d'opinion ou de phénomènes reliés aux opinions (*OPEM*) de l'utilisateur, exprimés envers l'interaction avec un agent, dans le but de détecter des PI [GEORGILADAKIS et collab., 2016; XIANG et collab., 2014].

Les indices lexicaux

KRAHMER et collab. [1999] étaient parmi les premiers à proposer des indices "négatifs" dans le tour de parole de l'utilisateur, c'est-à-dire des indices indiquant des problèmes d'interaction. Les indices qu'ils observent dans leur corpus de dialogues sont : un ordre marqué de mots (ex. phrases emphatiques), la réfutation, l'absence de réponse utilisateur, lorsque l'utilisateur fait des corrections, se répète ou ne fournit pas de nouvelles informations. Leur analyse du corpus annoté montre que l'apparition des corrections de l'utilisateur est un des indices les plus importants des problèmes d'interaction. L'analyse des règles, induites par l'algorithme RIPPER [COHEN, 1996], présenté par VAN DEN BOSCH et collab. [2001], confirme la pertinence des répétitions de l'utilisateur en tant qu'indice des PI. Cette analyse indique également que le vocabulaire marqué, tel que l'utilisation des termes archaïques, peut représenter un indice pour la détection des PI. Le même type d'analyse, fait par MEENA et collab. [2015], montre également la pertinence de la réfutation de l'agent.

Les indices lexicaux utilisés par les systèmes d'apprentissage automatique se limitent souvent aux descripteurs de "bas" niveau. Ce sont des "sacs-de-mots" [1] [LENDVAI et collab., 2002a; LUQUE et collab., 2017; VAN DEN BOSCH et collab., 2001], des "sacs-de-concepts" [2] [MEENA et collab., 2015], bigrammes des charactères [GEORGILADAKIS et collab., 2016], "n-grammes" de tours de paroles et TF-IDF [BEAVER et FREEMAN, 2016] ou "n-grammes" des actes de dialogue [HARA et collab., 2010]. Par ailleurs, LENDVAI et collab. [2002a] considèrent que l'approche "sac-de-mots" capture toutes les informations nécessaires sur "...la diversité sémantique, structure syntaxique, répétitions, omissions, corrections..." et permet d'éviter l'utilisation des approches de plus haut niveau. Toutefois, les dialogues que LENDVAI et collab. [2002a] utilisent sont assez longs : 8 paires adjacentes en moyenne. Nous estimons que dans les dialogues plus courts, le contexte n'est pas suffisant pour s'appuyer uniquement sur les indices de bas niveau.

La répétition ou la reformulation en tant qu'indice linguistique sont couramment utilisées pour la détection des PI. Elles sont détectées grâce à de multiples calculs de similarités :

— entre l'objet de la question Q de l'utilisateur et l'objet de sa question suivante Q+1 (qui selon l'expérience de CHAI et collab. [2006] ne sont pas très performantes pour cette tâche) ;

— entre deux énoncés utilisateur successifs [CHAI et collab., 2006; GEORGILADAKIS et collab., 2016] ;

— entre deux énoncés successifs de l'agent [GEORGILADAKIS et collab., 2016] ;

— entre les énoncés utilisateur et agent [CHAI et collab., 2006] ;

1. Un "sac-de-mots" est une approche introduite dans le domaine de la recherche d'information dans les textes par [SALTON, 1971]. Cette approche consiste en la représentation d'une phrase ou d'un texte comme l'ensemble de ses mots et de l'information de leur nombre d'occurrences dans le texte.

2. « concept » chez [MEENA et collab., 2015] est un champ que le système doit remplir avec les informations clés de l'utilisateur

— entre les entités nommées de deux énoncés utilisateur [CHAI et collab., 2006] ;

Les indices lexicaux utilisées en tant que descripteur pour la détection de la similarité entre deux textes sont présentés dans le tableau 3.1

Indice lexical	**Principe**
La similarité des traits communs et des différences de Lin [LIN et collab., 1998] utilisée par [CHAI et collab., 2006]	La similarité de [LIN et collab., 1998] est théorique et veut être universelle. LIN et collab. [1998] ne proposent pas de formule. Les auteurs disent qu'il est possible de la mesurer dans les systèmes probabilistes
Distance de Levenshtein chez [GEORGILADAKIS et collab., 2016; LISCOMBE et collab., 2005]	Mesure la distance d'édition. Applicable au niveau des caractères. Permet de gérer jusqu'à certain seuil les fautes de frappe et les erreurs d'orthographe
Coefficient de Dice utilisé par [GEORGILADAKIS et collab., 2016]	Cherche à détecter des digrammes communs. Proche de Jaccard, mais n'est pas une distance
Nombre de concepts en commun utilisé par [MEENA et collab., 2015]	Chaque concept est un élément d'information requis pour accomplir la tâche attendue (par exemple, l'utilisateur doit fournir le nom de la ville de départ lors de l'achat d'un billet). Les concepts se distinguent par type. Pour pouvoir utiliser cet indice, il est nécessaire d'avoir accès aux informations que le système extrait des énoncés utilisateur.
Nombre de relations de dépendance en commun dans les arbres de dépendance utilisé par [XIANG et collab., 2014]	Nécessite une annotation correcte en parties de discours

TABLEAU 3.1 – Les indices lexicaux utilisés pour la détection de la répétition ou reformulation de l'utilisateur

XIANG et collab. [2014] utilisent quatre types d'intentions de l'utilisateur de bas niveau pour la détection des PI : changer (lorsque l'utilisateur change de thème), réessayer (correspond à la répétition ou la reformulation), continuer (la question de l'utilisateur concerne le même thème que le précédent) et clarifier (l'utilisateur reformule son énoncé afin de mettre au clair son intention).

Les indices sémantiques

Dans la sous-section précédente nous avons passé en revue les mesures statistiques appliquées aux chaînes de caractères et permettant de modéliser la répétition ou reformulation. À côté de ces mesures, la similarité sémantique est également utilisée. Elle permet de capturer le sens de mots de façon plus large que celui d'un dialogue en cours. Elle peut être mesurée soit en ayant recours aux ontologies [XIANG et collab., 2014], soit en calculant la similarité géométrique entre les deux vecteurs de deux énoncés. La similarité cosinus est la plus utilisée à ces fins [MEENA et collab., 2015; GEORGILADAKIS et collab.,

2016]. GEORGILADAKIS et collab. [2016] utilisent également d'autres descripteurs sémantiques : une valeur binaire, indiquant l'existence de paraphrases en suivant la méthodologie de SOCHER et collab. [2011] basée sur des autoencodeurs récursifs, et le score de concrétisme[3] de chaque chunk, s'appuyant sur la base de données psycholinguistiques MRC [COLTHEART, 1981].

Il est intéressant de noter que dans les travaux les plus récents, le calcul de la distance cosinus varie en fonction de la méthode de calcul des vecteurs des énoncés. Ainsi, LOPES [2017] utilise les vecteurs de mots pré-entrainés[4] pour construire les vecteurs des énoncés, en prenant la moyenne de la somme des vecteurs de mots de l'énoncé.
SUGIYAMA [2017] utilise un modèle sequenciel (seq2seq) pour obtenir les vecteurs des énoncés.

En analysant les résultats de son travail, LOPES [2017] a souligné qu'un indice, tel que la similarité sémantique entre deux énoncés, peut être utilisé pour les dialogues orientés tâche aussi bien que pour les dialogues de conversation libre. En revanche, dans le premier cas, la similarité élevée indiquerait la présence d'un problème dans le dialogue (la répétition), alors que dans un dialogue de conversation libre, elle indiquerait la cohérence du développement d'un sujet et donc l'absence de problème. [LOPES, 2017].

L'historique de dialogue

Il existe de nombreuses façons de prendre en compte l'historique de dialogue. Ici nous ne mentionnons que ceux relatifs à notre sujet. Dans les travaux sur la détection/prédiction des PI le moyen le plus répandu de prise en compte de l'historique du dialogue est le suivi de l'ordre de succession des actes de dialogues [SCHMITT et collab., 2011; HIGASHINAKA et collab., 2010a; HARA et collab., 2010]. Ils peuvent être de nature générique, comme "ouvrir, fermer, informer, requêter, accepter, rejeter" [LANDRAGIN et ROMARY, 2004] ou spécifique au corpus, par exemple un acte intitulé "Solde" dans le corpus HMIHY[5] 0300 [LISCOMBE et collab., 2005]. Les chercheurs ont expérimenté la prise en compte de l'historique de dialogue avec la prise en compte des actes de dialogue n'appartenant qu'à l'utilisateur [HARA et collab., 2010; KIM, 2007], qu'au système [VAN DEN BOSCH et collab., 2001; HASTIE et collab., 2002; LENDVAI et collab., 2002b; HARA et collab., 2010] ou à l'utilisateur et au système [EL ASRI et collab., 2014; HARA et collab., 2010]. Néanmoins SCHMITT et collab. [2011] affirment que les actes de dialogue n'apportent pas suffisamment d'amélioration au modèle de prédiction de la qualité d'interaction humain-agent. Leurs tests montrent que les informations automatiques du système de dialogue oral, tels que la réussite de la reconnaissance vocale, la réussite de l'interprétation de la parole transcrite et l'action du système en cours sont les plus pertinentes pour un modèle SVM de classification des tours de parole.

BECHET et collab. [2004] représentent l'historique de dialogue sous forme d'une chaîne des états du dialogue[6]. BECHET et collab. [2004] l'utilisent comme l'un des indices pour la classification non-supervisée des dialogues du corpus "How may I help you?[GORIN et collab., 1997] afin d'obtenir des groupes de dialogues significatifs. Selon les résultats qu'ils ont obtenus, l'historique de dialogue n'est pas un élément significatif pour la discrimination des dialogues problématiques pour l'utilisateur lorsque les dialogues sont courts (quatre paires adjacentes en moyenne).

3. Le concrétisme est une variable sémantico-psycholinguistique qui évalue à quel point le sens d'un mot est abstrait ou concret [TROCHE et collab., 2017]

4. Google News 100B 3M words https ://github.com/3Top/word2vec-api

5. « How May I Help You? »

6. Un état de dialogue est un label attribué par l'unité de gestion de dialogue [BECHET et collab., 2004]

HARA et collab. [2010] utilisent l'historique de dialogue à base de "n-grammes"des actes de dialogue pour estimer la satisfaction des utilisateurs de façon supervisée et démontrent l'importance de la prise en compte de l'historique. La spécificité du corpus utilisé pour leur étude est que les utilisateurs "naïfs" avaient pour tâche de communiquer avec le système de recherche d'enregistrements musicaux pendant au moins quarante minutes ou d'écouter au moins cinq chansons qui devait résulter à au moins vingt dialogues de question-réponse. Ensuite, les utilisateurs répondaient à un questionnaire leur permettant d'exprimer leur niveau de satisfaction concernant le dialogue avec le système. En moyenne, les dialogues dont les utilisateurs ont été satisfaits étaient courts (trois - quatre paires adjacentes). Les dialogues insatisfaisants contiennent jusqu'à 107 paires adjacentes en moyenne. De plus, les utilisateurs ne pouvaient utiliser que des phrases très simples lors de leurs dialogues avec le système.

Cette contradiction dans les résultats de l'utilisation de l'historique de dialogue sous forme d'actes de dialogues en tant qu'indice pour la détection des problèmes d'interaction démontre que son efficacité varie en fonction des données et de la méthode (supervisée ou non-supervisée) appliquée aux données.

LISCOMBE et collab. [2005] prennent en compte l'historique émotionnel de l'utilisateur.

Lors du défi de détection des "breakdowns" DBDC3 [HIGASHINAKA et collab., 2017], où le "breakdown" est défini comme une difficulté à continuer la conversation, les équipes participantes prenaient en compte l'historique de dialogue, en calculant la similarité sémantique entre l'énoncé du système en cours, l'énoncé précédant de l'utilisateur et du système [KATO et SAKAI, 2017] ou tous les énoncés précédent du système et de l'utilisateur [LOPES, 2017].

Pour pouvoir suivre l'évolution du sujet dans un dialogue, SUGIYAMA [2017] a également calculé la distance de déplacement des mots : Word Mover's Distance [KUSNER et collab., 2015] définie comme la distance entre le plongement d'un mot d'un document jusqu'à un autre plongement d'un mot dans un autre document.

L'opinion et les phénomènes reliés à l'opinion en tant qu'indice

L'opinion et les phénomènes reliés à l'opinion (OPEM) en tant qu'indice sont créés de deux façons, en utilisant : 1) l'annotation manuelle [SCHMITT et collab., 2011; LISCOMBE et collab., 2005] ou 2) des dictionnaires [GEORGILADAKIS et collab., 2016; CAILLIAU et CAVET, 2010; ROY et collab., 2016; XIANG et collab., 2014], ainsi que des smileys [ROY et collab., 2016]. GEORGILADAKIS et collab. [2016] utilisent un dictionnaire des termes affectif de l'anglais [PALOGIANNIDI et collab., 2015] qui est une extension du dictionnaire ANEW [BRADLEY et LANG, 1999]. CAILLIAU et CAVET [2010] utilisent le lexique de l'évaluation Blogoscopie [VERNIER et MONCEAUX, 2010] pour détecter les sentiments afin d'identifier des dialogues problématiques dans les centres d'appels français.

En ce qui concerne la typologie des OPEM utilisés, ROY et collab. [2016]; XIANG et collab. [2014] n'utilisent que la valence positive et négative des termes. SCHMITT et collab. [2011] utilisent trois catégories émotionnelles de valence pour la prédiction de la qualité d'interaction : "pas en colère", "en colère" et "très en colère", mais selon les résultats que ces chercheurs ont obtenus, ces indices sont peu performants pour la prédiction de la qualité d'interaction. GEORGILADAKIS et collab. [2016] utilisent la valence, l'activation et la dominance. XIANG et collab. [2014] effectuent également la détection de la cible de l'OPEM pour distinguer les OPEM envers l'interaction des OPEM envers des sujets libres de discussion, en utilisant les arbres de dépendances.

3.1.2 Les approches pour la détection des problèmes d'interaction

La grande majorité des travaux pour la détection et la prévision des PI utilise une approche basée sur l'aprentissage automatique. RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [COHEN, 1995, 1996] est le programme d'apprentissage automatique supervisé le plus populaire dans les premiers travaux sur la prédiction des PI dans les dialogues oraux humain-machine. Il est, par exemple, utilisé par WALKER et collab. [2002]; LANGKILDE et collab. [1999]; LENDVAI et collab. [2002a]; VAN DEN BOSCH et collab. [2001]; MEENA et collab. [2015]; GEORGILADAKIS et collab. [2016]. L'algorithme de [COHEN, 1995] est une amélioration des systèmes d'apprentissage automatique des règles de classification de données. Il permet de réduire le taux d'erreur car il supprime une à une les règles qui en produisent le plus. L'avantage du programme est qu'il permet de classifier le corpus en deux classes rapidement et de pouvoir visualiser les règles qu'il a déduit pour la compréhension plus approfondie du phénomène recherché et l'intégration des règles déduites dans le système de l'agent [WALKER et collab., 2002].

Les premiers systèmes de prédiction des PI au niveau des dialogues, comme celui de WALKER et collab. [2002], sont suivis par les systèmes cherchant à détecter et prédire les PI au niveau des énoncés. Les modèles de Markov cachés (HMM) permettent de modéliser une suite d'opinions d'utilisateurs lorsque le corpus d'entraînement contient des informations sur l'avis de l'utilisateur au niveau des énoncés [ENGELBRECH et collab., 2009] et même lorsque les annotations manuelles de la satifaction utilisateur ne sont disponibles qu'au niveau du dialogue [HIGASHINAKA et collab., 2010b]. HIGASHINAKA et collab. [2010a] démontrent également l'efficacité des HMM par rapport aux méthodes utilisant les champs aléatoires conditionnels (CRF), expliqué par sur-apprentissage des CRF. SCHMITT et collab. [2011]; EL ASRI et collab. [2014]; MEENA et collab. [2015] démontrent l'efficacité des machines à vecteur de support (SVM) sur la même tâche. XIANG et collab. [2014] testent les SVM, la classification naïve bayésienne (NB), les arbres de décisions et les CRF. Ils obtiennent de meilleurs résultats ($F_1 = 62,19\%$) avec les SVM [7] en utilisant en entrée des descripteurs pour les sentiments et les intentions de l'utilisateur. Selon les auteurs, ce modèle s'avère donner également les performances les plus stables.

Les travaux les plus récents se tournent vers l'apprentissage des réseaux de neurones. PRAGST et collab. [2017] appliquent des réseaux de neurones récurrent (RNN) pour l'estimation de la qualité de l'interaction au niveau des énoncés. L'idée principale est de modéliser l'évolution séquentielle de la satisfaction de l'utilisateur et par conséquent, de prendre en compte une partie de l'historique de dialogue. Ils démontrent que les RNN peuvent être plus performants que les réseaux de neurones non-récurrents.

Les approches représentées lors de la compétition en détection des ruptures des dialogues dans les tchats japonais, tenue au Japon en octobre 2015 [FUNAKOSHI et collab., 2016; HIGASHINAKA et collab., 2016] sont représentatives des approches existantes pour cette tâche. Les détails des approches sont décrits par chacune des six équipes participant dans [KOBAYASHI et collab., 2015; MIZUKAMI et collab., 2015; SUGIYAMA, 2015; INABA et TAKAHASHI, 2015; HORII et ARAKI, 2015; TANIGUCHI et KANO, 2015] rédigés en japonais. Malheureusement, [FUNAKOSHI et collab., 2016; HIGASHINAKA et collab., 2016] n'indique pas quel équipe correspond à quels auteurs.
La tâche du défi de détection de "breakdown" de dialogue était d'identifier si les énoncés du système sont la cause des ruptures de dialogues. "Breakdown" est défini ici comme une situation où il est difficile de poursuivre la conversation. Le corpus d'apprentissage représente 1 146 dialogues en japonais entre les utilisateurs et un système de tchat non

7. Les résultats pour les autres modèles sont : J48 37,75% ; NB 43,35 ; CRF 60,16

axé sur les objectifs. Chaque dialogue contient 10 paires adjacentes. 100 dialogues ont été annotés par 24 annotateurs. Les 1046 dialogues restants ont été annotés par deux ou trois annotateurs. Les labels utilisés sont :

1. NB (Not a breakdown) : il est facile de continuer la conversation.
2. PB (Possible breakdown) : il est difficile de continuer une conversation fluide.
3. B (Breakdown) : il est difficile de continuer la conversation.

Pour le défi, de nouveaux dialogues ont été collectés par la méthode de crowdsourcing. Ils ont été annotés par 30 annotateurs. Le corpus de développement contenait 20 dialogues et le corpus de test contenait 80 dialogues. Nous énumérons ici ces approches en précisant les principaux descripteurs. Les organisateurs de la compétition ont proposé un système de référence utilisant les CRFs. Une des équipes participant a proposé une approche à base de mots clés et de règles. Une autre équipe a utilisé les SVM, en s'appuyant sur les vecteurs de la fréquence des mots dans l'énoncé du système et de l'énoncé précédent de l'utilisateur. Quatre équipes ont fait appel aux techniques de réseaux de neurones profonds, dont trois : une approche combinant les mémoires longues à court terme (LSTM) avec des réseaux neuronaux récurrents (RNN). Les vecteurs de mots dans un des systèmes ont été créés à l'aide de Word2Vec [MIKOLOV et collab., 2013a]. Les vecteurs de mots et de cooccurences ont été créés avec Sent2Vec [MIKOLOV et collab., 2013b] dans le second système. Le troisième utilisait les vecteurs de mots créés avec un modèle neuronal conversationnel [VINYALS et LE, 2015], LSTM et des vecteurs des sacs-de-mots. Le sixième et dernier participant utilisait les réseaux de neurones profonds avec pour entrée des actes de dialogue du système et de l'énoncé précédant de l'utilisateur, l'acte de dialogue de l'énoncé suivant du système, la perpléxité de l'énoncé du système calculé à base des "n-grammes" des mots des énoncés et les résultats de classification des questions utilisateur. Les meilleurs résultats sont obtenus par les équipes utilisant des réseaux de neurones. Le meilleur score d'exactitude (Accuracy) obtenu est de 0,643. Le meilleur score de F-mesure obtenu pour la tâche de classification entre "break-down" et "break-down" possible est 0,468, pour la classification entre "break-down" et non "break-down" 0,798 [HIGASHINAKA et collab., 2016].

Le troisième défi pour la détection de la rupture de dialogue (DBDC3 [8]) souligne la tendance initiée par l'augmentation de la quantité de données présentes sur Internet : le recours aux réseaux de neurones. DBDC3 s'est tenu aux États-Unis dans le cadre du sixième défi pour les technologies des systèmes de dialogue (Dialog System Technology Challenges 6) en décembre 2017. Les organisateurs ont décrit le défi dans [HIGASHINAKA et collab., 2017]. Contrairement aux défis précédents, la langue des données du défi a été non seulement le japonais mais également l'anglais.
L'objectif est resté le même que pour le défi précédent. Pour chaque paire d'un échange entre un utilisateur et un système, les participants devaient proposer une étiquette (NB, PB ou B) et la probabilité de distribution des étiquettes. Les données proposées pour le développement et l'évaluation des systèmes sur les dialogues en anglais consistaient en quatre corpus. Deux (Tick Tock et IRIS) sont les corpus des workshops sur les chatbots WOCHAT [9]. Le troisième corpus est un corpus CIC collecté lors du hackathon DeepHack d'une école d'été Turing à l'université de Physique et Technologie de Moscou (PTM). Le quatrième corpus est le corpus YI qui a été créé par les organisateurs, en demandant aux intervenants sur Amazon Mechanical Turk (AMT) [10] de discuter avec un chatbot déve-

8. Dialogue Breakdown Detection Challenge 3
9. http ://workshop.colips.org/wochat/data/index.html
10. https ://requester.mturk.com/

loppé par l'université PTM [11]. L'ensemble des corpus représente 600 dialogues ayant pour but une discussion libre. Chaque dialogue du corpus a une longueur de 20 paires adjacentes. Un tiers des dialogues a servi pour l'évaluation des systèmes. Les données ont été annotées par la méthode de crowdsourcing sur CrowdFlower [12]. Malgré l'effort notable fourni par les organisateurs pour offrir aux participants du défi une quantité de données conséquente, la taille du corpus de développement proposé pour le japonais (1546 dialogues) était pratiquement quatre fois plus grand que celui pour l'anglais. En effet, le corpus de développement pour le japonais bénéficiait des données de tous les défis précédent.

Au total, huit équipes ont pris part au défi : six ont travaillé sur l'anglais et quatre sur le japonais. Les meilleurs résultats pour la prédiction de la rupture de dialogue en anglais en termes d'exactitude (Accuracy) ont été obtenus par le représentant de l'institut suédois José Lopes 0,44. LOPES [2017] a utilisé des LSTM.

En terme de F-mesure, le meilleur résultat (0,36) sur la prédiction de la rupture des dialogues pour l'anglais a été obtenu par l'équipe de l'entreprise japonaise Sarl Nextremer Co.. Cette équipe a obtenu également le meilleur résultat de F-mesure (0,87) pour la prédiction des étiquettes PB et B pris comme une seule étiquette [HIGASHINAKA et collab., 2017]. Le modèle conçu par l'équipe est décrit chez IKI et SAITO [2017]. IKI et SAITO [2017] proposent un modèle basé sur le réseau de mémoire, les plongements des phrases calculés au niveau des caractères pour prendre en compte les mots inconnus et un mécanisme externe de mémoire pour intégrer des informations au contexte local des dialogues.

Les meilleurs résultats pour le japonais ont été obtenus par la méthode qui combinait un ensemble des régresseurs telles que le regresseur de forêts aléatoires, le regresseur d'arbres supplémentaires ou encore le regresseur de k-plus proches voisins. L'algorithme t-SNE (t-distributed stochastic neighbor embedding) a été utilisé pour réduire les dimensions des vecteurs [SUGIYAMA, 2017]. Cette méthode a été proposée par SUGIYAMA [2017] dont les résultats sont l'exactitude (Accuracy) = 0,61 et F-mesure = 0,67 pour la prédiction de l'étiquette "B".
En ce qui concerne la prédiction de "PB+B" comme une seule étiquette pour les dialogues en japonais, l'équipe de l'université de Waseda a obtenu le meilleur résultat : 0,83 de F-mesure [HIGASHINAKA et collab., 2017]. Cette équipe a utilisé la même approche que SUGIYAMA [2017], mais en prenant les vecteurs de fréquence de termes et les vecteurs de plongements de mots pour calculer les similarités entre des énoncés [KATO et SAKAI, 2017].

D'autres approches proposés lors du défi pour la prédiction de la rupture de dialogue sont :

— un modèle de l'entropie maximale de Soochow University pour l'anglais ;

— RNN couplés au modèle de l'attention entre des phrases [PARK et collab., 2017] pour l'anglais ;

— Bi-LSTM et les plongements de mots préentrainés avec GLOVE [XIE et LING, 2017] pour l'anglais ;

— CNN et LSTM pour le japonais [TAKAYAMA et collab., 2017]

11. https ://www.slideshare.net/sld7700/skillbased-conversational-agent-80976302
12. https ://www.crowdflower.com/

Analyse de l'opinion et des phénomènes reliés aux opinions pour la détection des problèmes d'interaction

Lors de la présentation des indices utilisés pour la détection des PI, nous avons présenté l'analyse des OPEM en tant qu'indice. Nous exposons ici les méthodes d'analyse des OPEM pour la détection des PI. Les systèmes où l'interface vocale est disponible fournissent des informations prosodiques qui peuvent être utilisées pour la détection des PI [ANG et collab., 2002; SEGURA et collab., 2016]. Puisque nous nous intéressons uniquement au texte, nous évoquerons ici des approches utilisées dans les systèmes de chatbots, ainsi que dans les systèmes où la parole transcrite est analysée.

Les méthodes utilisées pour le texte varient de celles basées sur des règles à celles utilisant l'apprentissage automatique. Ainsi, HIGASHINAKA et collab. [2015b] font la détection des sentiments dans les énoncés avec un "break-down" en utilisant des **règles conçues manuellement**. CAILLIAU et CAVET [2010] appliquent des patrons et des heuristiques pour déterminer la polarité des opinions et des sentiments exprimés par les interlocuteurs d'un centre d'appel français. XIANG et collab. [2014] utilisent l'aprentissage automatique basé sur des mots clés des dictionnaires de sentiments chinois et également des patrons. Afin d'extraire des descripteurs affectifs, pour ensuite les utiliser pour la détection des PI avec le classifieur RIPPER, GEORGILADAKIS et collab. [2016] calculent un score pour chaque mot de l'utilisateur en utilisant un dictionnaire affectif [PALOGIANNIDI et collab., 2015] et des statistiques de base de ces scores.

LISCOMBE et collab. [2005] utilisent un algorithme BoosTexter [13] [SCHAPIRE et SINGER, 2000] qui leur permet de combiner plusieurs types de descripteurs : les actes de dialogue, le contexte prosodique, lexical et dialogique afin d'attribuer aux énoncés utilisateur une classe négative ou non-négative. Pour modéliser les enchaînements des émotions dans des phrases d'un énoncé utilisateur d'un centre d'appel, ROY et collab. [2016] utilisent des CRF.

3.2 Méthodes de détection des opinions et des phénomènes reliés aux opinions

Les méthodes de détection des opinions et des phénomènes reliés aux opinions (OPEM) dépendent beaucoup du type de texte à analyser. Nous mettons ici l'accent sur des textes courts qui peuvent être mal structurés syntaxiquement et mal orthographiés. L'analyse des OPEM peut être réalisée au niveau d'un document, en présupposant qu'un document, tel qu'un revue de film, n'exprime qu'un seul OPEM [MORAES et collab., 2013; YESSENALINA et collab., 2010]. Il existe deux niveaux de détection plus précis : au niveau d'une phrase et au niveau des mots et des expressions, autrement appelée l'analyse à grain fin. L'analyse à grain fin consiste à détecter des éléments tels que définis, par exemple, par LIU [2012] : la cible, un aspect de la cible, la source d'opinion, le moment où l'opinion a été exprimée et l'opinion envers une caractéristique de la cible exprimée à ce moment. L'analyse au niveau de la phrase peut également inclure des éléments plus précis, comme la détection de la source de l'OPEM [KIM et HOVY, 2004]. Dans le cadre des interactions humain-agent virtuel, l'analyse des OPEM de l'utilisateur au niveau de l'énoncé est nécessaire pour adapter la réponse de l'agent [CLAVEL et CALLEJAS, 2016]. La détection de la cible de l'OPEM de l'utilisateur permettrait de distinguer les OPEM exprimées envers

13. un algorithme utilisant une technique d'apprentissage automatique par renforcement et permettant une classification de texte contenant plusieurs étiquettes.[SCHAPIRE et SINGER, 2000]

l'interaction de ceux relatifs au thème de la discussion.

La majorité des méthodes de détection des OPEM, présente dans la littérature, peut être divisée en trois groupes : les méthodes à base de règles, à base d'apprentissage automatique ou "machine learning" (ML) et des méthodes hybrides.

3.2.1 Les méthodes linguistiques

Les méthodes linguistiques s'appuient sur des dictionnaires [PENNEBAKER et collab., 2001; TABOADA et collab., 2011], des règles [GINDL et collab., 2013; DING et LIU, 2007] et des patrons [MAHARANI et collab., 2015; ZHANG, 2012; ZHANG et collab., 2010; PENNEBAKER et GRAYBEAL, 2001] prenant en compte la structure syntaxique des phrases [SAMHA, 2016]. Elles permettent de modéliser la propagation d'un sentiment dans les phrases [GINDL et collab., 2013], d'identifier l'orientation sémantique [14] des mots en fonction du contexte [DING et LIU, 2007], de détecter des cibles, leurs aspects [GINDL et collab., 2013; SAMHA, 2016] et de désambiguïser des cas comme les cibles multiples, comme dans la phrase "Le téléphone a un bon écran mais des mauvaises batries.", et les anaphores, comme dans "Hier j'ai acheté un nouveau téléphone. Il est le meilleur achat que j'ai jamais fait." [15] [GINDL et collab., 2013], un des défis du traitement automatique des langues (TAL).

Les principaux éléments que les règles linguistiques prennent en compte sont les interjections [OSHERENKO et ANDRÉ, 2009], la ponctuation [OSHERENKO et ANDRÉ, 2009; HUTTO et GILBERT, 2014], les intensifieurs [OSHERENKO et ANDRÉ, 2009; HUTTO et GILBERT, 2014], les conjonctions contrastives (ex. "mais") [HUTTO et GILBERT, 2014] et les négations [OSHERENKO et ANDRÉ, 2009; HUTTO et GILBERT, 2014]. LIU [2010] a formulé quatorze principales règles conceptuelles de polarité d'opinion dans une phrase qui peuvent être ensuite enrichies ou revisitées en fonction du domaine d'application. Les systèmes à base de règles sont capables de différencier des nuances de polarité. OSHERENKO et ANDRÉ [2009] détectent cinq niveaux de polarité, variant de très négatif à très positif, en passant par la polarité neutre. HUTTO et GILBERT [2014] détectent jusqu'à neuf polarités, y compris la polarité neutre. Le système décrit dans OSHERENKO et ANDRÉ [2009] contient 4 527 règles pour trouver des correspondances de mots désignant des émotions et 154 règles grammaticales. Les résultats obtenus sur le corpus des revues de films sont supérieurs à ceux de la référence statistique mais n'excèdent pas 55% en rappel ou précision ce qui illustre la difficulté de la tâche. HUTTO et GILBERT [2014] comparent leur système Vader avec des approches d'apprentissage automatique sur les corpus de tweets, des articles et des revues des produits et des films. Vader obtient le meilleur score de F-mesure de 96% sur le corpus de tweet, où le meilleur score obtenu par des méthodes d'apprentissage automatique est celui de Naïve Bayes : 84%. La clé de l'approche de HUTTO et GILBERT [2014] à base de règles est d'utiliser un dictionnaire des termes de sentiments comportant la valence des termes. La valence a été obtenue via des tâches de services collaboratifs (crowdsourcing).

MOILANEN et PULMAN [2007]; KUMAR et RAGHUVEER [2013]; SAMHA [2016] exploitent les relations de dépendance. Ils obtiennent de bons résultats sur des revues de produits.

Le système à base de règles de détection des préférences de l'utilisateur dans le texte, proposé par LANGLET et CLAVEL [2016] est un des rares à avoir été développé spécifiquement pour le contexte conversationnel humain-agent.

L'inconvénient de l'approche à base de règles est qu'elles ne peuvent pas détecter les

14. L'orientation sémantique ou la polarité d'un mot indique la direction dans laquelle le mot dévie de la norme de son groupe sémantique ou de son champ lexical [LEHRER, 1974]

15. les exemples sont tirés de [GINDL et collab., 2013]

cas où les connaissances du monde réel sont nécessaires et dépendent fortement de l'exhaustivité des dictionnaires utilisés [MOILANEN et PULMAN, 2007]. L'avantage est qu'elles ne nécessitent pas de grands volumes de corpus annotés manuellement. HUTTO et GILBERT [2014] rapportent également la vitesse d'exécution du programme supérieure à celle des systèmes à base d'apprentissage automatique.

3.2.2 Les méthodes à base d'apprentissage automatique

RAMÍREZ-TINOCO et collab. [2017] ont analysé les publications scientifiques des années 2010 - 2017 sur l'analyse de sentiments dans les réseaux sociaux. Les auteurs soulignent les difficultés liées aux corpus constitués à partir des réseaux sociaux qui comportent notamment des abréviations et des fautes d'orthographe. Sur cet aspect, leur domaine est proche du domaine du chatbot que nous étudions. RAMÍREZ-TINOCO et collab. [2017] affirment que près de 60% des travaux utilisent l'apprentissage automatique, car cela leur permet d'obtenir un meilleur score d'exactitude (Accuracy). Selon RAMÍREZ-TINOCO et collab. [2017], 30% des travaux utilisant les approches à base de dictionnaires le font afin de rendre les systèmes davantage généralisables et parce qu'elles ne nécessitent pas de corpus pré-annoté. En effet, la principale spécificité d'apprentissage automatique est que l'apprentissage s'effectue sur des données pré-étiquetées. Il existe deux principaux modèles d'apprentissage : supervisé et non-supervisé.

L'apprentissage supervisé

L'apprentissage supervisé appliqué au domaine de la détection des OPEM consiste à classifier du texte, par exemple, contenant une opinion positive ou négative. Un corpus de données annotées sert à l'entraînement du classifieur. L'étude effectuée par RAMÍREZ-TINOCO et collab. [2017] montre que les méthodes d'apprentissage supervisé les plus utilisées dans le domaine de la détection des OPEM sont SVM (68.42%), Naïve Bayes (52.63%), les arbres de décision (23.68%) et les réseaux de neurones (18.42%), comme un réseau de neurones artificiels dynamiques dans [GHIASSI et collab., 2013]. Selon la même étude, 96.77% des travaux se limitent à l'identification de la polarité sans aller plus loin.

Pour la détection des OPEM en français, les compétitions annuelles en fouille de textes reflètent bien les tendances des méthodes utilisées. Les résultats de l'édition 2007 du Défi Fouille de Texte consacré à la détection de l'opinion présentés dans [PAROUBEK et collab., 2007] montrent la popularité du classifieur SVM (la machine à vecteur de support). Mais les meilleurs résultats ont été obtenus par les équipes combinant plusieurs classifieurs. Lors du défi en 2017, dont les résultats sont décrit dans [BENAMARA et collab., 2017], bien que les méthodes des participants variaient, tous ont utilisé des approches de classification supervisée : Naïve Bayes, SVM, K plus proche voisins, les arbres de décision et les réseaux de neurones. Le meilleur résultat pour la détection de la polarité des tweets a été obtenu par le système utilisant des réseaux de neurones dont l'entrée a été initialisée avec des "sentiment embeddings" appris sur des tweets contenant des émoticônes et des termes annotés avec un dictionnaire de polarités [ROUVIER et BOUSQUET, 2017].

Un énoncé utilisateur peut contenir une ou plusieurs phrases. Afin de détecter des OPEM au niveau de l'énoncé, GALLEY et collab. [2004] utilisent des adjectifs positifs et négatifs comme l'un des descripteurs d'entrée du réseau bayésien permettant aux auteurs d'incorporer plusieurs dépendances entre des étiquettes pour la détection de l'accord ou du désaccord dans les discussions entre humains.

En ce qui concerne la détection des OPEM au niveau de la phrase, WILSON et collab. [2005] désambiguïsent la polarité des phrases en deux étapes et en utilisant un diction-

naire de plus de 8 000 entrées. Lors de la première étape, les textes sont classifiés en deux groupes : polarisé et neutre. Lors de la seconde étape, les textes polarisés sont étiquetés comme positifs, négatifs, les deux ou neutres. Les auteurs utilisent BoosTexter, un algorithme utilisant une technique d'apprentissage automatique par renforcement et permettant une classification de texte contenant plusieurs étiquettes [SCHAPIRE et SINGER, 2000]. WILSON et collab. [2009] augmentent le nombre des descripteurs utilisés dans leur approche afin de mieux désambiguïser les expressions neutres. Ils constatent que pour la majorité des algorithmes d'apprentissage automatique qu'ils testent, cela permet d'améliorer les résultats obtenus. Pour aller plus loin que la détection de la polarité de phrase, KIM et HOVY [2004] ont utilisé la détection des entités nommées pour identifier la source d'opinion. Ils ont défini la *région du sentiment* comme une région autour de la source d'opinion. Pour définir la polarité d'une phrase, ils prennent en compte la polarité des mots dans la région du sentiment. Les auteurs appliquent deux règles simples au calcul de la polarité de la région du sentiment : 1) deux polarités négatives s'annulent et 2) les mots "not" (ne...pas) et "never" (jamais) inversent la polarité. YANG et CARDIE [2014] proposent une méthode semi-supervisée permettant de prendre en compte le contexte aussi bien à l'intérieur qu'à l'extérieur de la phrase. Ils utilisent les champs aléatoires conditionnels (CRF) pour pouvoir modéliser des structures complexes des phrases et la régularisation postérieure [16] pour renforcer les CRF avec des contraintes linguistiques. Les contraintes linguistiques sont représentées dans la recherche de YANG et CARDIE [2014] par les patrons contenant des sentiments et les connecteurs discursifs, tels que "bien que", "toutefois". Ils permettent d'effectuer la classification avec le modèle CRF de façon semi-supervisée, c'est-à-dire dans les conditions des données annotées manuellement limitées. Selon l'analyse de l'état-de-l'art effectuée par CAMBRIA et collab. [2013], les approches purement statistiques sont plus performantes sur des textes plus longs qu'une phrase. PORIA et collab. [2013] proposent un système permettant de passer du niveau des mots clés au niveau des concepts en s'appuyant sur des bases de connaissance.

La détection des OPEM au niveau des caractéristiques d'une cible est surtout intéressante pour les opinions des utilisateurs sur des produits. MUKHERJEE et LIU [2012] proposent deux approches statistiques utilisant des mots d'utilisateurs comme des "semences" pour orienter la détection des caractéristiques des services et des sentiments liés à ces caractéristiques. SOCHER et collab. [2013] augmentent la barre de l'état-de-l'art de 80,0% jusqu'à 85.4% en score d'exactitude (Accuracy) pour la classification des phrases en positives et négatives. Les auteurs proposent un réseau de tenseurs neuronaux récursifs qui permet également d'améliorer l'analyse de sentiments à grain fin. Selon la revue de SUN et collab. [2017], les techniques sophistiquées sont nécessaires pour détecter les OPEM aux niveau des caractéristiques d'une cible résultant, le plus souvent, en approches semi-supervisées, comme l'algorithme combinant la maximisation de l'attente et un classifieur Bayésien [NIGAM et collab., 2000] utilisé par ZHAI et collab. [2011] parallèlement aux "connaissances du langage naturel" et non-supervisées (des expressions partageant les mêmes mots communs, la similarité lexicale basé sur WordNet [MILLER, 1998] et la corrélation positive et négative des expressions dans une phrase).

L'inconvénient de l'apprentissage supervisé est, d'une part, qu'il demande beaucoup de données annotées manuellement pour l'étape d'apprentissage. D'autre part, le modèle appris sur un corpus d'un domaine donnée est difficilement généralisable pour des données venant d'autres domaines [SUN et collab., 2017]. Malgré l'absence de la nécessité de construire des règles manuellement comme c'est le cas dans l'approche à base de règles,

16. une plateforme probabiliste pour l'apprentissage structuré et faiblement supervisé [GANCHEV et collab., 2010]

le réglage des paramètres peut s'avérer non moins fastidieux [BASARI et collab., 2013]. L'avantage de cette méthode est lorsque de grands corpus sont accessibles, elle trouve facilement des généralités permettant de classifier les données, en obtenant une bonne exactitude des résultats. Lorsque la granularité de l'OPEM à détecter n'est pas élevée, l'intervention d'un expert linguiste n'est pas nécessaire. Le temps de calcul dépend souvent de l'algorithme choisi.

L'apprentissage non-supervisé

L'apprentissage non-supervisé représente des algorithmes de classification des données en clusters selon des similarités. L'avantage principal des techniques non-supervisées est qu'elles n'ont pas besoin d'apprentissage sur un corpus pré-annoté, ce qui signifie que l'effort humain est réduit. TURNEY [2002] est un des premiers chercheurs à proposer une approche non-supervisée pour la détection des OPEM. L'auteur effectue la classification en utilisant les adjectifs et les adverbes comme des indicateurs de subjectivité. Il calcule l'orientation sémantique d'une phrase en calculant la valeur d'informations mutuelles instantanées (PMI [17]) de la phrase. TSAGKALIDOU et collab. [2011] identifient l'émotion exprimée dans un tweet en calculant la similarité sémantique entre un mot du dictionnaire des émotions et un tweet de l'utilisateur. L'intensité de l'OPEM est extraite directement d'un dictionnaire. Ensuite les auteurs appliquent un algorithme de partitionnement en k-moyennes. POPESCU et ETZIONI [2007] utilisent une technique d'étiquetage par relaxation [HUMMEL et ZUCKER, 1983] en lui fournissant des règles d'annotation des opinions et des potentiels termes d'opinion obtenus avec PMI. Le système obtient 86% en Précision et 89% en Rappel dans la tâche de classification de polarité.

BRODY et ELHADAD [2010] utilisent une allocation de Dirichlet latente (LDA) [18] [BLEI et collab., 2003] pour détecter les caractéristiques des cibles et leur influence sur les opinions exprimées par les utilisateurs dans des revues des produits. Leur travail a permis de dégager des nuances sémantiques des adjectifs en fonction de la caractéristique d'une cible. Pour certains adjectifs leur système arrive à détecter un sentiment dans le contexte des revues des restaurants, alors que des annotateurs humains les ont considéré comme objectifs et neutres, lorsqu'ils avaient une liste des adjectifs classés par thème (cuisine, services, vins, etc.). Leur approche permet également de résoudre le problème des fautes d'orthographe.

HEMMATIAN et SOHRABI [2017] analysent les performances des méthodes non-supervisées appliquées à la classification des opinions. Ils remarquent que soit ces méthodes sont rapides et frugales en mémoire, mais sensibles au bruit et produisant des résultats instables (ex. K-means), soit robustes et créant des clusters de bonne qualité mais demandent beaucoup de mémoire et de temps de calcul (algorithme agglomératif ou de division).

3.2.3 Les méthodes hybrides

Dans le but de tirer le meilleur parti des méthodes linguistique et de l'apprentissage automatique, les chercheurs créent des systèmes hybrides. C'est le cas du système de BRUN et collab. [2015] utilisant une grammaire pour la détection des termes et de la polarité des termes et SVM pour le calcul des polarités des cibles. Le classifieur permet aux auteurs d'éviter les erreurs qui peuvent être générées par la grammaire.

17. pointwise mutual information
18. LDA est un modèle générative probabiliste, permettant, par exemple, de créer une représentation d'un document. [BLEI et collab., 2003]

L'hybridation peut donc être faite entre les approches à base de règles et l'apprentissage automatique, mais aussi entre l'apprentissage automatique et un algorithme d'optimisation, par exemple. BASARI et collab. [2013] a couplé SVM avec un algorithme d'optimisation PSO [19] pour augmenter l'exactitude (Accuracy) de la classification des revues de films en positif ou négatif grâce à un meilleur choix de paramètres d'initialisation du système de SVM, car dans le cas de SVM, le choix des paramètres d'initialisation est difficile, selon les auteurs.

Pour proposer une approche alternative à l'apprentissage supervisé et réduire la dépendance du système aux données pré-annotées manuellement, APPEL et collab. [2016] combinent des règles sémantiques, telles que l'inversement de la polarité, le choix de la partie de la phrase où il faut prendre en compte le sentiment, en fonction des conjonctions (par exemple, "malgré", "sauf si"), avec des dictionnaires et des ensembles flous pour la détection des sentiments au niveau de la phrase. Les ensembles flous permettent de modéliser l'intensité de la polarité à l'image de la perception humaine, qui est subjective. L'inconvénient de la logique et des ensembles flous, souligné par l'auteur, est leur lent et faible adaptabilité aux nouvelles informations. Ce sont les règles sémantiques qui permettent à l'auteur de palier à ce problème. La méthode hybride de APPEL et collab. [2016] permet d'obtenir de bons résultats en Précision 84.24% et en exactitude (Accuracy) 88.02% sur un corpus de tweets. Les auteurs listent l'argot, les métaphores, la double négation et les paragraphes de complexité très élevée comme points problématiques pour le système.

Pour apporter une solution aux langues ayant peu de ressources linguistiques, MARTÍN-VALDIVIA et collab. [2013] proposent un système composé de plusieurs classifieurs pour la détection de la polarité dans les revues en espagnol. Ils combinent l'apprentissage supervisé et non-supervisé. Les auteurs utilisent la pondération des termes dans un document (TF-IDF) et SVM d'abord sur un corpus en espagnol et ensuite sur le corpus traduit en anglais. Par la suite, une approche non-supervisée s'appuyant sur la base de connaissances SentiWordNet [BACCIANELLA et collab., 2010] est appliquée uniquement pour le corpus en anglais. La dernière étape est l'application d'un algorithme d'empilement qui permet de choisir de meilleurs résultats à partir des trois classifications obtenues. Cette approche permet d'obtenir de meilleurs résultats que l'apprentissage supervisé et non-supervisé appliqué séparément sur le même corpus.

Selon CAMBRIA et HUSSAIN [2015], les méthodes existantes de la détection des OPEM basées sur les dictionnaires ne permettent pas de détecter des OPEM implicites communiqués par la sémantique latente. En revanche, les approches comme "Sentic computing", proposé par les auteurs, à base de concepts permettent de détecter les OPEM implicites dans des expressions grâce aux grandes bases de connaissances sémantiques. L'approche très innovante de CAMBRIA et HUSSAIN [2015] utilise des bases de connaissances et combine des réseaux de neurones et les machines d'apprentissage extrême (Extreme Learning Machine). En utilisant des règles linguistiques et les bases de connaissances, cette approche permet aux auteurs de modéliser la circulation des sentiments d'un concept à l'autre suivant les relations de dépendance entre les clauses. Malgré l'enrichissement de la base de connaissances par les experts et via une plateforme de crowdsourcing à travers des jeux sérieux, elle n'est pas exhaustive. Si ni la base de connaissances, ni les règles linguistiques ne permettent d'identifier la polarité des concepts extraits par

19. "Particle Swarm Optimization" [KENNEDY et EBERHART, 1995] : l'optimisation par essaims particulaires est un algorithme, où le système est initialisé avec des solutions aléatoires, appelées "particules". Les particules se déplacent dans l'espace vers de meilleures solutions, en augmentant leur rapidité de déplacement.

l'analyseur sémantique, l'apprentissage automatique est alors appliqué. Les connaissances du monde réel sont transmises au système d'apprentissage (les machines d'apprentissage extrême) via les concepts extrait précédemment par l'analyseur sémantique et transformés en vecteurs de phrase. Les quatre autres indices fournis en entrée sont l'information sur la polarité des concepts, lorsqu'il est possible de l'obtenir, sur l'appartenance des mots aux parties de discours, sur les dépendances dans l'arbre de dépendances et sur la présence de la négation. Les auteurs listent les limites de leur système qui sont tous liés au recours à la base de connaissances : manque de l'exhaustivité de concepts et de sens de concepts (seule l'utilisation habituelle des concepts dans la langue est répertoriée), les problématiques liés aux performances lors de l'extraction des connaissances et l'absence de l'historique de l'OPEM d'une phrase à l'autre dans le même texte.

Comme nous avons vu dans cette sous-section, les approches hybrides sont utilisées pour la détection des OPEM afin de pallier soit au manque de ressources linguistique, l'absence de grands corpus annotés manuellement, soit pour améliorer la précision et les performances des systèmes, souvent avec l'objectif de donner plus de détails sur l'intensité de la polarité de l'OPEM détecté.

3.3 Utilité des prétraitements des textes

Les prétraitements du texte, tels que la normalisation du texte ou l'annotation des catégories lexicales des mots, est une étape nécessaire pour rendre possible le fonctionnement des méthodes à base de règles. Quant aux méthodes d'apprentissage automatique, HADDI et collab. [2013]; BAO et collab. [2014]; SINGH et KUMARI [2016] démontrent également un apport tangible des prétraitements à l'amélioration des résultats de classification des textes selon leur valence (positive ou négative). Les prétraitements permettent de réduire le bruit [HADDI et collab., 2013; BAO et collab., 2014] et de gérer l'argot Internet [SINGH et KUMARI, 2016] lors de la tâche de l'analyse de sentiments. Lorsqu'il s'agit de la tâche de détection des problèmes d'interaction, le nombre de prétraitements varie du simple découpage de texte en phrases, mots et, éventuellement, les racines des mots, comme dans [KATO et SAKAI, 2017] pour la détection de rupture de dialogue, à l'annotation en parties de discours et en dépendances syntaxiques pour la détection de l'insatisfaction de l'utilisateur par l'interaction avec un chatbot comme dans [XIANG et collab., 2014].

Une des techniques de prétraitement destinée à réduire le bruit contenu dans les textes est l'utilisation d'un dictionnaire pour corriger les mots mal orthographiés. ANGIANI et collab. [2016] effectuent l'évaluation de l'influence des différentes techniques de prétraitements sur le résultat de l'analyse des sentiments dans des tweets. Une des techniques est l'utilisation de la bibliothèque externe Python PyEnchant [20] dans le but de normaliser l'orthographe des mots (corriger les mots mal orthographiés, remplacer les formes réduites des mots par leur forme complète) et remplacer les injures par un tag "injure". Selon ANGIANI et collab. [2016], ses résultats de classification supervisée (Naïve Bayes) montrent que l'utilisation du correcteur orthographique n'améliore pas le score obtenu mais augmente le temps de traitement. BEAVER et FREEMAN [2016] rapportent également le gain non significatif dans les résultat de la détection de demandes de transfert vers un conseiller humain par les utilsateurs lors de l'utilisation d'un correcteur orthographique TextBlob [LORIA et collab., 2014]. Le principe de fonctionnement de TextBlob, est de remplacer un mot inconnu par un mot connu ayant la probabilité la plus élevée et

20. http ://pythonhosted.org/pyenchant

dont "la distance d'édition est de 1-2 du mot" inconnu. Les auteurs proposent que le choix d'un correcteur orthographique prenant en compte le contexte d'une phrase permettrait d'avoir de meilleurs résultats.

Nous effectuons également des prétraitements nécessaires pour l'utilisation de l'approche à base de règles. De plus, nous évaluons l'apport d'un correcteur orthographique à la performance de notre système hybride.

3.4 Conclusion et notre positionnement

Nous avons présenté dans ce chapitre les approches pour la détection des problèmes d'interaction, de l'opinion et des phénomènes reliés aux opinions. Nous avons détaillé davantage les indices disponibles dans les interactions avec un chatbot : les indices lexicaux, sémantiques, l'historique du dialogue et également les OPEM. Notre position par rapport aux indices est d'utiliser les indices lexicaux de "haut" niveau, plus précis et plus clair en termes de compréhension humaine. En tant qu'indice sémantique, nous n'utilisons que la similarité sémantique ne nécessitant pas de bases de connaissances car ces dernières sont sous-développées pour le français. De plus, le développement de bases de connaissances demande un effort humain/temps supérieur à celui, nécessaire pour l'apprentissage non-supervisé des plongements de mots sur le corpus de dialogues. En ligne avec les travaux de XIANG et collab. [2014], nous utilisons les OPEM de l'utilisateur dont la cible est l'interaction comme un indice des problèmes d'interaction. Nous utilisons un dictionnaire d'émotions (le choix du dictionnaire est décrit dans le Chapitre 6 page 81), et les marqueurs des OPEM, présentés dans le Chapitre 4, Section 4.3.2 page 60. Nous modélisons également une relation, selon MARTIN et WHITE [2005], pour détecter la source et la cible de l'OPEM et prenons en compte la négation, comme décrit également dans le Chapitre 6 page 81. Nous modélisons le concept de l'interaction en tant que cible potentielle de l'OPEM, en constituant des dictionnaires et en nous appuyant sur les groupes syntaxiques de la phrase. Nous décrivons la détection de la cible potentielle à la page 96. La détection de la relation est décrite dans la Section 6.2.3 page 95.

En ce qui concerne le choix de l'approche, nous avons vu que lorsque le phénomène n'est pas encore suffisamment étudié, les chercheurs ont tendance à appliquer soit des approches à base de règles, soit des algorithmes, déduisant des règles pour mieux comprendre le phénomène étudié. Nous avons vu que l'apprentissage automatique supervisé est le plus utilisé pour la détection des problèmes d'interaction. Ors, cette approche demande un corpus conséquent annoté manuellement.
La situation est pratiquement similaire pour la détection des opinions et des phénomènes reliés aux opinions. Les réseaux de neurones montrent de très bons résultats pour la détection des opinions et des phénomènes reliés aux opinions lorsqu'il existe un grand corpus avec des annotations détaillées pour les apprendre. Ce n'est pas le cas du français. Les méthodes hybrides sont encore peu répandues mais peuvent apporter des solutions lorsque peu de ressources linguistiques sont disponibles.

Nous optons donc pour une méthode hybride : nous combinons une approche à base de règles avec une approche d'apprentissage non-supervisé des représentations des mots. Nous choisissons l'approche à base de règles linguistiques pour :

— mieux étudier le phénomène des problèmes d'interactions dans les dialogues humain-machine en français ;

— proposer un système ne nécessitant pas un grand corpus annoté manuellement.

L'apprentissage non-supervisé des représentations des mots sert à représenter la sémantique contextuelle des mots sous forme de vecteurs. Cette approche permettrait de tirer avantage d'un grand corpus non-annoté que nous avons en notre possession. Notre hypothèse est que la complétion d'une méthode à base de règles par une méthode à base d'apprentissage non-supervisé des représentations des mots comblerait le manque d'informations sémantiques de la première approche et réduirait le manque de précision de la seconde. Nous étudierons également l'apport d'un correcteur d'orthographe à la performance du système.

Deuxième partie

Partie 2 : Corpus, l’annotation et la stratégie de l’annotation

Résumé

Nous étudions les problèmes d'interaction entre un humain et un agent, ainsi que nos méthodes pour les détecter sur le corpus de tchat écrit entre une conseillère virtuelle et ses utilisateurs. Le corpus nous a été fourni par l'entreprise EDF. Le corpus étant privé, nous comparons ses caractéristiques avec des corpus de "conversation écrite" décrits dans la littérature. La comparaison avec des corpus de conversations entre humains révèle des similitudes de langage des utilisateurs. Cela nous indique que le langage des utilisateurs de notre corpus correspond au français tchaté.

Le corpus est analysé afin de déterminer la pertinence de la détection d'opinion dans un corpus humain-agent et les difficultés que ce type de texte peut contenir pour le traitement automatique de texte. L'analyse morphosyntaxique des énoncés utilisateur prouve que les utilisateurs mènent des conversations naturelles avec l'agent virtuel. De plus, les analyses des corpus de conversations avec des agents virtuels montrent que les utilisateurs communiquent avec la machine en exprimant leurs émotions. Un des moyens pour les utilisateurs de transmettre leur état émotionnel est la ponctuation, l'argot Internet (ex. lol) et l'utilisation exagérée des majuscules. Le texte des utilisateurs contient également des termes d'émotion que nous avons détectés en appliquant un dictionnaire. Ce qui confirme la pertinence de notre approche par la détection de l'opinion de l'utilisateur. Les conversations en ligne sont également caractérisées par un niveau élevé de fautes d'orthographe, ce qui influence les performances d'une analyse syntaxique automatique.

Afin de constituer une ressource de référence pour évaluer notre système, nous avons tenu une campagne d'annotation. L'annotation du corpus implique un choix mesuré des étiquettes. Nous avons proposé une taxonomie des problèmes d'interaction reflétant notre approche du point de vue de l'opinion de l'utilisateur. Nous distinguons des problèmes d'interaction implicites et explicites. La taxonomie a servi de base pour la constitution du guide d'annotation. Après une étude des techniques de représentation de l'information dans les guides d'annotation existants, nous avons choisi de guider l'annotateur par des questions, des définitions, des explications, des résumés d'information, des exemples et des contre-exemples et un arbre de décision. Les annotations ont été effectuées sous le logiciel GATE.

Chapitre 4

Le corpus des interactions écrites en français avec un chatbot

Sommaire

Dans le cadre de notre recherche, nous utilisons un corpus de tchat écrit humain-agent qui nous a été fourni par l'entreprise EDF, afin de tester la validité de la méthodologie que nous avons choisie. Il est également important pour nous de mesurer si l'analyse d'opinions et d'émotions est applicable au tchat humain - conseillère virtuelle.
Dans les sections suivantes nous présentons des corpus de « conversation écrite » [1] : des corpus décrits dans la littérature et le corpus en notre possession. D'abord, nous comparons leurs principales caractéristiques. Ensuite, nous présentons plus en détail notre corpus sous deux angles : sa pertinence pour la détection d'opinion et d'émotions et la difficulté que le langage utilisateur du corpus représente pour le traitement automatique des langues (TAL).

4.1 Corpus existants de la «conversation écrite»

Pour pouvoir mieux positionner notre corpus par rapport aux corpus existants de tchat et plus largement de « conversation écrite », afin de mieux le caractériser, nous allons présenter ici des corpus de « conversation écrite » soit entre humains, soit entre un humain et un chatbot. Nous nous limitons ici aux corpus existants en français puisque nous avons en notre disposition un corpus en français. Tout d'abord nous allons présenter deux corpus de conversations avec un agent virtuel. Le corpus décrit dans [EFRAIM et MOREAU, 2016] est un corpus de dialogues des utilisateurs avec des agents virtuels appartenant à des entreprises de domaines divers. Pour faciliter la référence à ce corpus, nous l'appellerons ici un corpus de dialogues des Agents Virtuels Orientés Domaine (AVOD). Le second corpus de ce type est le corpus HALPIN, orienté recherche documentaire et décrit par ROUILLARD et CAELEN [1998]. Pour pouvoir faire des parallèles entre le langage des utilisateurs dans les tchats avec un agent conversationnel et le langage dans les tchats entre humains, nous présenterons ensuite trois corpus de conversations entre humains :

— le corpus de français tchaté constitué par FALAISE [2005], orienté domaine libre,

— le corpus de tchat humain-humain issu d'un centre de contact d'assistance de l'entreprise Orange et analysé dans le cadre du projet DATCHA par DAMNATI et collab. [2016],

— et enfin, le corpus WebGRC qui est un corpus de messages de forums concernant l'activité EDF décrit dans [DUTREY et collab., 2012; DUTREY, 2011].

4.1.1 Corpus des conversations avec des agents virtuels

Les deux corpus de conversations avec des agents virtuels proviennent des agents virtuels qui sont conçus pour la recherche d'informations. Ainsi, **le corpus AVOD** est constitué des énoncés d'internautes extrait des dialogues avec dix-huit agents virtuels, présents sur les sites web de structures aussi bien privées que publiques (ex. agence de voyage, mairie, etc.). Selon ses auteurs, il contient "79 698 requêtes" utilisateurs, chacune correspondant à un tour de parole. Le nombre de tours de parole adressé à un agent virtuel varie fortement en fonction des agents (entre 64 et 20 000). 14% des requêtes utilisateur sont complexes [2] et 12,7 % des énoncés sont hors-sujet. 75% des énoncés contiennent moins

1. un terme utilisé par M. Marcoccia pour décrire le mode de communication au sein de forums et tchats [MARCOCCIA, 2000]

2. une requête complexe est celle "qui contient plus d'un acte de langage fondamental d'information ou de demande" [EFRAIM et MOREAU, 2016]

de 10 mots. Le corpus contient 645 637 mots. Le nombre moyen de mots utilisateur est inférieur à dix.

Le corpus HALPIN est un peu particulier car il représente une collection de dialogues écrits humain-machine créés lors d'un jeu. Le but du jeu est de trouver un livre dans le catalogue d'une bibliothèque. Malgré les conditions préétablies de l'expérience en laboratoire, le langage des requêtes des utilisateurs reste un langage naturel dont les caractéristiques nous intéressent. Le corpus HALPIN est composé de 897 dialogues. Le nombre de tours de parole est influencé par la configuration de l'expérience. En moyenne, un dialogue en contient 7. La Table 4.1 présente les statistiques disponibles des deux corpus. La différence entre les domaines couverts et la configuration des agents crée de fortes différences entre le nombre de formes uniques et le nombre de tours de paroles, mais le nombre d'erreurs d'orthographe reste équivalent, ce qui indique une complexité similaire des traitements automatiques.

Statistique	**Corpus AVOD (domaines multiples)**	**Corpus HALPIN (recherche de documentation)**
Nombre de formes uniques	25 500	527
Nombre de tours de parole moyen utilisateur par dialogue	-	7
Nombre maximum de tours de parole utilisateur par dialogue	-	63
Nombre de tours de parole moyen de l'agent par dialogue	-	6
Nombre de mots moyen par énoncé utilisateur	<10	-
Nombre maximum de tours de parole de l'agent par dialogue	-	54
Erreurs d'orthographe/mots	10,38% (sur l'échantillon de 8051 mots)	-
Erreurs de frappe/mots	-	13,09%

TABLEAU 4.1 – Statistique des corpus de conversations avec des agents virtuels.

EFRAIM et MOREAU [2016] constatent que le corpus que nous appelons ici AVOD, contient des marques d'expressivité telles que des majuscules et des signes de ponctuations multiples. Les auteurs les suppriment lors des prétraitements car le corpus est destiné à la recherche d'information. ROUILLARD et CAELEN [1998], au contraire, font une analyse du langage utilisateur assez poussée du point de vue comportemental. Ils observent sur des exemples tirés du corpus HALPIN que les utilisateurs perçoivent la machine comme un vrai interlocuteur. Les auteurs ont relevé des cas où les utilisateurs partagent avec elle leurs émotions, utilisent des expressions impératives, adaptent leur vocabulaire à celui de la machine, partagent leurs connaissances avec elle et même mettent en question les informations qu'elle leur fournit. Malheureusement, les auteurs ne donnent pas plus de précisions sur les caractéristiques linguistiques du langage de l'utilisateur.

Nous retiendrons de cette sous-section que les corpus des dialogues humain-agent se caractérisent par un langage naturel de l'utilisateur. Les utilisateurs font des énoncés plutôt courts. En moyenne, 12% des mots produits par les utilisateurs sont affectés par les fautes d'orthographes. Les utilisateurs communiquent souvent avec l'agent virtuel comme avec un humain et lui expriment leurs émotions.

4.1.2 Corpus des conversations entre humains

Pour pouvoir représenter les caractéristiques du langage utilisateur en fonction de son interlocuteur : humain ou agent virtuel, nous présentons dans cette section les corpus des conversations en ligne entre humains. Les trois corpus diffèrent en style : tchat généraliste Vs. tchat formel et tchat Vs. forum de discussion.

Le corpus de français tchaté collecté et décrit par FALAISE [2005] est un corpus de discussions sur des tchats aussi bien généralistes que spécialisés (programmation, politique, etc.) mais qui ne représentent toutefois pas un service et ne sont pas spécialisés dans une marque précise. Le langage des utilisateurs est donc informel. Le corpus du français tchaté contient "4 192 033 messages, couvrant environ 3 mois de conversations sur 105 canaux de tchat." C'est un grand corpus de 23 011 876 mots. Le nombre de mots par énoncé est de 5,5 mots en moyenne. Le nombre de formes différentes de mots dans le canal #18-25ans, d'où proviennent la majorité des messages, constitue 7% des mots du corpus.

FALAISE [2005] note également des spécificités qu'il qualifie de syntaxe de l'oral, par exemple, la topicalisation. Il met également en parallèle le découpage des messages en propositions dans les tchats et le découpage en groupes prosodiques dans le langage parlé. Il constate que 2/3 des mots sont correctement orthographiés.

Le corpus de tchat entre les clients et les agents de l'entreprise **Orange** est plus proche de notre corpus que le corpus de français tchaté, car lié à un domaine d'entreprise. Il se caractérise par un langage formel des échanges. Le tchat client - conseiller humain a la particularité de permettre à chacun des participants d'envoyer plusieurs messages lors de son tour de parole, ce qui n'est pas le cas des conversations avec un chatbot. Le corpus consiste en 276 conversations, 8 455 messages et 94 244 mots. En moyenne, il y a 10 tours de parole par conversation et 1,4 messages par tour de parole. La longueur des conversations peut être expliquée par le fait que les clients d'Orange contactent l'assistance surtout dans les cas de problèmes techniques qui peuvent demander des manipulations à effectuer. Selon DAMNATI et collab. [2015], décrivant la diversité lexicale du corpus, les clients produisent, en moyenne, 8,6 mots par message et l'agent un peu plus : 13,2 mots. G. Damnati et ses collègues expliquent la longueur des messages des agents par le fait que les agents ont la possibilité d'inclure automatiquement des phrases dans les dialogues lorsqu'il s'agit de fournir au client des instructions afin de résoudre un problème. Les clients produisent des énoncés plus courts mais avec plus de mots différents par rapport aux conseillers. Les mots différents représentent 12% des mots produits par les utilisateurs mais seulement 5% des mots des agents. La distribution des catégories syntaxiques entre le client et le téléconseiller (TC) est présentée dans la Figure 4.1.

DAMNATI et collab. [2016] ont aussi effectué une analyse du langage des interlocuteurs. 39,78% des messages des clients contiennent une erreur d'orthographe. Les agents en font beaucoup moins (15,38% des messages) ce qui est lié à leur professionnalisation. Seulement 8% des messages des utilisateurs ont moins d'1% de mots mal orthographié ce qui indique un niveau très élevé de fautes d'orthographes. L'analyse effectuée par les

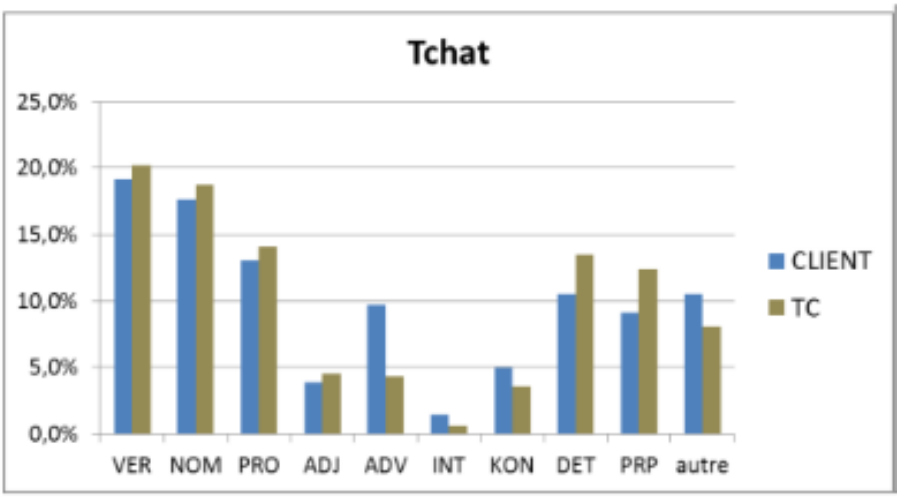


FIGURE 4.1 – Représentativité des parties de discours dans le corpus Orange. Figure tirée de [DAMNATI et collab., 2015]

auteurs du corpus sur les types d'erreurs d'orthographe a montré que ce sont des erreurs qui entravent l'analyse syntaxique automatique.

DAMNATI et collab. [2016] ont également analysé la présence des marques d'expressivité telles que la ponctuation et l'utilisation des majuscules dans le corpus Orange. Le résultat de l'analyse a montré que 44% des conversations contiennent au moins un indice d'expressivité dans les messages des clients. Ainsi, 12% des conversations contiennent des messages où le client a utilisé des majuscules soit pour insister sur une partie de sa phrase, soit pour exprimer sa déception ou incompréhension. 13% des conversations contiennent des points d'exclamation qui pour la plupart "ont une polarité négative". 13,76% des conversations contiennent au moins un message d'un utilisateur contenant des points de suspension exprimant la déception et même l'exaspération. Le corpus contient très peu d'émoticônes et ils sont majoritairement positifs.

Comme nous l'avons vu, le corpus d'Orange, ayant le style du langage davantage formel, contient moins de caractéristiques du langage de français tchaté. Les éléments présents restent néanmoins importants du point de vue de l'information émotionnelle.

Le corpus WebGRC est beaucoup plus petit et consiste en 2 970 discussions qui représentent 18 492 messages uniques. Les messages ont été collectés sur trois forums de discussion et concernent uniquement EDF. Les auteurs ont délibérément choisi uniquement les forums où les utilisateurs s'exprimaient sur la thématique de leur satisfaction vis-à-vis de l'entreprise. Malheureusement, les auteurs ne fournissent que cette information quantitative concernant leur corpus. Malgré la différence en taille et en sujets, le corpus du français tchaté et le corpus WebGRC sont proches par leurs caractéristiques du langage utilisateur. Ainsi, les spécificités lexicales du langage de tchat suivantes sont présentes dans les deux corpus : émoticônes, abréviations, graphies phonétiques, logogrammes, anglicismes, fautes de frappe et d'orthographe. DUTREY et collab. [2012] proposent également une typologie très détaillée des spécificités rédactionnelles du corpus WebGRC.

Nous retiendrons de cette sous-section que les corpus des dialogues entre humains se caractérisent par le langage français tchaté dont les caractéristiques sont moins pré-

sentes dans le cadre formel des dialogues entre un conseiller et un client. Les trois corpus présentés contiennent tous des fautes d'orthographe et des marques d'expressivité telles que des émoticônes ou la ponctuation multiple. La longueur des énoncés utilisateur est équivalente à celle des utilisateurs des tchats humain-agent virtuel.

4.2 Présentation du corpus Laura

Dans le cadre de notre recherche, nous disposons de données de tchat écrit humain - agent virtuel. Ces données ont été produites dans les conditions "in-the-wild" et collectées en interne à EDF. La taille du corpus est supérieure à celle des corpus présentés ci-dessus. Comme le corpus AVOD, les dialogues de notre corpus sont orientés vers le domaine de l'entreprise. Dans ce qui suit, nous adopterons les termes "*corpus LAURA*" pour désigner l'ensemble des données extraites des conversations entre les internautes et l'agent virtuel Laura ; "utilisateur" pour un internaute utilisant l'interface de l'agent virtuel Laura.

Le corpus contient l'ensemble des interactions entre les utilisateurs et l'agent virtuel LAURA collectés entre janvier 2014 à mai 2015. La conseillère virtuelle du site de l'entreprise EDF répond aux questions des utilisateurs concernant soit la navigation sur le site *EDF Particulier*, soit les services EDF. Un dialogue se compose a minima d'une paire adjacente (PA) qui, à son tour, contient un énoncé utilisateur et un énoncé agent. Le corpus de l'entreprise EDF a été anonymisé. Toutefois il n'est pas distribué et reste la propriété de l'entreprise.

L'une des particularités de notre corpus est l'apparition d'énoncés utilisateur semi-automatisés. Ils représentent 23% des énoncés utilisateur. En effet, l'agent virtuel LAURA a été configuré pour suggérer à l'utilisateur des listes de réponses adaptées sous forme de liens Internet en relation avec les thématiques abordées. Ainsi, l'utilisateur peut sélectionner une réponse en adéquation avec sa requête. Cette sélection sera considérée comme étant son propre énoncé. Dans ce cas-là, les spécificités relatives au tchat sont réduites : fautes de frappes, erreurs orthographiques, émotions, etc. sont inexistantes. Il est difficile de distinguer un énoncé rempli semi-automatiquement car il ne comporte pas d'hypertexte. La TABLE 4.2 donne les principaux chiffres qui nous permettent de caractériser le corpus LAURA. Le nombre d'énoncés utilisateur est égal à celui de l'agent pour former des paires adjacentes. Toutefois, certains énoncés utilisateur et agent sont vides. Dans les énoncés utilisateur cela signifie qu'aucun texte n'a été saisi. Pour l'agent, lorsque l'utilisateur est redirigé vers une page Internet, le système de chatbot remplit l'énoncé de l'utilisateur avec le lien correspondant. Ce type d'énoncé utilisateur ne demande pas de réponse de la part de la conseillère virtuelle et donc n'est pas suivi de réponse. La ligne de l'énoncé agent existe, mais elle est vide dans ce cas-là. 3% des énoncés utilisateur contiennent plus d'une phrase. Lorsque l'énoncé de l'agent ou de l'utilisateur contient qu'un lien Internet ou lorsque l'énoncé utilisateur ne comporte que la ponctuation, y compris les émoticônes, nous considérons que l'énoncé en question ne contient aucun mot. Le nombre de participants correspond à celui des dialogues, tout en sachant que nous ne pouvons pas identifier si la même personne intervient plusieurs fois dans d'autres discussions avec l'agent virtuel LAURA.

Le corpus a été découpé en plusieurs sous-corpus. Nous décrirons ici d'abord le sous-corpus "LauraDev" qui a servi pour la découverte des caractéristiques du corpus et le développement des premières règles de détection des problèmes d'interaction. Les deux sous-corpus annotés manuellement en problèmes d'interaction seront présentés dans la Section 6.1.3 page 84 et la Section 7.1 page 112.

Statistique	Nombre total	Moyenne	Min	Max	Déviation standard	Médianne
Dialogues	1 813 934					
Paires adjacentes	6 046 695					
Paires adjacentes par dialogue		3,33	1	153	2,23	3
Énoncés utilisateur, total	6 046 695					
Énoncés utilisateur non-vides	5 108 694					
Énoncés agent, total	6 046 695					
Énoncés agent non-vides	6 045 099					
Énoncés agent contenant une URL	1 836 822					
Nombre de mots utilisateur	30 130 016					
Nombre de mots utilisateur par énoncé non vide		6	0	193	6,83	4
Nombre de mots agent	168 109 350					
Nombre de mots agent par énoncé non vide		27,81	0	278	28,6	20

TABLEAU 4.2 – Statistique du corpus LAURA.

4.3 Statistiques descriptives

Pour le corpus de développement **"LauraDev"**, nous avons choisi des interactions du mois de janvier, avril, juillet et octobre de l'année 2014, pour avoir une distribution homogène des exemples de dialogues sur une année. Ce corpus contient 628 228 dialogues. Le nombre de dialogues par mois varie de 268 441 dialogues en janvier à 153 627 en avril. La croissance du nombre de dialogues en janvier est liée à la réception de la facture annuelle par les clients de l'entreprise ce qui crée de nombreuses interrogations de leur part. Le tableau 4.3 page 58 permet de voir qu'en dépit de l'existence des valeurs extrêmes, les dialogues sont en moyenne courts et contiennent peu de mots de l'utilisateur. Ces informations du corpus LauraDev correspondent bien aux caractéristiques moyennes pour le corpus Laura.

Dans cette section, nous nous intéressons à la qualité du corpus du point de vue de sa pertinence vis-à-vis de la tâche de la recherche d'opinion. Nous nous intéressons tout particulièrement aux usages de l'utilisateur du chatbot. Nous analysons les énoncés utilisateur et comparons les résultats aux corpus présentés dans la Section 4.1. Nous analysons également les difficultés potentielles que ce corpus humain-agent peut comporter pour un traitement automatique. La majorité des analyses présentes dans cette section ont fait l'objet de la publication [MASLOWSKI, 2016].

	Nb Interactions	**Nb mots User**	**Nb mots Laura**	**Durée dialogue**
min	1	0	0	00 :00 :00
1er décile	2	0	0	00 :00 :09
1er quartile	2	0	8	00 :01 :40
2ème quartile	3	3	18	00 :10 :31
3ème quartile	4	7	35	00 :11 :09
9ème décile	6	12	77	00 :12 :58
max	153	160	274	4 jours 22 :24 :29

TABLEAU 4.3 – Caractéristiques des dialogues. On notera que des erreurs dans le traitement de la session utilisateur sont la cause de plusieurs valeurs aberrantes induisant des durées de dialogue de plusieurs jours.

4.3.1 L'utilisateur, tchate-t-il?

Afin de vérifier la pertinence de la recherche d'opinions et d'émotions au sein du corpus Laura, nous vérifions que les énoncés utilisateur représentent des phrases suffisamment bien construites. Nous avons analysé les énoncés utilisateur en les comparant dans un premier temps avec les énoncés agent et dans un second temps avec des données issues des tchats entre les humains. Ensuite, nous avons analysé les annotations en parties de discours effectuées par l'outil TreeTagger [SCHMID, 1994].

La comparaison des énoncés utilisateur avec des énoncés agent ont montré que le vocabulaire des utilisateurs est très focalisé car leurs phrases sont synthétiques : « mon chauffage est collectif que dois-je renseigner? » (voir TABLE 4.4).

Enoncés	**Nombre moyen de mots par énoncé**	**Formes de mots distinctes/nombre de mots total**
Enoncé utilisateur	6	1%
Enoncé agent	30	0,02%

TABLEAU 4.4 – Caractéristiques des énoncés.

Selon FALAISE [2005], les utilisateurs de tchats entre humains ne font pas de phrases plus développées. Comme nous l'avons vu dans la Section 4.1.2, le nombre moyen des mots dans leurs énoncés est équivalent (5,5). En revanche, les énoncés utilisateur du corpus Orange sont en moyenne un peu plus long (8,6 mots en moyenne). Ce détail peut être lié au fait que les utilisateurs doivent décrire l'action qu'ils accomplissent lors de l'assistance technique. En continuant la comparaison des énoncés utilisateur du corpus Laura-Dev avec le corpus décrit par FALAISE [2005], nous remarquons que le nombre des formes distinctes est approximativement 10 fois plus restreint que dans le canal #18-25ans du corpus du français tchaté. Cela peut s'expliquer par la finalité des interactions qui est, sauf exception, d'obtenir rapidement des informations et non de soutenir une discussion comme dans le cas d'un tchat entre deux humains sur un canal thématique, même s'il arrive qu'ils échangent des informations. FALAISE [2005] explique « ... la finalité du tchat demeure le dialogue, et les tchateurs sont soucieux, jusqu'à un certain point, de la clarté de leur messages. » Le fait que le langage des utilisateurs du corpus Orange est plus riche, ne contredit pas notre conclusion. Le but de la conversation des clients Orange est la résolution de leur problème technique et la présence d'un humain "au bout du fil" les

encourage à donner le plus de détails possibles pour aider le conseiller humain. L'agent virtuel au contraire, encourage l'utilisateur à s'exprimer par des phrases courtes : "Votre phrase est trop longue, pouvez-vous être plus concis."

Néanmoins, pour s'assurer que les utilisateurs ne réduisent pas leurs énoncés à des mots clés nous procédons à l'analyse morphosyntaxique. Après avoir annoté le corpus avec le logiciel TreeTagger [SCHMID, 1994], nous avons calculé le nombre d'occurrences des étiquettes syntaxiques. Le résultat obtenu pour les étiquettes syntaxiques "NOM", "VER", "ADV" et "ADJ" a été comparé avec celui de DUTREY et collab. [2012] sur le corpus WebGRC "brut" et avec le corpus Orange. Il s'avère que la distribution, entre les corpus WebGRC et LauraDev, des quatre types d'étiquettes syntaxiques choisies est comparable (voir FIGURE 4.2 page 59). La différence entre le nombre des étiquettes de corpus LauraDev et le corpus WebGRC (p-value = 1) et LauraDev et le corpus Orange (p-value = 0, 08) n'est pas significative (p-value est supérieur à 0,05).

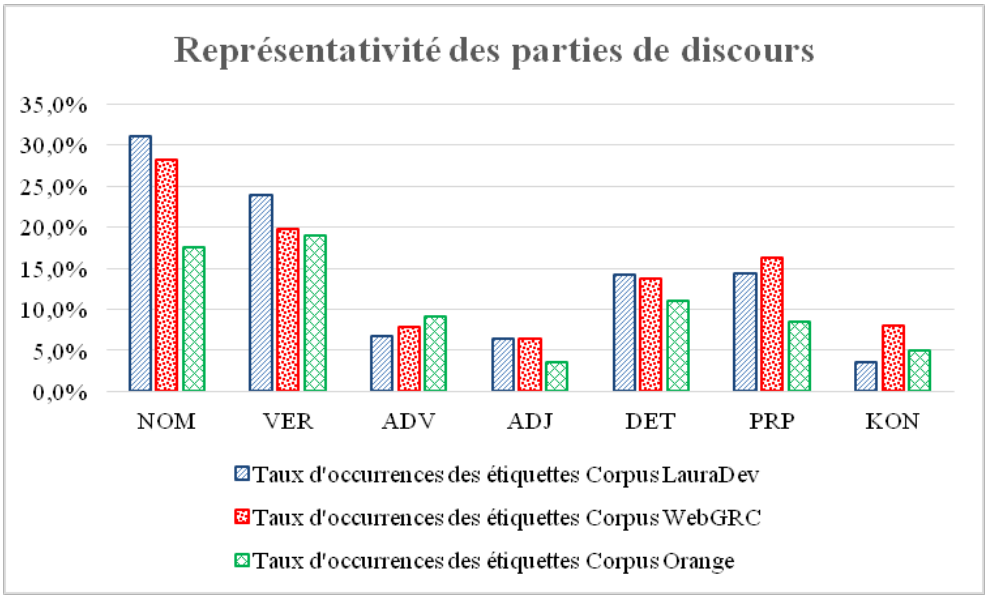


FIGURE 4.2 – Représentativité des parties de discours dans les corpus LauraDev, WebGRC et DATCHA (Orange)

Ce fait et le nombre relativement élevé d'occurrences de mots-outils (le taux d'occurrences d'étiquettes "DET" et "PRP" s'élève par exemple à 15%), nous permet de déduire que le langage des utilisateurs de l'interface de l'agent virtuel Laura est suffisamment proche du langage naturel pour ne pas être considéré comme une simple requête dans un navigateur Internet et semble ainsi plus propice à l'expression d'opinion. Les utilisateurs du corpus Orange utilisent plus de verbes et d'adverbes, et moins de noms et d'adjectifs que dans les deux corpus EDF. Cela peut-être également lié à la description d'actions liées au problème rencontré avec le matériel. La faible présence d'adjectifs peut être également liée au style formel des messages. Dans le corpus LauraDev les deux styles sont présents. Il arrive d'ailleurs que certains clients tentent d'utiliser l'interface de Laura pour soumettre leurs e-mails. De plus, les utilisateurs s'adressent souvent à la conseillère virtuelle comme à une vraie personne, comme dans l'exemple suivant « Bonjour Madame, pourriez-vous m'expliquer, je vous prie, le montant de ma dernière facture ? ». Où dans un style informel, lorsqu'il concerne des conversations hors-sujet : "quel âge as tu".
Ainsi, l'analyse en parties de discours de la structure du texte nous parait favorable pour effectuer une recherche au sein de notre corpus de lexique émotionnel et affectif.

4.3.2 Réalisations explicites de l'opinion dans le corpus

Nous utiliserons l'acronyme OPEM pour désigner les termes OPinion et EMotion afin de rassembler tous les phénomènes liés à l'opinion. Les marqueurs d'OPEM peuvent nous permettre de détecter des opinions et des phénomènes reliés aux opinions négatives dont la cible [MARTIN et WHITE, 2005] est l'interaction. Par conséquent, nous nous intéressons, dans un premier temps, aux marqueurs spécifiques du web, tels que les caratères échos, les interjections et les émoticônes, et, dans un second temps, aux marqueurs lexicaux.

Caractères échos, interjections et émoticônes

Il existe des marqueurs qui permettent d'exprimer des OPEM dans les textes écrits sans pour autant les exprimer explicitement par des mots descriptifs. Cela peut être l'utilisation exagérée de majuscules [YATES et collab., 1993], d'onomatopées [ANIS, 2003], la répétition de caractères [PANCKHURST, 2006], des sigles comme « lol » (version anglaise) ou « mdr » (version française) [LORENZ et MICHOT, 2012], et de façon plus non-verbale : la ponctuation multiple [DUTREY et collab., 2012] et les émoticônes [MARCOCCIA et GAUDUCHEAU, 2007], [LORENZ et MICHOT, 2012]. Les onomatopées et la répétition de caractères simulent la prononciation de sons [ANIS, 2003; PANCKHURST, 2006]. Les sigles « lol » et « mdr » expriment le rire et sont utilisés en parallèle par les utilisateurs français [LORENZ et MICHOT, 2012]. Les majuscules successives permettent à l'utilisateur soit d'accentuer un mot dans une phrase : "modification du calendrier de paiement de plus de 80 %, POURQUOI,;;;;", soit d'imiter un cri : "je me suis abonné par téléphone hier"/"JE ME SUIS ABONNE PAR TELEPHONE HIER". Ce dernier cas est similaire à la tendance remarquée chez des utilisateurs des systèmes vocaux : leur effort vocal s'intensifie lorsque le système ne les comprend pas [SHRIBERG et collab., 1992].

Nous avons tout d'abord étudié l'usage de ces types de marqueurs potentiels d'OPEM : (1) la ponctuation multiple et les caractères échos, (2) les interjections et (3) les émoticônes. Nous nous appuyons sur la définition de COUGNON [2015] décrivant (1) comme un phénomène qui « consiste en une répétition de caractères à valeur d'expressivité, d'intensité, mais également en vue d'apporter du son ». Nous avons considéré les signes "!" et "?" volontairement dupliqués au moins 2 fois d'affilée comme une ponctuation multiple. Nous avons mis en place pour les caractères échos une méthode permettant de détecter les répétitions de plus de 2 caractères. Nous avons utilisé la liste disponible sur Internet [3] pour (2). Nous nous appuyons sur la définition de (3) de DRESNER et HERRING [2010] [4] et MARCOCCIA et GAUDUCHEAU [2007] [5] et utilisons le dictionnaire fourni par la société DataGenetics [6]. Ce dictionnaire contient 2 242 émoticônes classés par ordre de fréquence décroissante. La FIGURE 4.3 page 61 présente la répartition de ces différents types de marqueurs dans notre corpus. Notre corpus comporte en moyenne 0,04 marqueur par énoncé utilisateur. Le marqueur le plus représenté (68%) est la ponctuation multiple mais l'usage de ce marqueur reste cependant marginal car seul 1% des énoncés

3. http ://www.aidenet.eu/grammaire28.htm et http ://www.aproposdecriture.com/wp-content/uploads/2014/06/Liste-des-onomatop%C3%A9es.pdf

4. "The term "emoticons" - a blend of "emotion" and "icons" - refers to graphic signs, such as the smiley face, that often accompany textual computer-mediated communication (CMC)."DRESNER et HERRING [2010] [Le terme "émoticônes" - mélange d'émotion et d'icônes - fait référence à des signes graphiques, tels que le smiley, qui accompagnent souvent la communication textuelle assistée par ordinateur.]

5. Les smileys sont "des moyens de pallier l'absence de face à face" dans "la communication médiatisée par ordinateur (CMO)". Ce sont "des signes de ponctuation expressive". MARCOCCIA et GAUDUCHEAU [2007]

6. http ://www.datagenetics.com/blog/october52012/index.html

utilisateur comprend une ponctuation multiple.

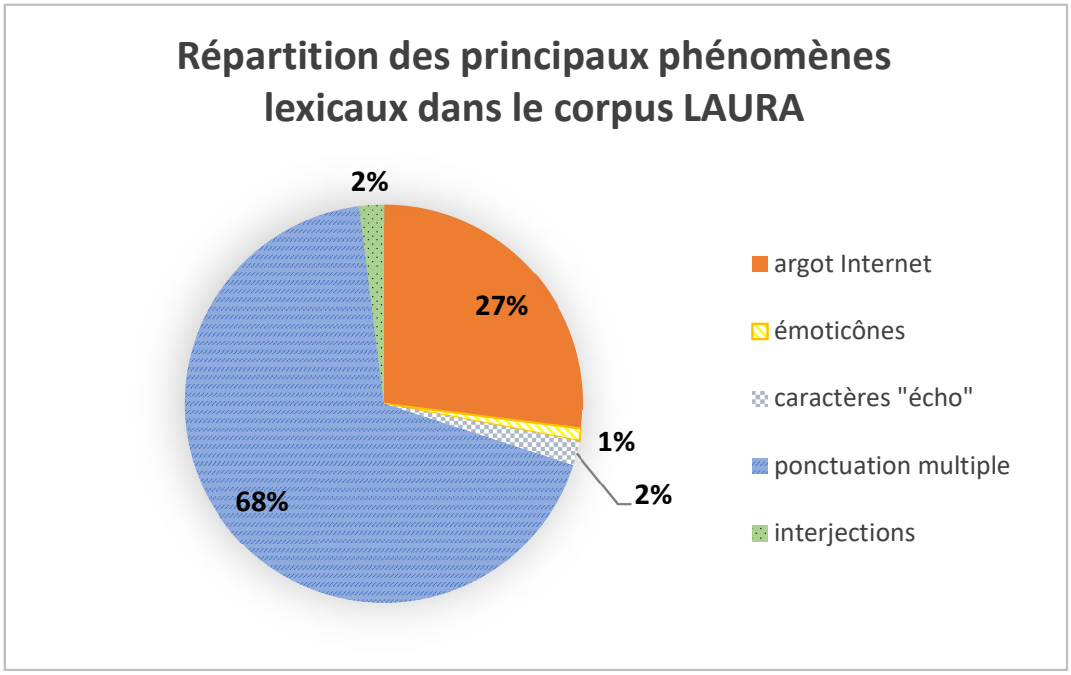


FIGURE 4.3 – Répartition des principaux phénomènes lexicaux parmi ceux détectés dans le corpus Laura

Parallèlement à la ponctuation “écho”, nous avons aussi détecté 4 451 formes de mots contenant le phénomène de caractères « écho » parmi lesquelles 40% sont des formes distinctes. La majorité des cas est issue de fautes de frappes, comme le mot « comm-ment » (212 occurrences dans le corpus). La seconde forme la plus fréquente de caractères “écho” de notre corpus est l’onomatopée “pff” (157 occurrences). On la retrouve sous une trentaine de formes différentes (« Pfff », « ppffff », « pfffffffffffffffffff », etc.), dont l’inten-sité nous renseigne sur le degré émotionnel que l’utilisateur souhaite communiquer. Ce phénomène a été aussi retrouvé dans d’autres termes : « merciiiiiiiiiii», « mmddddr ». Il est à noter que ces répétitions de caractères se trouvent souvent sur des interjections qui peuvent être porteuses d’OPEM assez précises, comme « pfff » et « zut », exprimant une réaction négative ou étant porteuse d’ironie (“pfffffffffffffffffff mon dieu il est beau le progrés”). Notre corpus contient 3 892 occurrences d’émoticônes soit moins de 0,001% d’émoticônes par énoncé utilisateur et couvre 3% des émoticônes présents dans la liste de référence. L’usage des émoticônes semble donc marginal dans ce corpus mais leur étude reste importante pour notre objectif final d’analyse des opinions et des émotions de l’uti-lisateur. Les émoticônes “standards” comme " :)", ";)" et " :(" sont les plus populaires chez les utilisateurs de l’agent virtuel Laura (« :)» 34% d’émoticônes, «;) » 10% ; « :(» 9%). Notons également l’utilisation assez fréquente de l’émoticône "<3" par les utilisateurs de l’agent virtuel Laura (7%). Cette émoticône est un marqueur de familiarité et hors du contexte de la relation “client-conseiller” attendue (“tu devrais changer de coupe de cheveux <3”). Nous trouvons aussi l’aspect “hors contexte” dans des énoncés utilisateurs ouvertement

agressifs tels que décrits par DE ANGELI et CARPENTER [2005]. En nous reposant sur la typologie de la relation entre les émoticônes et le contenu verbal de MARCOCCIA et GAUDUCHEAU [2007], nous avons procédé à une analyse plus approfondie de l'usage des émoticônes pour identifier différents types de relations entre le contenu textuel et les émoticônes présents dans notre corpus : l'information redondante (« ok merci beaucoup et bonne journée Laura :-) »); l'aide à la compréhension (« Bonjour, j'aimerai savoir comment payer ma facture sur internet :) »); l'atténuation (« Laura!!! fait un effort ma poulette!!!! :) »); l'ironie («c'était pourtant clair :-) »; la familiarité : « vous avez quel âge? :) »). La prise en compte du contexte de plusieurs énoncés d'un dialogue peut permettre la désambiguïsation du sens des émoticônes, ce que ne permet pas une approche utilisant des « sacs de mots » [FERRARI et collab., 2008]. 4% d'énoncés utilisateur contiennent au moins un marqueur, y compris l'argot Internet détaillé dans la Section 4.3.3. Il est possible que la rareté des phénomènes décrits ci-dessus renforce leur significativité dans un énoncé. Bien que marginale, la présence de ces phénomènes scripturaux dans notre corpus souligne l'appartenance de ce dernier au langage Web. Nous modélisons ces phénomènes lors de la conception de nos règles décrites dans la Section 6.1.4.

Lexique d'émotions dans le corpus

Les énoncés utilisateur peuvent aussi contenir des mots décrivant les OPEM envers l'interaction. Pour pouvoir les détecter nous avons choisi d'utiliser un dictionnaire d'OPEM. Parmi ceux disponibles en langue française, nous avons choisi d'utiliser la version française du dictionnaire pour LIWC [PIOLAT et collab., 2011b] qui contient des radicaux de mots classés par catégories thématiques. La version française de dictionnaire LIWC comporte approximativement 4 500 radicaux de mots. Il a été constitué sur la base du schéma PANAS, Positive And Negative Affect Schedule [WATSON et collab., 1988], et des dictionnaires de mots communs. Les mots sont divisés en 5 catégories, subdivisées en plusieurs sous-catégories liées aux processus psychologiques. Ce sont non seulement des catégories contenant des mots exprimant des émotions mais aussi des catégories grammaticales. Toutes les catégories sont regroupées sous cinq catégories principales : les processus linguistiques et psychologiques, préoccupations personnelles, dimensions du langage oral et ponctuation [PENNEBAKER et collab., 2007].

Nous avons choisi de nous focaliser sur une sous-catégorie de la catégorie processus psychologiques en lien avec notre problématique : la sous-catégorie « processus affectifs ». Cette dernière comprend elle-même deux sous-catégories : émotions positives et émotions négatives (à nouveau subdivisée en Anxiété, Colère et Tristesse). Une analyse préliminaire a montré que certains termes du vocabulaire sont très ambigus, surtout dans le contexte « métier » de notre corpus. Par exemple, le mot « puissance » qui fait partie de sous-catégories « affect » et « émotion positive» figure dans le contexte suivant : « déterminer la puissance à souscrire? » qui ne porte aucune coloration émotionnelle. Nous prenons en compte cette spécificité dans les règles pour le système final décrit dans la Section 6.1.4.

Pour étoffer notre analyse préliminaire, nous combinons le lexique d'OPEM avec quelques règles basiques d'inversion de polarité. En utilisant ces règles, nous annotons le corpus en deux catégories : positive (EmoPos) et négative (EmoNeg), par exemple “c est pas grave” comporte une annotation EmoPos et “J'ai fais tout ça, j'ai reçu un premier mail je me suis identifier et maintenant ça ne marche plus!” : une annotation EmoNeg. Ces annotations nous permettent d'avoir une idée préliminaire sur la couverture du corpus Laura-Dev par le lexique en dépit des ambigüités non-résolues. Ainsi, environ 25% de dialogues

du corpus contiennent une annotation Emo*[7] : 15% d'EmoPos et 10% d'EmoNeg. Nous considérons ce résultat comme prometteur pour l'utilisation du dictionnaire LIWC pour détecter les OPEM.

4.3.3 Difficultés pour le traitement

La détermination du type de langage des utilisateurs utilisé dans le corpus est essentielle pour prévoir les difficultés que nous rencontrerons lors de traitements plus complexes comme l'annotation automatique des termes exprimant des opinions et des émotions. Issu du tchat sur Internet, notre corpus porte des caractéristiques du langage français tchaté, telles que décrit par FALAISE [2005] : les "émoticons, les abréviations, une graphie phonétique, des allongements vocaliques qu'on préfère appeler ici "caractères écho", phénomènes phonético-graphiques ("2main" pour "demain") et sémantico-graphiques ("Micro$oft" pour "Microsoft"), onomatopées, xénismes (plus souvent des anglicismes), des noms d'utilisateur et des fusions de mots". L'argot Internet défini par MOREAU [2001] comme « un mélange de mots anglais, de sigles, d'onomatopées et d'abréviations, d'orthographe phonétique et d'expressions détournées. », peut représenter un défi lors de l'analyse automatique du texte [DUTREY et collab., 2012].

Pour mesurer l'importance de sa présence au sein de notre corpus, nous appliquons la liste de 490 termes de l'argot Internet de Wiktionnaire[8] sur notre corpus. 47% des formes d'argots du Wiktionnaire sont représentées dans le corpus LauraDev, comme les abréviations « svp » pour « s'il-vous-plaît » et « bjr » pour « bonjour » ou « kwa », forme phonétique de « quoi ». En prenant en compte l'ensemble des caractéristiques du langage utilisateur décrit dans cette section et dans la Section 4.3.2 (voir Figure 4.3), nous pouvons conclure qu'il appartient au type de langage défini par ANIS [2003] et plus récemment par MAHRER [2017] comme le langage "écrit parlé" caractéristique des tchats en ligne.

En outre, le corpus contient un grand nombre de fautes de frappe et d'orthographe. DUTREY et collab. [2012] ont démontré que les déviances orthographiques influent sur les performances de taggeurs morphosyntaxiques lors de l'annotation en parties de discours (POS) et la lemmatisation. En ce qui concerne les performances de TreeTagger [SCHMID, 1994] sur notre corpus, 12% de mots sont étiquetés comme `<unknown>`. NASR et collab. [2016] soulèvent également le problème du traitement du tchat client-agent humain dans le corpus Orange. Ils démontrent l'étendue de l'influence d'un corpus bruité sur les performances d'un taggeur morphosyntaxique obtenant normalement des résultats d'annotation en POS correspondant à l'état-de-l'art.

RAMOS [2005] démontre également l'utilité des indices contenus dans le langage de tchat pour la recherche d'opinion et d'émotion. Il semble donc plus pertinent de considérer les termes d'argot comme une information à conserver, puisque leur forme (« pb » pour problème, « lol ») pourrait transmettre une information supplémentaire sur l'état émotionnel de son auteur, plutôt que comme un bruit, car non conforme aux « normes » dictionnairiques.

Malgré les inconvénients que l'argot Internet crée pour la reconnaissance correcte des mots lors de leur annotation morphosyntaxique, le défi de données collectés "in-the-wild" représente un réel intérêt puisque les entreprises doivent faire face à ce type de données dans le cadre de leurs activités. Du point de vue scientifique, les données de tchat humain - agent sont encore peu étudiées. L'étude du comportement des outils existants tels qu'un correcteur orthographique sur ce type de données apporte de nouvelles

7. A cet étape, nous ne prenons pas encore en compte la cible des OPEM.
8. http ://fr.wiktionary.org/wiki/Annexe :Liste_de_termes_d%E2%80%99argot_Internet

connaissances à la communauté TAL.

Nous pouvons envisager les perspectives suivantes liées au besoin d'amélioration des annotations morphosyntaxiques :

— expérimenter les approches proposées pour l'annotation automatique des parties de discours dans des langues peu dotées mais proches de langues dotées, par exemple, pour annoter en parties de discours un texte en Irlandais, LYNN et collab. [2014] appliquent une approche inter-langagière par un transfert directe. Cette méthode consiste en l'utilisation d'un modèle déléxicalisé de l'arborescence de la langue source pour apprendre un modèle d'analyse pour la langue cible.

— utiliser les approches proposées pour l'annotation en parties de discours des textes des apprenants du français. THOUËSNY [2009] propose d'utiliser des dictionnaires et des règles pour corriger des annotations de TreeTagger. L'auteur établit également des références croisées entre les étiquettes correctes et les étiquettes erronées.

4.4 Conclusion

Au sein de ce chapitre nous avons présenté des corpus appartenant au genre "conversation écrite". Le corpus "LAURA" que nous utilisons dans notre étude est un corpus de tchat écrit entre une conseillère virtuelle présente sur le site web de l'entreprise EDF et ses utilisateurs. Nous l'avons comparé aux corpus connus par la communauté scientifique afin de le positionner. Malheureusement, les corpus des conversations avec des conseillers virtuels ou humains décrits ici ne sont pas en accès libre puisqu'ils appartiennent à des entreprises privées. Cependant, ces comparaisons ainsi qu'avec des corpus de tchat ouverts permettent de mieux cerner les caractéristiques de notre corpus. Le tableau 4.5 page 65 résume les caractéristiques quantitatives disponibles pour la majorité des corpus décrits dans cette section.

Nombre de/ Pourcentage de	**Corpus**					
	Laura (texte utilisateur)	**AVOD**	**HALPIN**	**Français tchaté**	**Orange**	**WebGRC**
Dialogues	1 813 934	-	897	-	276	2 970
Mots	30 130 016	645 637	-	23 011 876	94 244	1 635 743
Mots différents/mots totals	0,6%	4%	527	7%	12%	-
Messages	-	-	-	4 192 033	8 455	18 492
Enoncés utilisateur	6 046 695	79 698	-	-	-	-
Tours de parole par dialogue (moyenne)	3	-	7	-	10	6
Mots par énoncé utilisateur (moyenne)	6	<10	-	5,5	8,6	-
Erreurs d'orthographe ou de frappe/mot	>12%	10,38%	13,09%	33%	10%	4%
Ponctuation multiple	3%	présent	-	-	13%	14%
Émoticônes	0,06%	présent	-	19%	très peu	1%

TABLEAU 4.5 – Caractéristiques principales des corpus de la "conversation écrite"

Le corpus "LAURA" est le plus grand corpus de ceux présentés dans ce chapitre. Ces corpus appartiennent au type de données "in-the-wild" et ont le même style de langage utilisateur correspondant au français tchaté (sauf corpus HALPIN, collecté dans le cadre d'une expérience en laboratoire). Ce type de données contient des indices et un lexique d'opinion et d'émotion. De plus, il pose un défi pour les outils de prétraitement des données textuelles telles que les tagueurs morphosyntaxiques. Nous nous sommes concentrées sur les données de tchat en français, mais il existe également des travaux dans d'autres langues, par exemple pour l'espagnol [RAMOS, 2005] ou l'anglais [EISENSTEIN, 2013].

Chapitre 5

Stratégie d'annotation

Sommaire

Dans le chapitre précédent nous avons présenté le corpus LAURA et ses spécificités, selon l'étude du sous-corpus LauraDev. Un corpus de référence, un sous-corpus du corpus Laura, sera présenté dans le chapitre 7 page 111. Dans le présent chapitre nous décrivons la stratégie de son annotation.

La stratégie d'annotation du corpus englobe trois éléments principaux : le choix des informations à représenter sous forme d'étiquettes, décrit dans la Section 5.1, la constitution du guide d'annotation et le protocole d'annotation. Ces éléments visent à fournir suffisamment d'éléments pour l'évaluation des performances du système d'annotation automatique vis-à-vis de son objectif final. Les annotations manuelles doivent être produites de façon la plus impartiale et cohérente possible. Nous décrivons notre stratégie de constitution du guide d'annotation dans la section 5.2, en expliquant la relation entre la typologie que nous proposons et les annotations et la méthode de représentation des instructions à suivre pour l'annotateur. La dernière section 5.3 indique les éléments de déroulement de la campagne d'annotation.

5.1 Taxonomie des problèmes d'interaction

Suite à l'étude des typologies existantes des problèmes d'interaction (voir la Section 2.1), au choix de l'approche de détection des PI centrée sur l'utilisateur et l'étude du corpus, nous proposons une nouvelle taxonomie des problèmes d'interaction (voir la Figure 5.1).

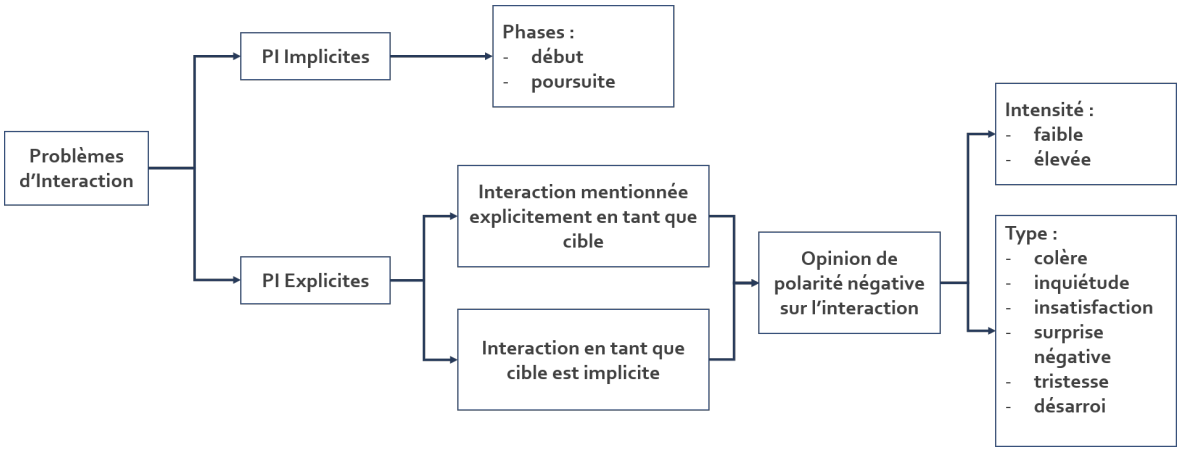


FIGURE 5.1 – Taxonomie des problèmes d'interaction

La taxonomie nous permet de distinguer des problèmes d'interaction explicites PIE (une expression de l'émotion ou de l'opinion négative de l'utilisateur à l'égard de l'interaction) et implicites PII (d'autres indices linguistiques : les répétitions de l'utilisateur, la demande de contact avec un conseiller humain de la part de l'utilisateur ou des demandes de renseignements du type "comment tu fonctionnes ?"). Les PIE sont représentés par une relation composée d'un triplet : source - opinion - cible. La modélisation de l'opinion ou de l'émotion de l'utilisateur est basée sur la théorie de l'évaluation de MARTIN et WHITE [2005]. Nous avons choisi ce modèle en fonction de l'analyse des approches existantes exposées dans [CLAVEL et CALLEJAS, 2016]. Nous rappelons que nous utilisons l'acronyme *OPEM* qui désigne **OP**inion et **EM**otion afin de rassembler tous les phénomènes liés à l'opinion. L'interaction en tant que cible, peut être mentionnée par l'utilisateur *explicitement* (e.g. "tu es virtuelle, tu ne peux pas m'aider") ou *implicitement* (e.g. **Agent :** "Veuillez

m'excuser, je n'ai pas compris ce que vous venez de dire." **User :** "pffff").

La taxonomie est prévue pour pouvoir approfondir les détails du phénomène de problème d'interaction. Elle prévoit la possibilité de distinguer des PI implicites (PII) selon la phase d'avancement du problème : début et poursuite. Suite à l'étude de l'état-de-l'art des typologies de l'opinion et des émotions (voir la Section 2.2), nous avons choisi dans le cas des problèmes d'interaction explicites (PIE), de distinguer l'intensité de l'OPEM et le type d'émotion négative, lorsque cela est possible. L'intensité est représentée par deux valeurs : faible et élevée. Nous proposons de détecter six types d'émotions négatives : la colère, l'inquiétude, l'insatisfaction, la surprise négative, la tristesse ou le désarroi. L'insatisfaction étant une émotion par défaut, lorsqu'il n'est pas possible de préciser un autre type d'émotion. Dans la section suivante nous présentons la réalisation de cette taxonomie dans un guide d'annotation manuelle.

5.2 Stratégie de constitution du guide d'annotation

Pour pouvoir évaluer notre système, nous avons besoin d'un corpus de référence, non encore utilisé pour le développement, annoté manuellement en problèmes d'interaction. Dans le cadre de la campagne d'annotation manuelle, nous avons conçu un guide d'annotation. Il est important de noter que le nombre de phénomènes à annoter prévu par ce guide est plus important que ceux que le système est capable de détecter. En effet, il est intéressant pour une entreprise d'avoir un corpus de référence riche en annotations pour pouvoir le réutiliser pour différentes tâches. Ainsi, le guide prévoit l'annotation de l'OPEM ayant pour cible non seulement l'interaction mais aussi les produits et services de l'entreprise.

5.2.1 Le choix du type d'annotation

Il existe différents types d'annotations en fonction de la nature des phénomènes à annoter, leur granularité et l'ordre de réalisation des annotations. Nos choix sont contraints par notre approche pour la détection des problèmes d'interaction centrée sur l'opinion de l'utilisateur, la taxonomie des problèmes d'interaction que nous avons proposée et le logiciel GATE que nous utilisons pour effectuer les annotations manuelles. Pour l'annotation des OPEM, nous adaptons la démarche préconisée par le standard EARL défini par le réseau d'excellence HUMAINE [1] qui utilise simultanément l'approche catégorielle et dimensionnelle pour l'annotation des émotions. L'approche catégorielle s'applique lorsque nous identifions le type de l'OPEM, correspondant aux émotions. L'approche dimensionnelle est pertinente dans notre cas pour l'identification de l'intensité faible ou forte d'une OPEM.

Pour simplifier la tâche de l'annotateur et rendre le processus d'annotation davantage fiable, nous avons également eu recours à l'annotation séquentielle : dans un dialogue l'annotateur annote d'abord tous les cas de relations, c'est-à-dire, celles, dont la cible est l'interaction (les problèmes d'interaction explicites) et celles, dont les cibles sont les produits et les services, et ensuite les problèmes d'interaction implicites. Pour éviter des chevauchements des frontières des annotations et assurer la consistance des annotations, nous conseillons à l'annotateur de commencer l'annotation de la relation par ses éléments : OPEM, sa cible et sa source.

1. http://emotion-research.net/projects/humaine/earl/schemadesign

5.2.2 Le choix des étiquettes et de leurs frontières

En prenant en compte la typologie des problèmes d'interaction, nous avons proposé les étiquettes suivantes :

— <probleme-interaction> pour annoter les problèmes d'interaction implicites avec un attribut "début", lors de la première occurrence ou "poursuite", lors des occurrences suivantes, permettant de tenir compte des phases d'avancement des PI implicites ;

— <relation> pour annoter les relations : source - OPEM - cible. Les relations correspondent aux PI explicites et les opinions et les phénomènes reliés aux opinions exprimés envers les produits et les services. Les frontières de cette étiquette sont identiques à celles de l'OPEM. L'attribut de cette annotation est "id" représenté par un chiffre qui doit être identique pour tous les éléments d'une relation et unique au niveau d'un dialogue ;

— <OPEM> avec des attributs "id", identique à l'attribut "id" de la source, cible et relation correspondantes, une polarité positive ou négative, l'intensité élevée ou faible et le type. Ce dernier attribut est optionnel. Sa valeur peut être "surprise négative", "tristesse/regret", "colère", "inquiétude/anxiété", "désarroi/désemparé". Les termes supplémentaires qui ne font pas partie de la taxonomie, ont été rajoutés pour éclaircir la signification des valeurs pour l'annotateur ;

— <source> avec un seul attribut "id" ;

— <cible> pour l'annotation des cibles de type produit ou service, toujours explicites, et dont l'attribut unique est également "id" ;

— <cible-interaction> pour annoter les expressions faisant une référence explicite à l'interaction. L'attribut de cette annotation est l'"id". Lorsque l'interaction en tant que cible est implicite, l'étiquette est absente.

Dans la littérature la position et les frontières des étiquettes correspondant à l'OPEM et aux problèmes d'interaction varient, en fonction de l'approche choisie (voir le Chapitre 2 page 11 et le Chapitre 3 page 29 de l'Etat-de-l'art), d'un document aux expressions, de l'énoncé agent, à l'énoncé utilisateur.

Dans le cadre de notre recherche, les frontières de l'annotation "relation" ont été déterminées par plusieurs facteurs. Premièrement, le logiciel GATE ne permet pas d'effectuer l'annotation discontinue d'un phénomène. Il existe pourtant des cas où la cible de l'OPEM est à l'extérieur de l'énoncé utilisateur courant. Nous choisissons néanmoins de réaliser l'annotation avec GATE pour conserver la compatibilité entre les formats des annotations automatiques et des annotations manuelles. Deuxièmement, visant l'annotation de l'OPEM à grain fin, la simplification du processus d'annotation et la compatibilité des annotations manuelles avec celles du système, nous avons choisi de réduire les frontières de l'annotation <relation>[2] à celle de <OPEM>, puisque <OPEM> est le seul élément obligatoire d'une relation. Le lien entre les éléments de la relation est fait grâce aux identifiants uniques ("id") au sein d'un dialogue. Pour les mêmes raisons de simplification et de rapidité d'annotation manuelle, lorsqu'une cible-interaction est implicite, la relation ne comporte aucune information concernant la cible. Cela est possible grâce au fait que la cible produits/service est, elle, toujours explicite.
Quant à l'annotation des problèmes d'interaction implicites, elle est réalisée au niveau de **l'énoncé utilisateur**. Ce choix est fait en fonction de la nature de ce type de problème :

2. Lors de l'évaluation, nous considérons que les frontières de la relation sont identiques aux frontières de l'énoncé utilisateur.

il n'est souvent pas possible de dissocier des éléments plus petits, comme, par exemple, lorsqu'on considère la répétition comme l'un des indices des PI. La présence d'un PI implicite sur un énoncé utilisateur n'exclut pas la possibilité de la présence d'un ou plusieurs PI explicites sur le même énoncé utilisateur.

5.2.3 Les techniques de représentation de l'information dans le guide d'annotation

La conception d'un guide d'annotation est une tâche très importante puisque la qualité des annotations en dépend. Il est important de présenter les informations de façon claire et brève à l'annotateur.

En nous inspirant des guides d'annotation des émotions pour le français de GIANOLA [2014] et des concepts métier de LAGARDE et PERADOTTO [2013], nous avons d'abord conçu un arbre de décision sur la base de notre taxonomie des PI. L'arbre de décision complet contient six niveaux de profondeur, vingt-deux branches dans sa largeur maximale et quatre sorties. Cet arbre est trop complexe pour être utilisé directement par un annotateur car difficile à visualiser mais il nous a aidé à structurer le guide. Le résumé de cet arbre est présenté dans la Figure 5.2 page 71.

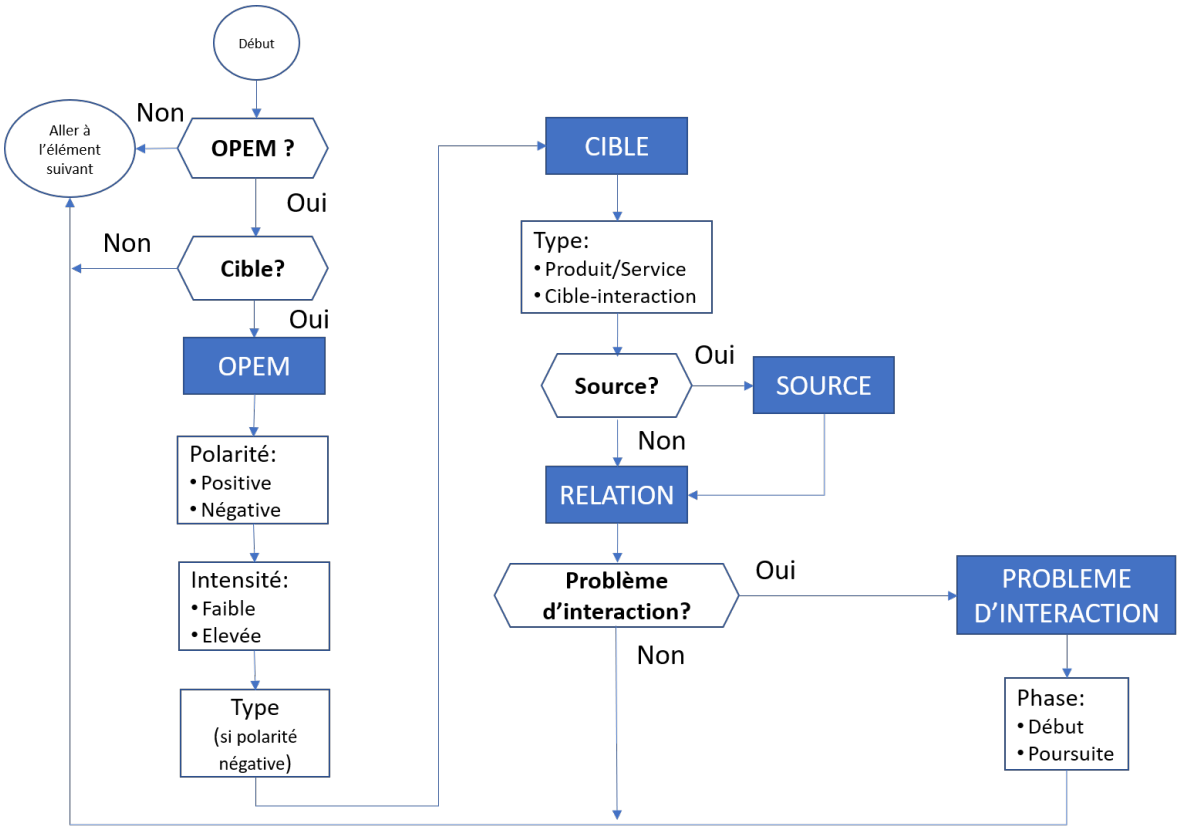


FIGURE 5.2 – L'arbre de décision à l'usage de l'annotateur pour l'annotation de l'énoncé utilisateur (extrait du guide d'annotation)

Pour bien guider l'annotateur et éviter les biais dans le processus de l'annotation, cet arbre de décision simplifié est complété par des questions à l'instar de LANGLET et CLAVEL [2015] et des résumés d'information. Les questions proposées à l'annotateur sont de deux types : celles qui guident l'ordre de l'action, par exemple "La source de l'OPEM est-elle

clairement exprimée?" et celles visant à clarifier l'aspect à annoter, par exemple "Quelle est la source de l'OPEM exprimée?". Les questions du second type sont formulées d'une manière plus intuitive et sont suivies d'explications comportant des exemples. Le guide d'annotation ne présente pas à l'annotateur les règles de détection que nous utilisons dans notre système pour ne pas influencer son jugement. Pour chaque type d'étiquette nous proposons des exemples et des contre-exemples. Le cas des répétitions des énoncés utilisateur permet d'illustrer les contre-exemples : cette répétition représente souvent un indice du fait que l'utilisateur n'a pas obtenu une réponse souhaité et donc d'un problème d'interaction implicite. Néanmoins, l'utilisateur peut également recourir à la répétition afin d'accéder au menu des propositions de l'agent virtuel et faire un choix différent du précédent. Dans ce cas, à chaque fois l'utilisateur clic sur un lien différent proposé par l'agent.
De même que chez PAROUBEK [2016], le surlignage en couleur indique l'emplacement et les frontières de l'annotation attendue.

Comme dans les guides d'annotation de la campagne DEFT 2015 [3] et du projet ANR uComp [4], nous proposons à l'annotateur une classe générique d'émotions (qui correspond à un type d'OPEM) et, lorsque cela est pertinent nous lui proposons une ou des classes spécifiques pour aider à la compréhension (voir le Tableau 5.1 page 73). Nous proposons également des définitions simplifiées ou des explications accompagnées d'un exemple.

Le guide d'annotation final comporte 27 pages et 5 pages d'annexe listant le vocabulaire métier. Le logiciel GATE [CUNNINGHAM et collab., 2011] a été utilisé pour effectuer les annotations manuelles afin de les rendre compatibles avec les annotations du système développé également sous GATE. Pour limiter l'erreur humaine, nous avons choisi un mode "bloqué" pour le schéma d'annotation. Ce mode ne permet pas à l'annotateur d'introduire de nouveaux types d'annotation, ne faisant pas partie du schéma proposé.

3. https://deft.limsi.fr/2015/guideAnnotation.fr.php?lang=fr

4. https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwjQp5mE8fTWAhVI6xQKHelQATgQFggnMAA&url=https%3A%2F%2Fnouba.dsi.cnrs.fr%2Ffournisseur%2Fpu_pck_sys_download.p_download%3Fls_file_var%3DF_225618058%2Fguide_annotation_ose.pdf&usg=AOvVaw1VLGpMuhGHq4hNvzcL_wxJ

ÉTIQUETTE GÉNÉRIQUE	CLASSE SPÉCIFIQUE	DÉTAILS	EXEMPLES
Colère		une forte insatisfaction	« 1 honte cette taxe CTA »
Inquiétude	inquiétude/ anxiété	l'utilisateur n'est pas rassuré vis à vis du manque d'information	« bonjour, je viens de payer ma facture par CB au téléphone, mais je n'ai toujours pas reçu de mail de confirmation de paiement. »
Insatisfaction		type par défaut	Utilisateur : « je n'arrive pas à me connecter » Laura : « Bonjour Madame N; Si vous avez rencontré un problème au moment de la connexion, il peut s'agir d'un problème technique. Nous faisons le maximum pour vous garantir le minimum de désagréments. » Utilisateur : « ce serait bien agréable de le savoir car voilà deux fois que me l'on change mon mot de passe. Pas très satisfaite »
Surprise négative		l'utilisateur ne s'attendait pas à ce qui s'est produit	« je suis surpris de l'augmentation de mes kw »
Tristesse	tristesse/ regret/ déception	l'utilisateur est déçu par ce qui s'est produit mais cela n'a pas d'un impact fort sur lui	« c'est dommage que vous arrêtez le tarif T, je l'aimais bien »
Désarroi	désarroi/ désemparé	l'utilisateur est désespéré et demande à l'aide. Degré supérieur à l'inquiétude	« j'ai perdu mon mot passe mais en plus il me dit que mon identifiant et faux je ne comprends pas ...j'ai bloqué mon compte. aider moi merci d'avance. »; « je suis une mère isolée avec deux enfants. Mon électricité est coupée, quoi faire? »; « quoi faire après avoir reçu une relance injustifiée de facture impayée? »

TABLEAU 5.1 – Types de l'OPEM négative (termes émotionnels). Le tableau tiré du guide d'annotation qui prévoit l'annotation de l'OPEM vers deux types de cibles : l'interaction et les produits et services

5.3 Protocole d'annotation

Les contraintes de budget et de qualité nécessaire pour annoter le corpus ont influencé le choix de l'annotateur humain. Un seul annotateur a été recruté. Cet annotateur est un annotateur-sémiologue expérimenté notamment dans l'annotation des corpus métier de l'entreprise EDF. Une session de test a été organisée pour permettre à l'annotateur de se familiariser avec le logiciel et le guide d'annotation. Cela a permis d'affiner les définitions de types des OPEM de manière à ce que leur signification soit claire pour l'annotateur.

Pour résoudre les incertitudes éventuelles de l'annotateur sur la manière d'annoter l'énoncé utilisateur, nous avons introduit dans notre schéma d'annotation l'étiquette "indéfini". A l'instar de SADOUN [2016], nous avons ensuite organisé plusieurs sessions de discussions entre l'annotateur, l'encadrante de la thèse côté entreprise et l'auteur de cette thèse. Le but de ces sessions était majoritairement d'écarter les annotations d'OPEM là où l'énoncé client n'en comporte pas, par exemple : "Je n'arrive pas à me connecter." Dans cet énoncé utilisateur, on peut supposer que la situation décrite peut créer une opinion négative chez l'utilisateur. Nous avons mis en garde l'annotateur de ne pas extrapoler des OPEM lorsque le texte est factuel et le contexte n'est pas suffisant pour prouver le contraire. Dans d'autres cas, l'incertitude portait surtout sur la frontière des annotations. La frontière des annotations n'est à l'heure actuelle pas utilisée pour l'évaluation de notre système, puisque le système effectue les annotations des PI au niveau de l'énoncé, mais pourra être employée lors de recherches futures.

La Figure 5.3 illustre le processus d'annotation.

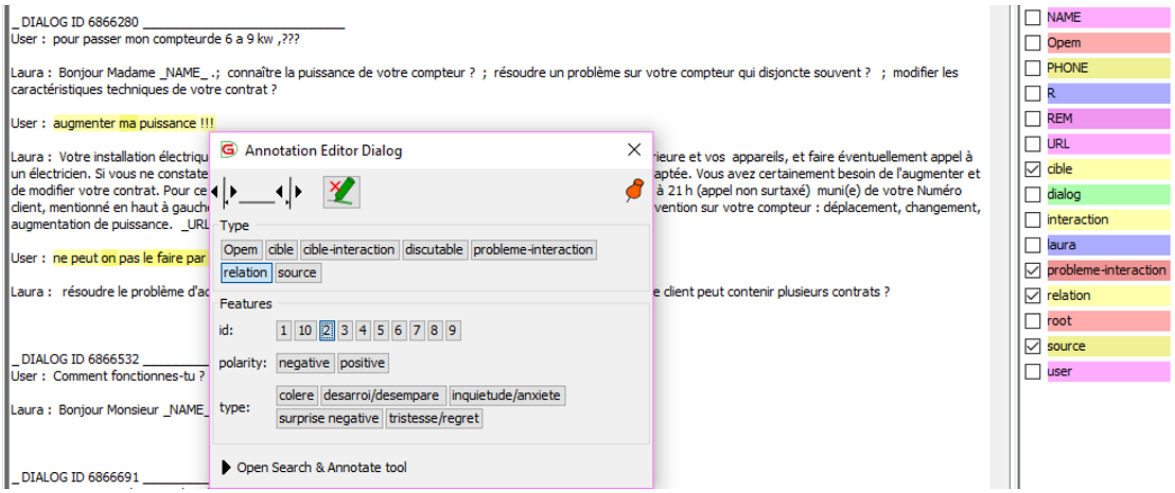


FIGURE 5.3 – Copie d'écran du GATE en cours de l'annotation

Analysons la difficulté de la tâche d'annotation, en nous basant sur les critères d'estimation proposés par FORT et collab. [2012]. Ils proposent six critères : la distinction, la délimitation, l'expressivité, la dimension de l'ensemble des étiquettes, le degré de l'ambiguïté et le contexte. La distinction des unités à annoter est complexe car des OPEM, autant que l'interaction en tant que cible peuvent être exprimées par un mot ou par une expression. Cette difficulté est augmentée par le fait que la quantité d'unités à annoter dans le texte est réduite par rapport au nombre des dialogues sans problèmes d'interaction. Les pré-annotations automatiques des tours de parole et des méta-données peuvent à la fois aider l'annotateur mais également le déconcentrer. La difficulté liée à la définition des frontières est rattachée au choix des mots devant faire partie de l'expression de l'OPEM

ou celles mentionnant la cible. Le choix d'annotation de la relation uniquement au niveau de l'OPEM allège cette difficulté. Néanmoins, l'analyse faite a posteriori montre que l'annotateur a dû rarement changer les frontières des annotations. Le niveau de l'expressivité du langage de l'annotation peut être considéré comme moyen puisque l'annotateur doit non seulement choisir une étiquette mais aussi relier des étiquette avec une relation. Nous avons restreint l'ensemble des étiquettes proposées à l'annotateur pour réduire la complexité de la tâche d'annotateur. Comme nous l'avons vu dans l'état-de-l'art, l'annotation des problèmes d'interaction est hautement ambiguë, car cette une tâche très subjective. De plus, la part du contexte lors des décisions prisent par l'annotateur est importante puisqu'il doit lire l'ensemble de dialogue, se référer au guide d'annotation et à la liste du vocabulaire métier pour annoter les cibles "produits-services". Selon les retours de l'annotateur sur les difficultés qu'il a rencontré lors de l'annotation, différencier les problèmes d'interaction implicites des problèmes d'interaction explicites est le point le plus complexe. L'annotateur a participé à deux sessions d'annotation. Il a annoté deux corpus de la même taille qui proviennent de la partie des données du corpus Laura que nous n'avons pas étudié préalablement. Ces deux corpus sont constitués de manière aléatoire. Le premier corpus annoté nous sert de corpus de développement. Nous le décrivons dans le chapitre 6.1.3 page 84. Le corpus qui a servi à évaluer notre système est décrit dans le chapitre 7 page 111.

5.4 Conclusion

Notre positionnement vis-à-vis des définitions des problèmes d'interaction et leurs typologies dans les travaux de l'état-de-l'art, ainsi que l'analyse du corpus présenté dans le Chapitre 4 page 51, nous amènent à proposer une typologie des problèmes d'interaction. Cette typologie vise à évaluer l'efficacité du déroulement d'un dialogue du point de vue de l'utilisateur. L'hétérogénéité de la nature des phénomènes constituant les problèmes d'interaction représente une difficulté majeure aussi bien lors de la création d'un schéma d'annotation, que pour l'annotateur, par exemple, les expressions des OPEM par rapport aux répétitions. Il est donc nécessaire de faire des concessions qui augmentent la difficulté des traitements automatiques afin de diminuer le travail de l'annotateur humain. Ainsi nous ne mentionnons pas la cible implicite lors des annotations, pour faciliter le travail de l'annotateur humain (cf. la Section 5.2.2 page 70).

Les annotations manuelles obtenues pour le corpus de référence, nous permettront d'évaluer les performances de notre système, décrit dans le Chapitre 7. Les annotations manuelles du corpus de développement permettent également de percevoir la distribution des problèmes d'interaction dans le corpus, décrit dans la Section 6.1.3 page 84.

Troisième partie

Partie 3 : Détection automatique des problèmes d'interaction

Résumé

Notre sujet de recherche intervient dans la continuité de travaux sur l'analyse des interactions humain-humain menés au laboratoire de recherche de l'entreprise EDF et des interactions humain-agent du laboratoire LTCI. Notre méthodologie pour le développement du système se décompose ainsi :i) le choix des approches en fonction de l'état-de-l'art, des besoins métier et les ressources disponibles, ii) l'étude du corpus afin de valider les approches choisies, iii) le développement des modules du système et leur test permettant le choix d'une meilleure configuration. Ainsi, nous préférons l'utilisation de l'outil GATE en tant que plateforme de développement car il est modulable. Il permet le développement des règles linguistiques en y intégrant des ressources extérieures et l'utilisation des approches à base d'apprentissage automatique.

En ce qui concerne la méthodologie de développement des règles linguistiques, celles-ci sont conçues de manière itérative en incluant des allers-retours entre la recherche d'indices des problèmes d'interaction dans le corpus de développement, l'étude des indices mentionnés dans l'état de l'art et les tests de performance des règles basées sur ces indices dans ce même corpus. L'étude de la pertinence des indices pour la détection des problèmes d'interaction a montré que la ponctuation multiple et l'argot Internet sont les indices les plus performants pour la détection des problèmes d'interaction explicites, alors que la répétition et la reformulation sont plus appropriés pour les problèmes d'interaction implicites. Cette étude a permis également d'écarter des indices tels que les retours des utilisateurs via un questionnaire de satisfaction, les métadonnées de la performance du système et le temps de réponse des utilisateurs car non-pertinents.

Pour adresser les problématiques soulevées par l'étude de l'état de l'art et du corpus, nous testons plusieurs solutions techniques et leur compatibilité. Ainsi, l'approche hybride comprenant une approche à base de règles et une approche à base de représentation sémantique des mots, apprise de façon non-supervisée, permet de détecter davantage de répétitions et reformulations de l'utilisateur. Les règles linguistiques opèrent à plusieurs niveaux : au niveau du contexte de l'énoncé de l'utilisateur, du contexte d'une paire adjacente et d'un dialogue. Les prétraitements assurent le fonctionnement des règles manipulant des lemmes. Le correcteur orthographique apporte beaucoup de bruit. La désambiguïsation au niveau des énoncés est effectuée via la détection des opinions positives et des énoncés utilisateur hors-sujet. La détection d'émotions repose sur le lexique du dictionnaire LIWC mais des études plus approfondies permettront d'éclaircir un apport possible du lexique de Blogoscopie plus proche de notre corpus.

Nous testons également plusieurs approches pour le calcul de la similarité sémantique et obtenons de meilleurs résultats en utilisant la mesure cosinus entre deux vecteurs d'énoncés utilisateur. Les vecteurs des énoncés sont obtenus avec la somme des vecteurs de mots dans l'énoncé. Les vecteurs de mots sont appris de manière non-supervisée en appliquant le modèle word2vec sur le corpus non-annoté. L'étude des résultats montre que le meilleur rappel est obtenu par le système combinant les règles linguistiques, le correcteur d'orthographe et le calcul de la similarité sémantique. De plus, nous décrivons les caractéristiques des énoncés détectés comme similaires et les éléments qui restent difficiles à traiter pour le système.

Chapitre 6

Système hybride de Détection Automatique des Problèmes d'Interactions (DAPI)

Méthodologie du développement d'un système de détection automatique des problèmes d'interaction à base d'automates

Sommaire

Depuis quelques années, les agents virtuels et les chatbots sont devenus un outil populaire auprès des entreprises pour la gestion de la relation client afin d'alléger la tâche des conseillers humains.

Dès l'apparition des premiers systèmes vocaux automatiques, la détection et la prédiction des problèmes d'interaction sont devenus un enjeu important pour les entreprises. Traditionnellement, les logs du système traçant, principalement, les scores des erreurs de transcription de la parole, de la compréhension du langage parlé et l'historique de dialogue [WALKER et collab., 2002] sont utilisés comme des indices pour la détection et la prédiction des problèmes d'interaction. Ces indices sont donnés en entrée des systèmes d'apprentissage automatique supervisés. Il existe relativement peu de travaux pour les systèmes automatiques de tchat et se concentrent surtout sur un type spécifique des PI, la détection de "breakdown".

Ce chapitre décrit la méthodologie des choix effectués avant la mise en place du système dans la Section 6.1. Nous présentons les deux approches dont le système de détection des problèmes d'interaction rencontrés dans un dialogue écrit humain-agent virtuel est composé dans les sections 6.2 et 6.3.

6.1 Choix méthodologiques

Nous consacrons cette section à la description des éléments qui ont guidé nos choix lors du développement du système. Ces éléments proviennent de l'état de l'art de la communauté scientifique, des études effectuées au sein de la R&D de l'entreprise concernant la relation client et au sein du laboratoire LTCI (le Laboratoire de Traitement et Communication de l'Information) concernant l'interaction humain - agent virtuel. Dans cette section nous donnons les éléments expliquant le choix d'une approche hybride (Section 6.1.1 page 82), de l'outil de travail (Section 6.1.2 page 83) et des indices pour la détection de problèmes d'interaction (Section 6.1.3 page 84).

6.1.1 Choix d'une approche hybride mêlant règles et apprentissage par représentation.

Notre analyse de l'état de l'art nous permet de choisir le modèle théorique de l'"appraisal" de MARTIN et WHITE [2005] qui décrit une opinion comme une évaluation d'une cible par une source, parmi les théories existantes présentées dans le Chapitre 2. Le sujet de l'analyse d'opinions des clients d'EDF a été déjà abordé par CAILLIAU et CAVET [2010]; CLAVEL et collab. [2013] concernant les centres d'appel. Leurs travaux représentent un premier pas vers la détection des conversations problématiques entre un client et un conseiller humain en français. Le corpus des dialogues avec la conseillère virtuelle a attiré l'intérêt de SUIGNARD [2010] qui a effectué une analyse préliminaire du corpus de tchat avec l'agent conversationnel Laura. Notre choix d'utiliser une approche symbolique s'appuie, entre autres sur les travaux de LANGLET et CLAVEL [2015] qui utilisent une approche à base de règles linguistiques pour la détection des expressions de préférence dans le contenu verbal d'un utilisateur parlant avec un agent conversationnel animé (ACA) en face à face.

Suite à l'étude des approches existantes de détection des opinions et des émotions présentées dans le Chapitre 3 et les éléments exposés ci-dessus, nous proposons de favoriser l'approche hybride composée des règles linguistiques et de la représentation sémantique des mots apprise par une méthode non-supervisée. Contrairement aux systèmes

d'apprentissage automatique supervisé, notre approche permet d'avoir un système fonctionnel en l'absence d'un grand corpus annoté manuellement.

Pour valider l'approche à base de règles, il est nécessaire d'acquérir des données quantitatives et qualitatives sur la nature du corpus. Nous avons mené l'analyse de ce dernier selon l'état-de-l'art des corpus de « conversation écrite» décrits dans le Chapitre 4. Nous avons récolté des données quantitatives concernant des phénomènes du "français tchaté" présents dans le corpus, tels que l'utilisation de l'argot Internet et de la ponctuation multiple (Chapitre 4 Section 4.2 et Section 4.3), et l'analyse manuelle des extraits de corpus a fourni des exemples du langage formel et informel des utilisateurs pour l'analyse qualitative du corpus.

L'étude du corpus nous a permis d'évaluer les difficultés, telles que le traitement automatique des fautes de frappe ou des phrases mal construites, aussi bien que les opportunités qu'il représente, comme les indices de problèmes d'interaction. Les spécificités linguistiques du corpus présentées dans le Chapitre 6.1.3 nous ont conduit à confirmer l'approche symbolique comme un des éléments du système hybride et permettant de mieux étudier ce type de données. Dans cette optique nous avons choisi l'outil de travail GATE que nous présentons dans la Section 6.1.2. L'analyse des difficultés pour le traitement automatique nous a convaincu également de la nécessité de compléter l'approche symbolique avec la représentation sémantique des mots apprise par une méthode non-supervisée.

6.1.2 Choix de l'outil GATE

En fonction de l'approche à base de règles linguistiques que nous avons choisie, nous avons cherché un outil open source qui nous permettrait à la fois de combiner les outils de prétraitement textuel déjà disponibles pour le français, et d'utiliser des ressources linguistiques, de développer des règles linguistiques de détection des problèmes d'interaction dans les dialogues et d'avoir la possibilité d'utiliser également une approche statistique dans le futur.

La plateforme GATE CUNNINGHAM et collab. [2011] développée par l'équipe de l'université de Sheffield conjugue toutes ces qualités et bénéficie d'une grande communauté d'utilisateurs. Les modules proposés par ce logiciel ouvrent de nombreuses possibilités de traitement de texte pour l'extraction d'information. Ils permettent, entre autres, la tokenization, la lemmatisation, l'analyse syntaxique, l'étiquetage morphosyntaxique, l'annotation, aussi bien automatique (à base de règles, dictionnaires ou d'apprentissage automatique), que manuelle, l'évaluation des résultats d'annotation. La plateforme est développée en Java et peut s'intégrer à une chaîne de traitements de fichiers textuels. Elle permet également de tester des règles linguistiques simples avant de les implémenter, ce qui permet d'explorer davantage le corpus.

Concernant le domaine de la détection des opinions et des sentiments, GATE est l'un des outils utilisés par la communauté de chercheurs non seulement pour le prétraitement de texte et l'annotation à base de dictionnaires [MULHOLLAND et collab., 2016] avant d'effectuer la détection par des méthodes d'apprentissage automatique supervisé à base d'autres outils [TANG et collab., 2017], mais également pour la détection des sentiments à base de règles [KIEU et PHAM, 2010] ou de la classification automatique supervisée [FUNK et collab., 2008].

Nous avons donc arrêté notre choix sur la plateforme GATE pour effectuer les prétraitements du texte, tel que la tokenization, l'intégration d'un logiciel externe d'annotation en parties de discours (TreeTagger[SCHMID, 1995]) et le développement des règles pour

l'annotation des problèmes d'interaction avec l'utilisation de dictionnaires. Il existe deux plugs-in GATE pour l'analyse morpho-syntaxique pour le français : TreeTagger et Stanford [TOUTANOVA et collab., 2003]. Malheureusement, le taggeur morpho-syntaxique de Stanford ne propose pas de lemmatisation. C'est la raison pour laquelle nous avons arrêté notre choix sur l'outil TreeTagger.

6.1.3 Analyse du corpus de développement : choix des indices des problèmes d'interaction

Notre méthodologie de développement d'un système à base de règles repose sur un processus itératif entre l'analyse du corpus de développement *(LauraDev)*, qui fournit des indices pour la conception des règles, les tests des règles ainsi conçues qui fournissent de nouvelles informations sur la nature du corpus et l'étude de l'état-de-l'art concernant des patrons existants. Enfin, le corpus de référence annoté manuellement permet d'avoir une idée claire de l'efficacité du système. Lorsque cette séquence est terminée, le corpus de référence devient un corpus de développement, que nous appelons ici *DevCorpus*. Ensuite la séquence des action est répétée.

Puisque nous avons déjà décrit le corpus *LauraDev* dans le chapitre 4 page 51, nous nous concentrons ici sur le corpus *DevCorpus*. Ce corpus de développement contient plusieurs niveaux d'information permettant d'évaluer l'intérêt des spécificités linguistiques du corpus pour la détection des problèmes d'interaction. Le premier niveau contient les données des retours clients vis-à-vis de leur interaction avec un agent conversationnel (la métadonnée "feedback"). Le deuxième niveau contient la métadonnée "failed" fournie par l'agent virtuel lui-même et indiquant une conversation en échec. Le troisième niveau contient les annotations manuelles des problèmes d'interaction qui ont été obtenues lors de la première session de travail de l'annotateur. Nous donnons la distribution des annotations dans ce corpus dans la Figure 6.1. Selon WALKER et collab. [1997], pour évaluer la performance d'un agent conversationnel, il est important de prendre en compte la corrélation entre les informations contenues dans les logs du système de l'agent et les retours des utilisateurs sur leur satisfaction de l'interaction avec l'agent. Comme nous l'avons déjà mentionné, les logs des systèmes à entrée vocale contiennent des informations telles que les scores des erreurs de transcription de la parole, de la compréhension du langage parlé et l'historique de dialogue. Dans notre cas, les deux éléments des logs qui peuvent jouer un rôle similaire, sont la métadonnée "failed" et les retours des utilisateurs (la métadonnée "feedback"). ARTSTEIN et collab. [2009] confirment l'existence de cette corrélation pour un agent conçu pour la recherche d'information sur un sujet prédéfini et destiné à un large public d'utilisateurs. Or, la corrélation n'existe pas toujours. SILVERVARG et JÖNSSON [2011] rapportent que dans leur cas spécifique d'évaluation d'un chatbot conçu pour accompagner l'apprentissage des adolescents à l'école, cette corrélation est absente. L'absence de corrélation est expliquée par la spécificité de l'âge des sujets de l'expérimentation.

Notre analyse cherche à déterminer la présence d'une corrélation entre les retours utilisateurs, la métadonnée "failed" et les annotations humaines. Cette analyse nous permet d'évaluer la pertinence des retours utilisateurs et de la métadonnée "failed" pour l'évaluation des indices des problèmes d'interaction dans l'absence des annotations humaines.

Choix de l'information considérée par le système

Nous présentons ici les statistiques du corpus *DevCorpus* annoté manuellement. Il est annoté par un annotateur, entre autres, en problèmes d'interaction de la même façon que le corpus de référence. Les informations statistiques du corpus sont données dans le Tableau 6.1. La distribution des énoncés contenant un problème d'interaction dans les

Nombre de	**DevCorpus**
Dialogues	3 000
Paires Adjacentes (PA) = énoncés	8 576
Dialogues contenant au moins un PI	741 (25%)
Énoncés contenant un PI	1 284 (15%)
Énoncés contenant un PI dans un dialogue problématique (en moyenne)	2

TABLEAU 6.1 – Les informations statistiques sur le corpus de développement DevCorpus

dialogues est présentée dans la Figure 6.1. 25% des dialogues du corpus contiennent des

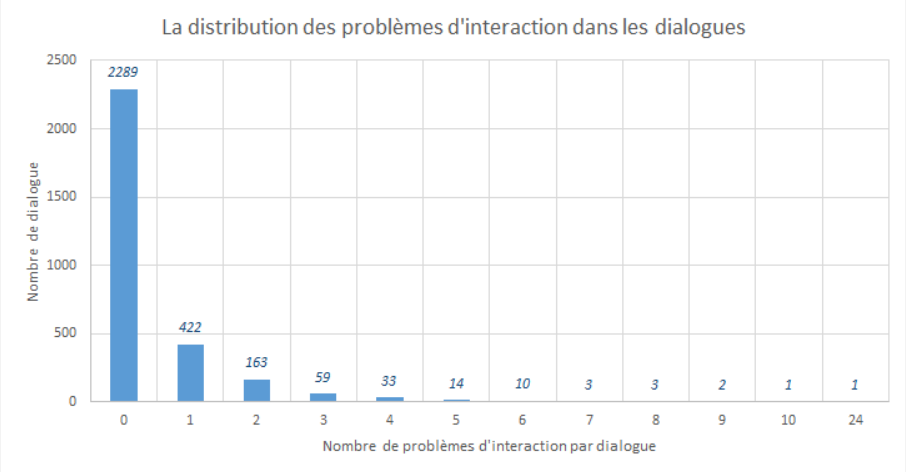


FIGURE 6.1 – La distribution des problèmes d'interaction dans les dialogues

problèmes d'interaction. 13,9% des dialogues ne contiennent qu'un seul PI. Le ratio des dialogues contenant au moins un PI est relativement bas et la majorité des dialogues n'en contiennent qu'un.

Le *DevCorpus* comporte également des métadonnées indiquant l'auto-évaluation des performances du système de l'agent : "OnlyDirectMatches", signifiant que l'utilisateur a cliqué sur un des liens proposés par l'agent, "EndingWithDirectMatch", signifiant qu'à la fin du dialogue l'utilisateur a cliqué sur un des liens proposés par l'agent, "Failed", signifiant que l'utilisateur a quitté le dialogue sans cliquer sur un lien proposé par l'agent. La figure 6.2 montre que le système de l'agent conversationnel détecte 13% des dialogues échoués. Cela correspond à plus de la moitié (57%) des dialogues contenant un PI. La détection des dialogues échoués ne se fait qu'à la fin du dialogue. Les dialogues, ayant la métadonnée "Failed", contiennent en moyenne 2,8 paires adjacentes. Cela signifie deux choses :

1. que ces dialogues contiennent au moins un énoncé contenant un problème d'interaction ;

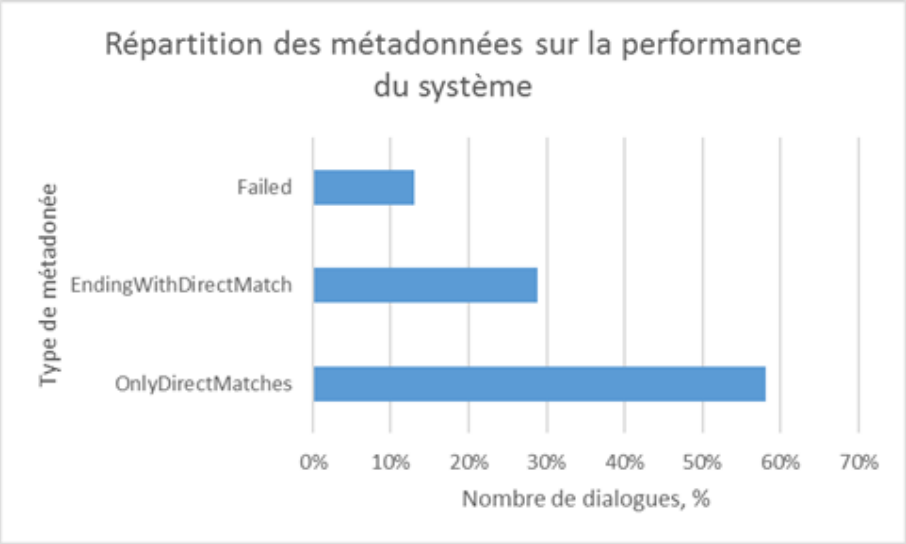


FIGURE 6.2 – La distribution des métadonnées dans le corpus *DevCorpus*

2. et que l'énoncé contenant un problème d'interaction n'est pas forcément le dernier énoncé du dialogue.

De plus, l'apparition d'un PI ne signifie pas forcément que le dialogue échoue. La métadonnée "Failed" ne permet donc pas l'identification immédiate d'un problème d'interaction et ne fournira pas d'information opportunément pour l'adaptation de la stratégie de l'interaction de l'agent virtuel. Nous cherchons à identifier un PI dès son apparition c'est-à-dire au plus proche du temps réel. L'agent pourrait ensuite proposer à l'utilisateur une réponse pertinente.

Un questionnaire de satisfaction est proposé à l'utilisateur par le système de l'agent afin de recueillir son avis. En fonction de la réponse de l'utilisateur, le système enregistre l'une de ces quatre informations : "withoutAnswer" lorsque l'utilisateur n'a rien répondu, "positive", "négative" ou vide, lorsque le questionnaire n'a pas été proposé. La Figure 6.3 illustre la distribution dans le corpus des différents types de réponse.

Les utilisateurs répondent rarement au questionnaire (9% des dialogues). Le nombre de réponses positives (5%) est légèrement supérieur aux réponses négatives (4%).

Pour identifier le degré de la pertinence des métadonnées du système et des retours client pour l'identification des problèmes d'interaction, nous avons étudié la corrélation entre ces éléments et les annotations manuelles dans les dialogues contenant les retours des utilisateurs. Le Tableau 6.2 page 86 illustre les résultats obtenus.

	Problème d'interaction	**Feedback Positif**	**Feedback Négatif**	**Métadonnée "Failed"**
Problème d'interaction	1,00	-0,05	0,02	0,24
Feedback Positif	-0,05	1,00	-0,05	-0,09
Feedback Négatif	0,02	-0,05	1,00	-0,04
Métadonnée "Failed"	0,24	-0,09	-0,04	1,00

TABLEAU 6.2 – Corrélation de Pearson entre les annotations manuelles, les retours des utilisateurs et la métadonnée du système de l'agent

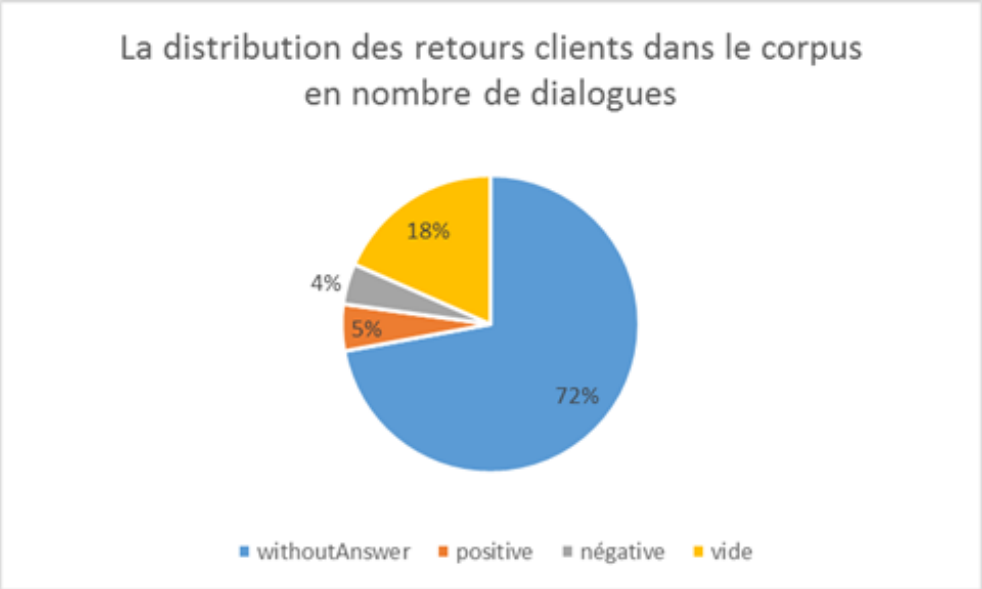


FIGURE 6.3 – La distribution des retours client dans le corpus de développement *DevCorpus*

Le Tableau 6.2 montre que la corrélation entre les problèmes d'interaction et les retours des utilisateurs est quasi nulle. La corrélation entre la métadonnée "Failed" et les problèmes d'interaction est également très faible. La corrélation entre les retours des clients et la métadonnée "Failed" est pratiquement absente également. L'absence de corrélation entre les retours des utilisateurs peut être expliquée par deux points. Le premier point est que la métadonnée "Failed" correspond à l'incapacité de l'agent à répondre à la demande du client, indépendamment du bon déroulement de la discussion. Le second point est que les feedbacks sont trop peu nombreux pour être représentatifs.

Nous considérons les annotations humaines comme une référence car elles ont été réalisées par un annotateur expert (pour plus de détails sur la stratégie d'annotation, voir le Chapitre 5). Nous concluons que les retours client et les métadonnées du système ne sont pas applicables pour l'évaluation de la pertinence des spécificités linguistiques du corpus pour la détection des problèmes d'interaction.

Pertinence des indices et des règles pour la détection des problèmes d'interaction

Suite à l'analyse manuelle des exemples des dialogues issus du corpus *LauraDev* ne contenant pas d'annotations manuelles décrite dans la Section 4.3 du Chapitre 4, nous identifions, comme décrit également dans le chapitre 5 dédié à l'annotation du corpus, deux types d'indices des PI : d'une part, des marqueurs d'émotions négatives et d'autre part, des répétitions et des reformulations. Pour le premier type d'indices, nous proposons d'utiliser les spécificités linguistiques du corpus liées au langage Internet. Nous avons décrit l'étude de la présence des émotions dans le corpus sans prendre en compte leur cible dans la Section 4.3.2 page 60. Nous étudions ici leur pertinence pour la détection des émotions dont la cible est l'interaction. Pour le second type d'indices, notre objectif est de détecter deux énoncés similaires dans un dialogue. Il existe beaucoup de mesures linguistiques de similarité. MCGILL [1979] a notamment effectué une comparaison de 67 mesures de similarités dans le cadre d'un travail sur un système de recherche d'information. Nous optons pour les distances linguistiques classiques de la détection des répétitions. En étudiant nos données, nous avons identifié principalement les cas suivants :

— Les énoncés comportant des fautes de frappes ou d'orthographes qui peuvent engendrer des incompréhensions de la part de l'agent.

Exemple 2 *Faute de frappe dans un énoncé utilisateur.*
User [U1] : *facture papier*
Agent : *(...)*
User [U2] : *facture paier*

Nous avons émis l'hypothèse que la distance de Levenstein [LEVENSHTEIN, 1966] est bien adaptée et suffisante pour traiter ce cas. De plus, elle est disponible nativement en Java et ne demande donc pas de développement spécifique.

— Les énoncés étant des reformulations ayant plusieurs mots communs avec un énoncé précédent.

Exemple 3 *Reformulation de l'utilisateur*
User [U1] : *releve confiance*
Agent : *(...)*
User [U2] : *comment opter pour le releve confiance*

Pour ce cas, nous avons émis l'hypothèse que l'application de la distance de Jaccard [JACCARD, 1908] améliorée de BRUNET [2003] permettra de détecter une grande partie des reformulations utilisateur. La distance de Jaccard améliorée de BRUNET [2003] détecte non seulement l'intersection des mots mais est également moins sensible aux grands écarts de longueur entre deux énoncés utilisateur.

— Les reformulations contenant des répétitions des termes métier.

Exemple 4 *Répétition des termes métier dans les reformulations de l'utilisateur*
User [U1] : *comment faire pour releverle comteur de* `<business_term>gaz</business_term>` *pas de cle pour ouvrir*
Agent : *(...)*
User [U2] : *commen fair pour relever le conteur de* `<business_term>gaz</business_term>??`

Comme nous l'avons vu dans l'état-de-l'art, MEENA et collab. [2015] proposent d'utiliser les concepts détectés par le chatbot dans l'énoncé utilisateur comme des indices de la répétition. Notre corpus ne contient pas de métadonnées permettant de connaître les informations extraites par le système de chatbot. Par conséquent, notre hypothèse est que la répétition du vocabulaire métier dans les énoncés utilisateur indique la répétition ou la reformulation.
Nous cherchons à identifier la pertinence de ces indices en nous appuyant sur les annotations manuelles des problèmes d'interaction.

Pertinence des spécificités du langage Internet

Les spécificités du langage Internet dont nous cherchons à identifier la pertinence pour la détection des PI sont les suivantes : les caractères écho, les majuscules, la ponctuation multiple, les smileys, les interjections et l'argot Internet.

Nous ne trouvons que deux exemples où deux indices sont utilisés dans le même énoncé : l'utilisation très fréquente des majuscules et des signes d'exclamation. Nous trouvons également un énoncé contenant deux termes d'argot Internet. Dans tous les

autres cas d'énoncés contenant un PI et l'un des indices, l'utilisation de ces spécificités est singulière. Au total, 3,74% des énoncés problématiques contiennent des indices. La ponctuation multiple et l'argot Internet sont les indices de ce groupe les mieux représentés et représentent 3% des énoncés contenant un PI. Les smileys et les interjections ne sont pas présents.

D'une part, l'ensemble de ces indices semble être peu pertinent pour la détection des PI. D'autre part, ils sont contenus dans 9% des énoncés annotés avec une relation. Le Tableau 6.3 présente le nombre des annotations de chaque type dans le corpus.

Type de problème d'interaction	Nombre d'énoncés
Total	1284
implicites	1185
explicites	99 (8%)

TABLEAU 6.3 – Nombre d'annotations manuelles dans le DevCorpus

Pertinence des indices pour la détection des répétions et reformulations

Nous avons appliqué des règles au DevCorpus utilisant les distances de Levenshtein et Jaccard en prenant en compte l'historique du dialogue. Afin de détecter des répétitions et des reformulations de l'utilisateur, nous avons également appliqué une règle basée sur la répétition des termes métier dans les énoncés utilisateur au sein du dialogue.

Type de règle	Nombre d'énoncés correctement détectés (pourcentage par rapport au nombre des PI total)	Bruit
Distances	419 (32,63%)	241
Termes métier	265 (20,64%)	220

TABLEAU 6.4 – Nombre d'énoncés contenant un PI détectés pour chaque type de règles

Le Tableau 6.4 montre le résultat de la détection des PI par les règles. Les deux règles combinées permettent de détecter plus de la moitié des PI. Par conséquent, nous les considérons comme pertinentes et que notre hypothèse les concernant est vérifiée. En revanche, ces règles ne détectent pas les reformulations de l'utilisateur sans autres mots identiques communs et ne contenant que des changements de la forme de mots, comme « payer » et « payement ». Nous avons vu dans l'état-de-l'art qu'il est possible de détecter les reformulations également en utilisant les mesures de la similarité sémantique. Nous décrivons nos choix et nos analyses dans le chapitre 6.3.

La Figure 6.4 page 90 illustre la couverture de la détection des énoncés contenant des PI par les indices que nous avons présentés ci-dessus. Elle permet de bien visualiser que même en supposant que tous les spécificités du langage Internet se situent dans les PI explicites (ce qui n'est pas toujours vrai), Ces indices permettraient de détecter qu'une moitié de ce type des PI. Les distances de Levenshtein et de Jaccard permettent de détecter un tiers des PI. Une part importante (43%) reste non-couverte pas les indices.

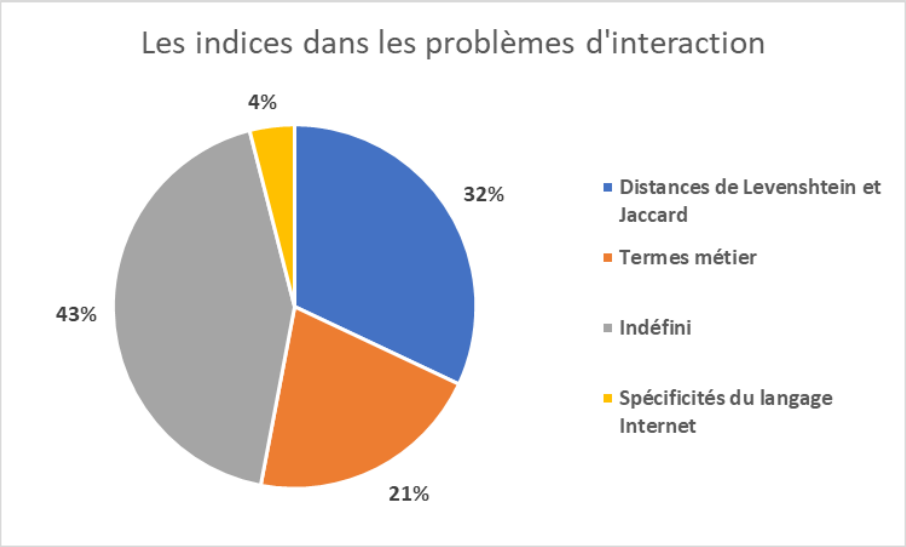


FIGURE 6.4 – La part de chaque type d'indices dans les annotations manuelles des PI

6.1.4 Architecture générale du système DAPI

Le système DAPI (pour Détection Automatique de Problèmes d'Interaction) est développé afin de détecter des problèmes dans des interactions écrites de type "tchat" entre une conseillère virtuelle et son utilisateur. L'architecture globale du système est présentée dans la Figure 6.5.

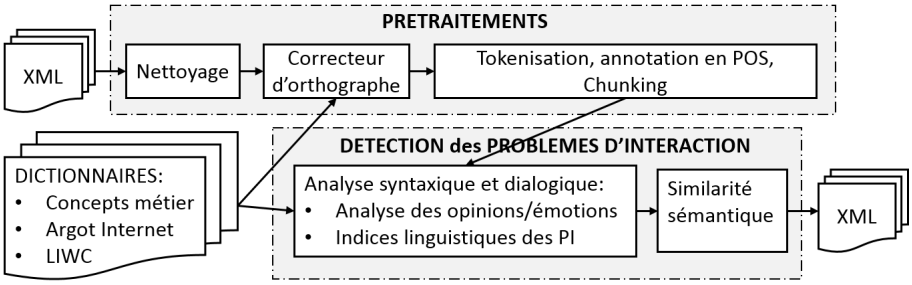


FIGURE 6.5 – Le schéma du système DAPI

L'hybridation du système consiste en l'utilisation de règles linguistiques basées sur des transducteurs d'automates à états finis et en l'apprentissage non supervisé des représentations sémantiques de mots. Le système fonctionne en trois étapes : tout d'abord les prétraitements (Section 6.2.1) chargés de l'anonymisation des données et de leur mise en forme pour les traitements dans le logiciel GATE, suivie de la détection des problèmes d'interaction consistant en l'analyse syntaxique et dialogique à base de règles et le calcul de la similarité sémantique, basée sur les vecteurs des énoncés.

Les règles linguistiques pour la détection des problèmes d'interaction (PI) sont conçues manuellement à l'aide du logiciel GATE [CUNNINGHAM et collab., 2011] et son module JAPE [1] [CUNNINGHAM et collab., 2000]. JAPE propose des transducteurs à états finis sur des annotations et permet d'utiliser des expressions régulières. Nos règles linguistiques reposent sur l'historique d'un dialogue et les spécificités du français tchaté. Elles cherchent

1. Java Annotation Patterns Engine

à détecter l'opinion ou l'émotion de l'utilisateur pendant l'interaction et ont pour modèle la théorie de l'évaluation de MARTIN et WHITE [2005], que nous avons défini dans la Section 2 page 11. Comme dans les chapitres précédents, nous utilisons ici l'acronyme OPEM lorsque nous parlons de l'opinion et des phénomènes reliés aux opinions.

Les règles linguistiques s'appuient également sur les éléments suivants :

- le lexique LIWC pour le français (PIOLAT et collab. [2011a])
- la liste des termes d'argot Internet [2] (ex. *lol*)
- le dictionnaire des termes métier créé en entreprise contenant 400 entrées et leurs variantes orthographiques
- des listes de lexiques conçues manuellement :
  - les émoticons de base
  - les sources potentielles de l'opinion (les variantes de pronoms personnels à la première personne, car on cherche à détecter l'opinion de l'utilisateur)
  - les verbes et des expressions d'opinion (20 entrées)
  - les expressions des concepts spécifiques (ex. remerciements, salutations, demandes (30 entrées)).

Afin de détecter l'interaction en tant que cible de l'OPEM, nous avons regroupé des expressions la désignant sous un concept. Le concept d'interaction regroupe les concepts des adverbes interrogatifs (ex. comment), des verbes et des adjectifs pouvant être utilisés pour exprimer la critique (ex. efficace, se demander), des verbes indiquant des échanges entre deux interlocuteurs (ex. comprendre, répondre), soit au total, près de 165 entrées. Le lexique positif et les insultes de LIWC sont typés selon leurs probabilité de n'être utilisés que pour des discussions hors-sujet (ex. amour) ou lors d'un échange sur un thème métier (ex. l'utilisation d'une insulte pour exprimer son irritation).

L'apprentissage de la sémantique des mots du corpus par l'apprentissage des représentations vectorielles de mots, type plongements lexicaux, permet de détecter des répétitions et des reformulations de l'utilisateur. Ces derniers représentent des indices de problèmes d'interaction.
Dans les sections suivantes nous décrivons d'abord les règles qui représentent l'approche symbolique à la détection des PI (Section 6.2) et ensuite les plongements lexicaux qui représentent l'approche non-supervisée (Section 6.3).

6.2 Approche symbolique à la détection des problèmes d'interaction

6.2.1 Prétraitements

Les prétraitements des dialogues ont pour but d'effectuer leur anonymisation et leur mise en forme pour une manipulation simplifiée dans l'outil GATE. Les prétraitements contiennent aussi une désambiguïsation des énoncés.
La désambiguïsation des énoncés hors-sujet et des énoncés contenant un OPEM de polarité positive simplifie les règles de détection des problèmes d'interaction. La détection des énoncés hors-sujet permet de distinguer les énoncés où l'utilisateur utilise l'agent

2. http ://fr.wiktionary.org/wiki/Annexe :Liste_de_termes_d%E2%80%99argot_Internet

pour du "small talk". Les énoncés positifs ne représentent pas un problème d'interaction, au contraire, ils témoignent de son bon déroulement. Nous ne voulons donc pas les confondre avec les énoncés contenant un problème lors de l'annotation.

D'un point de vue technique, les prétraitements sont hétérogènes. Le nettoyage et les corrections orthographiques sont effectués par un script Python. L'enrichissement du texte et la désambiguïsation des énoncés sont faits à l'intérieur de la plateforme GATE.

Les fichiers XML produits par le chatbot sont soumis aux prétraitements suivants : le nettoyage et la mise-en-forme des fichiers xml, la correction d'orthographe, l'enrichissement du texte avec des informations morphosyntaxiques.

Nettoyage et mise en forme des fichiers xml

Cette étape est nécessaire pour pouvoir faire une analyse morphosyntaxique du texte. Les hyperliens, les messages système, mais aussi les noms et les adresses sont source de bruits et d'erreurs pour les règles linguistiques. Les prétraitements les transforment donc en balises XML pour qu'elles puissent être directement interprétées par les règles dans GATE. Les fichiers bruts comportent également beaucoup d'informations spécifiques à l'agent virtuel et des données personnelles de l'utilisateur connecté. Ces données contiennent notamment le nom de la personne, le type du navigateur utilisé et son système d'exploitation. Nous ne conservons que les informations du système de l'agent indiquant : les identifiants et les types des dialogues et des paires adjacentes. Les types des paires adjacentes permettent de déterminer si la réponse de l'utilisateur a été écrite manuellement ou correspond à une saisie semi-automatique, telle que le choix d'un lien proposé.

Traitement des erreurs d'orthographe

Un correcteur d'orthographe est utilisé en support des règles utilisant des concepts. Nous décrirons ce type de règles dans la section 6.2.3.

Nos règles font appel aux lemmes des formes dictionnairiques. Cette solution est étroitement liée aux résultats d'annotation par TreeTagger. Il arrive souvent que les tagueurs syntaxiques ne soient pas capables d'identifier le lemme d'un mot lorsque le mot est mal orthographié. Pour résoudre ce problème, nous introduisons un correcteur d'orthographe.

Lors du choix d'un correcteur d'orthographe, nos critères sont les suivants : la compatibilité avec les éléments fondateurs de notre chaîne (GATE ou Python), la disponibilité de la version française, la facilité d'intégration dans la chaîne et le type d'erreurs corrigées.

Lors du choix d'un correcteur d'orthographe, nous considérons les solutions proposées pour le logiciel GATE et pour le langage Python dans notre chaîne de traitements. Le logiciel GATE propose un plug-in pour le traitement des tweets en français. Ce plug-in contient un normalisateur d'orthographe French Normalizer [MAYNARD et collab., December 2016]. Python propose un correcteur d'orthographe via la librairie PyEnchant [3]. Nous avons testé les deux solutions sur 130 énoncés (1058 mots) des utilisateurs choisis au hasard dans le corpus de développement. Les énoncés contenaient 61 mots erronés. Les résultats ont été analysés manuellement et sont présentés dans le Tableau 6.5. L'analyse manuelle des résultats montre que les apostrophes et les termes métiers, particulièrement les abréviations, représentent les difficultés principales pour le correcteur PyEnchant. L'application de l'algorithme permettant de filtrer les termes métier avant d'appliquer la correction, améliore considérablement le résultat de PyEnchant et le rapproche

3. http ://pythonhosted.org/pyenchant/

du résultat du normalisateur de GATE. L'ajout du dictionnaire des termes métier au normalisateur GATE permet de baisser le nombre de fausses corrections de 24 à 21 mais ne permet pas d'augmenter le nombre de corrections correctes.

Correcteur d'orthographe	**Nombre de mots corrigés**	**Nombre de mots corrigés correctement avec raison**	**Mots erronés non-identifiés**	**Nombre de mots corrigés par erreur**
French Normalizer (GATE)	38	14	47	24
PyEnchant (Version sans dictionnaire métier)	94	8	53	86
PyEnchant (Version avec dictionnaires métier)	49	15	46	34

TABLEAU 6.5 – Résultats du test de la correction d'orthographe.

Les deux correcteurs d'orthographe ne font pas les mêmes types de corrections, sauf le seul exemple d'argot présent dans le texte ("svp") normalisé par les deux correcteurs. A part ce cas, le normalisateur de GATE ne corrige que l'absence d'accents. Alors que PyEnchant a réalisé un nombre égal de corrections d'accents manquants (7 cas) et de fautes de frappe/d'orthographe ("méson" - "maison") (7 cas).

Le type d'erreur corrigée n'a pas le même impact sur les performances du taggeur. Sur 100 lemmes non-reconnus par TreeTagger lorsque l'on n'applique aucun correcteur, seuls 8 lemmes supplémentaires sont reconnus après l'application du French Normalizer : 4 sur les mots corrigés avec raison et 4 sur les mots corrigés par erreur. Lorsque le texte est prétraité par PyEnchant : 26 lemmes supplémentaires sont identifiés : 15 sur les mots corrigés avec raison et 11 sur les mots corrigés par erreur. Nous envisageons de corriger certains cas de faux positifs de reconnaissance des lemmes par TreeTagger qui sont liés à l'utilisation du dictionnaire des termes métier lors de l'application de PyEnchant.

Étant donné ces résultats obtenus sur un petit échantillon de données, nous choisissons de continuer nos tests avec le correcteur d'orthographe PyEnchant combiné avec le dictionnaire des termes métier, car celui-ci obtient les meilleurs résultats. Dans la Section 7.2 nous décrivons nos tests sur l'apport du correcteur d'orthographe aux performances globales du système.

Pour pouvoir conserver néanmoins les indices contenus dans le langage Internet, nous vérifions que les mots ne font pas partie des dictionnaires de l'argot Internet, des émotions et des termes métier avant d'appliquer le correcteur d'orthographe. Nous faisons appel à PyEnchant pour déterminer d'abord si un mot fait partie des mots connus. Si le mot n'est pas dans la liste du dictionnaire PyEnchant, nous vérifions alors qu'il n'est pas contenu dans les lexiques mentionnés ci-dessus. Pour vérifier si ce n'est pas une insulte ou un terme métier mal orthographié, nous utilisons la librairie "difflib" [4] de Python, qui est une extension de l'algorithme de Ratcliff et Obershelp [5] (RATCLIFF et METZENER [1988]). Pour chaque mot du texte n'étant pas dans le dictionnaire, nous recherchons le

4. https ://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher
5. Calcule la similitude de deux chaînes comme le nombre de caractères correspondants divisé par le

terme du dictionnaire le plus proche ayant une similarité, calculée par l'algorithme, et supérieure à 0,65. Nous utilisons ensuite ce nouveau terme en remplacement de l'original. Lorsqu'aucune correspondance n'est trouvée, nous corrigeons le mot avec PyEnchant.

Annotation du texte avec des informations morphosyntaxiques

Comme nous l'avons déjà mentionné dans le début de la section 6.1.4, nous avons besoin des informations sur les parties de discours et les informations syntaxiques, telles que les chunks, pour pouvoir créer des règles linguistiques à base d'automates. L'annotation en question est effectuée par le tagueur morphosyntaxique TreeTagger(SCHMID [1994]). C'est un tagueur probabiliste qui a fait ses preuves dans la communauté TAL (ALLAUZEN et BONNEAU-MAYNARD [2008]). La plateforme GATE permet d'intégrer également le tagueur morphosyntaxique de Stanford [TOUTANOVA et collab., 2003], mais ce dernier ne donne pas d'information sur les lemmes des mots pour le français [6].
7 règles JAPE [7] effectuent l'enrichissement des annotations avec des informations morphosyntaxiques supplémentaires pour des cas spécifiques, telles que la distinction des verbes auxiliaires "être" et "avoir" d'autres verbes. Les verbes auxiliaires sont utilisés ensuite pour détecter des expressions qui peuvent signaler la présence d'une opinion, par exemple, "être sûr" ou "avoir impression".

6.2.2 Étape lexicale (dictionnaires)

Lors de l'étape lexicale, le texte est annoté par des dictionnaires : l'argot Internet, décrit dans le Chapitre 4 Section 4.3.3 page 63, les listes des émotions et des insultes de LIWC (nous décrivons nos motivations quant au choix de ce dictionnaire ci-dessous) et plusieurs lexiques constitués à la main : émoticônes de base, sources potentielles d'opinion (premières variantes de pronoms personnels), verbes d'opinion et expressions (20 entrées), expressions de différents concepts (gratitude, salutations, demande) (30 entrées).

Dans le système DAPI, la détection d'émotions est basée sur le dictionnaire d'émotions. Nous avons présenté les dictionnaires d'émotions disponibles pour le français dans le chapitre 2.2 et plus particulièrement le dictionnaire LIWC et la distribution des termes d'émotion de ce dictionnaire dans notre corpus de développement dans le chapitre 4.3.2.

Avant de choisir la méthode de gestion de la polarité d'OPEM à base de syntagmes de LANGLET et CLAVEL [2015], nous cherchons à déterminer si cette méthode est pertinente pour la détection d'émotions dans notre corpus de développement. La projection du vocabulaire des émotions LIWC sur les syntagmes annotés par l'annotateur en constituants faisant partie de TreeTagger SCHMID [1994] montre que 67% (nous n'appliquons aucune désambiguïsation de vocabulaire d'émotion lors de ce test) des termes d'émotions de LIWC présents dans le corpus, dont 48% des termes d'émotions positives, sont contenus dans des syntagmes. Plus de la moitié des syntagmes contenant un terme d'émotion

nombre total de caractères dans les deux chaînes. Les caractères correspondants sont ceux dans la plus longue suite commune plus, récursivement, les caractères correspondants dans la région sans correspondances de chaque côté de la plus longue suite commune. (Définition : Compute the similarity of two strings as the number of matching characters divided by the total number of characters in the two strings. Matching characters are those in the longest common subsequence plus, recursively, matching characters in the unmatched region on either side of the longest common subsequence. (https://xlinux.nist.gov/dads/HTML/ratcliffObershelp.html))

6. Il existe d'autres tageurs morphosyntaxiques pour le français, tels que LIA_tag http://pageperso.lif.univ-mrs.fr/frederic.bechet/download.html ou MElt [DENIS et SAGOT, 2009]. Nous ne les avons pas testé.

7. un module de GATE (CUNNINGHAM et collab. [2000])

sont des groupes verbaux. Nous confirmons donc l'approche de la gestion de la négation à base de syntagmes.

Ici nous considérons le dictionnaire LIWC résultant d'un travail solide de trois ans de PIOLAT et collab. [2011b], comme le dictionnaire d'émotions le plus complet et correct (par rapport aux dictionnaires constitués automatiquement) à ce jour. Les auteurs PIOLAT et collab. [2011b] caractérisent le dictionnaire LIWC comme un dictionnaire qui doit être adapté dans le cas d'une étude très spécifique, mais qui convient également à l'étude "des écritures expressives (expression du ressenti émotionnel...)". L'avantage de ce dictionnaire est qu'il propose des sous-catégories d'émotions négatives : "l'anxiété, la colère, la tristesse", ainsi qu'une liste de jurons, qui correspondent à trois des six types d'émotions que nous avons déterminés dans notre taxonomie.

En considérant l'aspect "conversation écrite" de notre corpus il parait également intéressant d'effectuer un test du lexique d'évaluation "Blogoscopie" VERNIER et collab. [2010], que nous avons écarté lors de l'analyse préliminaire en nous basant sur la variété des thématiques de blogs dont il était extrait. La comparaison de la projection du lexique de Blogoscopie sur le corpus a montré qu'il y a 1,7 fois plus de syntagmes contenant un terme du lexique de Blogoscopie que de LIWC. Les syntagmes les plus représentatifs sont des groupes verbaux à l'infinitif, les syntagmes coordonnés et des groupes adjectivaux. Ce qui nous amène à revenir sur les conclusions de notre première analyse, ayant écarté ce dictionnaire. L'annotation du corpus avec les lemmes du lexique de Blogoscopie et leur analyse manuelle a montré que pour avoir une idée précise de l'intérêt de ce dictionnaire, il faut l'intégrer dans le système pour effectuer la désambiguïsation. Ce que nous envisageons de faire à l'avenir.

Nous avons considéré également d'autres ressources, tels que des ontologies de type SentiWordNet BACCIANELLA et collab. [2010], mais nous n'avons pas pu identifier de ressource accessible pour le français.

Les sections suivantes présentent la détection des problèmes d'interaction en fonction de l'étendue du contexte dialogique pris en compte.

6.2.3 Détection des problèmes d'interaction en fonction du contexte

Lorsque le texte des dialogues a reçu les prétraitements et est annoté par les dictionnaires, les problèmes d'interaction sont annotés par les règles au sein de GATE puis par des traitements dans un script Python. La détection des PI est faite au niveau du contexte de l'énoncé utilisateur, du contexte d'une paire adjacente et le contexte du dialogue. En effet, le contexte de l'énoncé utilisateur et d'une paire adjacente permet de détecter des expressions d'OPEM d'utilisateur. Le contexte du dialogue donne des éléments pour la détection aussi bien des expressions d'OPEM de l'utilisateur, que des répétitions et des reformulations de l'utilisateur.

Contexte de l'énoncé utilisateur

Au niveau du contexte de l'énoncé utilisateur, nous cherchons à détecter des problèmes d'interaction explicites. Nous détectons une OPEM de polarité négative envers l'interaction, sa source (dans notre cas représenté par l'utilisateur) et sa cible : l'interaction. Ces trois éléments représentent une relation (voir la description de l'annotation manuelle dans le Chapitre 5 page 67). Pour détecter la relation, nous nous appuyons sur des indices contenus dans le langage de l'utilisateur (ponctuation, lexique) et sur des concepts (ex. concept d'interaction).

Chaque élément potentiel d'une relation est annoté et seulement ensuite nous cherchons à rassembler ces éléments dans une relation selon les patrons de relations que nous décrivons dans ce chapitre. Lorsqu'une relation est détectée, nous annotons l'énoncé utilisateur en sa totalité. Nous décrirons ici d'abord la détection de la source, de l'OPEM et de la cible et ensuite de la relation.

Source La source potentielle est détectée à l'aide du lexique contenant des formes du pronom personnel à la première personne, puisque nous ne nous intéressons qu'aux OPEM appartenant à l'utilisateur.

OPEM avec une cible explicite Pour détecter des PI, nous nous concentrons sur la détection d'OPEM à polarité négative. L'OPEM potentiel est modélisé par treize règles de 3 à 4 niveaux de complexité, y compris quatre règles de quatre niveaux de complexité pour la gestion de la négation. La gestion de la négation s'appuie sur l'approche à base de chunks, proposé par LANGLET et CLAVEL [2015]. Lorsqu'un chunk contient un terme d'émotion ou une expression d'opinion, le chunk verbal est annoté comme OPEM$chunk$VN, le chunk nominal OPEM$chunk$NP et adjectival OPEM$chunk$AP.

Les termes d'émotion sont annotés à l'étape lexicale. Les expressions d'opinion sont annotées en s'appuyant sur des dictionnaires des verbes d'opinion (ex. "penser", "préférer") et des mots d'opinion (ex. "avis", "douteux"). Les règles permettent d'annoter des expressions tels que "il est évident", "être sûr", "il (me) semble", "je me doute", "avis sur".

Chaque expression est initialisée avec la polarité zéro, sauf les expressions contenant un verbe d'opinion. Ces dernières héritent de la polarité du verbe.
La gestion de la négation sera expliquée dans la Section 6.2.4.

Cible La cible potentielle de l'OPEM est modélisée par les pronoms personnels de la deuxième personne ("vous", "tu") dans les phrases interrogatives commençant par l'adverbe "comment" ou ses expressions synonymiques, une liste de noms péjoratifs féminins comme "blonde", "bonniche" et un concept d'interaction qui permet de prendre en compte des expressions avec lesquelles l'utilisateur peut adresser l'interaction. Le *concept de l'interaction* est modélisé du point de vue de :

— la compréhension mutuelle lors de l'interaction, par les verbes "comprendre", "expliquer" et leurs synonymes ;

— l'efficacité de l'agent dans l'exercice de son travail (les synonymes des noms "l'efficacité", "utilité" et des verbes "se demander", "remercier") ;

— la pertinence de la réponse de l'agent (des synonymes des noms "question", "réponse" et des verbes "relire", "répondre").

Les noms synonymiques de l'"utilité", tels que "avantage", "service" ou encore "besoin" ne sont utilisés qu'au singulier pour limiter les cas ambigus (comparez "votre service" Vs. "vos services", où "vos services" serait employé plus souvent dans le sens "les services de l'entreprise" en général). Les verbes sont divisés en deux groupes :

- les verbes pouvant être employés aussi bien pour la critique que pour exprimer la demande d'explications suite à l'incompréhension de la part de l'utilisateur, par exemple le verbe "expliquer" : "vous n'expliquez rien" ou "expliquez-moi encore une fois" ;
- les verbes servant à exprimer un retour positif par rapport à l'interaction ("remercier").

Pour détecter des expressions comme "de le savoir" dans une phrase "ça serait bien de le savoir", nous nous appuyons sur l'annotation des groupes prépositionnels de TreeTagger.

La cible potentielle inconnue Lorsque la cible potentielle est exprimée par un pronom démonstratif, nous la considérons comme une cible inconnue $Cible_{type:unknown}$. La désambiguïsation des anaphores est décrite dans la Section 6.2.4. Pour pouvoir s'affranchir du problème de l'orthographe lors de l'appel des concepts, nous avons introduit un correcteur d'orthographe comme étape de prétraitement que nous avons décrit dans la section 6.2.1.

OPEM avec une cible implicite Pour pouvoir détecter des OPEM avec une cible implicite, nous utilisons des indices typographiques, tels que la ponctuation multiple, des émoticônes et des expressions d'insatisfaction (ex. "Laisse moi tranquille"). L'avantage des points d'interrogation multiples et des insultes est que leur polarité n'est pas ambigüe et toujours négative. Malgré la présence très rare de certains indices tels que les émoticônes, nous préférons les utiliser pour n'ignorer aucune opinion ou émotion négative d'un utilisateur concernant l'interaction.

Ensuite, si les éléments potentiels d'une relation, une source, une cible ou OPEM, correspondent à un patron de relation, l'énoncé de l'utilisateur est annoté.

Relations Les relations sont modélisées par six patrons de type [$Source$OPEM$Cible$] qui varient en fonction de la présence d'un des trois éléments du triplet dans une phrase. Le sixième patron modélise l'annotation d'une relation dans des phrases complexes. La liste des patrons est décrite ci-dessous. Chaque élément faisant partie d'un patron de relation, tels que $Source$, OPEM et $Cible$ sont des éléments potentiels.

Régles d'annotation des PI en fonction des patrons de relation :

1. [$Cible$OPEM] Exemple : "votre réponse est insatisfaisante".
2. [$Cible_{type:vous}$ $Cible_{type:verbe}$ $Négation$]
   La combinaison de la cible potentielle exprimée par un pronom personnel à la deuxième personne, une cible potentielle exprimée par un verbe et une négation. Par exemple, "Vous ne répondez à rien.".
3. [$Source$ $Demand$ $Cible$]
   Soit U un énoncé utilisateur.
   Soit $Demand$ un concept qui détecte des demandes impératives des utilisateurs, telles que "Je veux une réponse". IF $Source$ est suivie de $Demand$ & $Demand$ et suivie de $Cible$, WHERE $Source$ & $Demand$ & $Cible$ ∈ U, THEN $Relation$
4. [OPEM_$cible_implicit$]
   Soit OPEM_$cible_implicit$ une opinion ou une émotion exprimé envers une cible implicite.
   IF OPEM_$cible_implicit$ ∈ U, THEN $Relation$ Ce patron correspond, par exemple, aux énoncés très courts de l'utilisateur comme "??".
5. [$Cible_{type:unknown}$ OPEM],
   par exemple "ça ne m'aide pas", où "ça" se réfère à la réponse de l'Agent
6. [$Thanks$ $CoordConj$ $Relation$]
   Soit $Thanks$ un concept de l'appréciation exprimée par l'utilisateur. Soit $CoordConj$ une conjonction de coordination comme "mais".

IF T$hanks$ est suivie de C$oordConj$ & C$oordConj$ est suivie de R$elation$, WHERE T$hanks$ & C$oordConj$ & R$elation$ ∈ U, THEN R$elation$, par exemple "Merci mais ça ne m'aide pas".

Contexte d'une paire adjacente

Lorsqu'il n'est pas possible d'identifier la cible d'une OPEM au sein du contexte de l'énoncé utilisateur, nous nous appuyons sur le contexte de l'énoncé agent. Ce contexte peut être utilisé aussi bien pour la détection d'une opinion ou émotion de l'utilisateur envers l'interaction, que d'un PI implicite.

Relation Nous définissons alors une liste des phrases standards de l'agent qui expriment soit son incapacité à donner une réponse attendue, soit son invitation à revenir envers lui en cas de besoin. Lorsque l'une de ces phrases, par exemple, "Je suis désolée", est suivie d'une insulte de l'utilisateur ou d'une demande d'aide, nous annotons l'énoncé utilisateur comme une **relation**.

Problème d'interaction implicite Au niveau d'une paire adjacente, nous cherchons à identifier les PI en nous appuyant sur deux indices :

1. demande spontanée de mise en relation ;
2. expressions d'insatisfaction par rapport à la réponse de l'agent

Demande spontanée de mise en relation Nous définissons une demande spontanée de mise en relation de la part de l'utilisateur comme un ensemble de lemmes suivants :

1. "contacter", "téléphoner", "téléphone" ;
2. "conseiller"

Au moins un mot de chaque groupe ou un"appel" définit comme tel, doit être présent dans un énoncé utilisateur. Les règles de détection des PI basées sur le concept de **la demande spontanée de mise en relation** sont les suivantes :

1. les demandes de mise en relation (similaire à BEAVER et FREEMAN [2016]) et des questionnements au sujet du mode de fonctionnement de l'agent sont considérés comme problématiques s'ils ne se trouvent pas dans le premier énoncé utilisateur ;
2. Si l'agent prend l'initiative de la proposition de la mise en relation, l'énoncé de l'utilisateur n'est pas considéré comme problématique.

Expressions d'insatisfaction par rapport à la réponse de l'agent Nous utilisons les expressions d'insatisfaction de l'utilisateur pour détecter les PI. La règle est la suivante :
SI l'agent exprime son incapacité à apporter de l'aide à l'utilisateur ET l'énoncé utilisateur contient des caractères écho et/ou interjections ("pfff") ou des expressions argotiques servant à clore la discussion (ex. un mot anglais "bye"), ALORS l'énoncé utilisateur sera annoté en tant que contenant un PI.

Contexte d'un dialogue

Les répétitions et les reformulations de l'utilisateur représentent un indice de problème d'interaction comme nous l'avons indiqué dans le Chapitre 6.1.3. Nous appliquons plusieurs approches différentes pour la détection des répétitions et des reformulations de l'utilisateur :

Détection des répétitions utilisateur à l'aide des distances linguistiques Pour détecter la répétition de l'utilisateur, nous calculons la distance linguistique entre les énoncés de l'utilisateur à l'aide de l'algorithme de Levenshtein [LEVENSHTEIN, 1966] et de Jaccard améliorée [BRUNET, 2003]. Les seuils choisis empiriquement sur le corpus de développement sont $\leqslant 4$ pour Levenshtein et $\leqslant 0,85$ pour Jaccard. La distance finale pour un énoncé est la distance minimale entre l'énoncé courant et chaque énoncé précédent.

La distance de Levenshtein permet de prendre en compte les fautes de frappe. La distance de Jaccard classique permet de détecter des répétitions des énoncés lorsque la majorité des mots utilisés sont communs entre deux énoncés comparés. L'avantage de la distance de Jaccard améliorée est qu'elle permet de détecter la répétition même si un des énoncés est court et le deuxième est long, par exemple, soit l'énoncé e_1 et A un ensemble des mots appartenant à l'énoncé e_1.
$e_1 = \{A\}$
Soit le deuxième énoncé de l'utilisateur e_2 et A, B, C des ensembles de mots appartenant à l'énoncé e_2, de la façon suivante :
$e_2 = \{B, C, A\}$
L'ensemble $A \in e_1$ et $A \in e_2$, alors l'énoncé e_2 répète en partie l'énoncé e_1.

Exemple 5 *Annotation de la distance de Jaccard*
User [Utterance(U) 1] : *releve confiance*
Agent : *(...)*
User [U2] : *URL*
Agent : *(...)*
User [U3] : *comment opter pour le releve confiance*`<jaccard_distance: 0.66; levenshtein_distance: 22.0>`
Agent : *(...)*
User [U4] : *releve confiance*
`<jaccard_distance: 0; levenshtein_distance: 0>`
Agent : *(...)*

Dans l'**Exemple** 5, le troisième (U3) et le quatrième (U4) énoncés utilisateur sont annotés comme contenant un PI puisqu'ils contiennent une expression commune : "releve confiance".

Exemple 6 *Annotation de la distance de Levenshtein*
User [U1] : *facture papier*
User [U2] : *facture païer*
`<jaccard_distance: 1.0; levenshtein_distance: 2.0>`

Selon le score de la distance Levenshtein, dans l'**Exemple** 6, le second énoncé utilisateur contient un PI. Nous constatons ici que les deux distances permettent de détecter des problèmes d'interaction et sont complémentaires, puisque la distance de Levenshtein permet de détecter des fautes de frappe.

Nous avons aussi testé la pénalité affine gap, également connue sous le nom d'algorithme de Smith-Waterman [SMITH et WATERMAN, 1981]. Elle est utilisée pour faire l'alignement de séquences, c'est-à-dire, pour détecter les insertions et les modifications entre

deux chaînes d'objets. Nous utilisons l'implémentation python de l'algorithme de Smith-Waterman proposée par Forest Gregg et Dedupeio [8] (2016). Nous testons cet algorithme sur 100 dialogues annotés manuellement. Pour un énoncé, on prend la mesure minimale entre tous les énoncés précédents.

$$M = \frac{score(reference, test)}{score(reference, reference)} \tag{6.1}$$

Dans l'équation 6.1 M est la mesure de la pénalité "affine gap" normée. Le résultat du test de la pénalité "affine gap" est présenté dans le Tableau 6.6 page 100.

Mesure	**Distances de Jaccard+Levenshtein**	**Pénalité "Affine gap"**
Rappel	0,33	0,33
Précision	0,94	0,88
F-mesure	0,49	0,48
Exactitude	0,57	0,56

TABLEAU 6.6 – Résultats de l'annotation des PI avec les distances Jaccard+Levenshtein Vs. la pénalité "Affine gap" sur un échantillon de 100 dialogues.

Nous avons également calculé l'accord inter-annotateurs entre les distances Jaccard couplé avec la distance de Levenshtein et la pénalité "affine gap". L'observation des accords entre ces deux annotateurs est de 0,92. La mesure du kappa de Cohen [COHEN, 1960] est alors à 0,76, ce qui correspond à un accord fort. Cette résultat signifie que la mesure de la pénalité "affine gap" donne des résultats comparables à l'utilisation de la distance de Jaccard couplée avec la distance de Levenshtein. Toutefois, il y a des différences entre les PI détectés par la mesure de la pénalité "affine gap" et les mesures de distances Jaccard+Levenshtein. Puisque l'"affine gap" semble être une mesure pertinente pour la détection des PI dans notre corpus, dans nos perspectives, nous ferons une étude plus détaillée des résultats obtenus. Nous étudierons sur la totalité du corpus annoté les résultats que nous pouvons obtenir :

— en remplacement des mesures de distance que nous avons adoptées par la pénalité "affine gap" ;

— en utilisation conjointe de la pénalité "affine gap avec Jaccard améliorée.

Détection des répétitions et des reformulations utilisateur à l'aide de termes métier La répétition est également détectée grâce à la présence des concepts ou des termes métier dans les énoncés de l'utilisateur. Les deux règles prennent en compte l'historique d'un dialogue. Pour trouver les concepts métier, nous nous appuyons sur la liste des termes métier constituée dans l'entreprise et des patrons pour des termes métier multi-mots (ex. espace client).

La première règle s'appuie sur la répétition du même concept métier dans les énoncés utilisateur. Un énoncé utilisateur est annoté comme contenant un PI, s'il contient un concept métier déjà présent dans un des énoncés utilisateur précédent.

Exemple 7 *Detection de PI basée sur la répétition des concepts métier*
User [U1] : *je régler par carte bleu je ne le trouve plus*
Agent : *XXX met plusieurs [URL] modes de paiement à votre disposition. (...)*

8. https ://github.com/dedupeio/affinegap

User [U2] : *[URL]*
Agent : *Je viens de vous rediriger vers la page demandée*
User [U3] : *je veut régler par carte bancaire*
Agent : *(...)*

Dans l'**Exemple** 7 les termes métier "carte bleue" et "carte bancaire" appartiennent au concept "Carte Bancaire" et sont donc synonymiques.

La seconde règle utilise la ponctuation multiple et la présence constante des termes métier pour détecter des problèmes d'interaction. La présence de la combinaison des termes métier avec la ponctuation multiple dans l'énoncé en cours et des termes métier dans un des énoncés précédents permet de désambiguïser la ponctuation multiple liée à l'expression d'OPEM envers un produit de celle, liée à l'expression d'OPEM envers l'interaction. L'algorithme de la règle est le suivant :

SI _ l'énoncé de l'utilisateur contient un terme métier suivi de la ponctuation multiple (ex.!!,??), ET _ un terme métier est présent dans un énoncé précédant, ALORS _ l'énoncé utilisateur est annoté `<probleme-interaction>`.

Exemple 8 *Détection de PI basée sur la répétition de concepts métier et la présence de ponctuation multiple*
User [U1] : *comment faire pour releverle comteur de* `<business_term>`*gaz*`</business_term>` *pas de cle pour ouvrir*
Agent : *(...)*
User [U2] : *commen fair pour relever le conteur de* `<business_term>`*gaz*`</business_term>` *??*

6.2.4 Focus sur le traitement des ambiguïtés

Ici nous décrivons notre approche pour la gestion des ambiguïtés au niveau des énoncés, au niveau de la polarité et au niveau lexical.

Détection des énoncés hors-sujet

Les interactions hors-sujet sont des interactions sur des sujets autres que métier ou de type « small talk »[9] (« Laura, a-tu des enfants? »). La désambiguïsation des énoncés utilisateur hors-sujet est effectuée à plusieurs niveaux du contexte de dialogue. **Au niveau de l'énoncé utilisateur**, la détection des énoncés hors-sujet est basée sur le dictionnaire des jurons (LIWC). Nous classons les jurons par type : ceux qui ne s'utilisent majoritairement que dans le contexte hors-sujet (comme ceux, relevant de la thématique sexuelle) et ceux qui peuvent être utilisés dans les deux cas. Si l'énoncé utilisateur contient un juron de type "hors-sujet", alors l'énoncé est annoté <hors-sujet>.
Au niveau du contexte d'un dialogue, la détection des énoncés hors-sujet est basée sur l'absence de termes métier dans les énoncés de l'utilisateur. En effet, notre hypothèse est que lorsque l'utilisateur n'utilise pas de termes métier, son discours peut être qualifié comme du "small talk", par exemple, "a tu des enfants? quel age a tu?". Si deux énoncés successifs utilisateur ne contiennent pas de termes métier, alors le second énoncé est hors-sujet. Les cas suivants ne sont pas pris en compte : les énoncés contenant uniquement des formules de salutation, les énoncés annotés en relation et ceux, annotés en PI

9. Conversation qui ne concerne pas de choses importantes, souvent entre des personnes qui ne se connaissent pas. [traduit du dictionnaire de Cambridge]

implicite lorsque l'utilisateur demande à l'agent comment il fonctionne (voir la description des règles linguistique de la Section 6.2.3).

Détection des énoncés contenant des OPEM positives envers l'interaction

L'utilisateur peut exprimer une opinion positive concernant l'interaction avec l'agent. Il peut le remercier du service rendu. Nous ne pouvons pas considérer ces cas lors de l'annotation des OPEM de polarité négative.

La détection d'une OPEM de polarité positive envers l'interaction est réalisée à l'aide du dictionnaire LIWC (voir Chapitre 6.2.2 page 94) et d'une liste d'expressions de remerciements. Ces règles sont également liées aux règles de traitement de la négation décrites dans la Section 6.2.4.

Nous ignorons les phrases de type "Merci + (...) + votre réponse", "remercier de/pour" + "votre réponse", "attente de votre réponse".

Gestion de la négation

À la suite de l'annotation en chunks, nous procédons à la gestion de l'influence de la polarité du chunk verbal OPEM$chunk$VN sur les chunk nominaux OPEM$chunk$NP et adjectivaux OPEM$chunk$AP. Cela dépend également de la présence de la négation au sein de OPEM$chunk$VN. Si la polarité de OPEM$chunk$VN est "0" et la négation est présente, alors elle inverse la polarité des éléments qui suivent. Si la polarité du OPEM$chunk$VN est négative, alors la polarité totale de la phrase est également négative. Si la polarité de OPEM$chunk$VN est positive, alors la polarité totale dépend de la polarité des OPEM$chunk$AP et OPEM$chunk$NP. Si un de ces éléments a une polarité négative, alors la polarité totale est négative.
Si OPEM$chunk$AP et OPEM$chunk$NP sont précédés par un chunk verbal ne contenant pas d'OPEM mais une négation, la polarité est inversée.

Soit une OPEM contenant les expressions CVN_{op} et CAP_{op} et/ou CNP_{op} des chunks verbaux, adjectivaux et nominaux contenant une OPEM.
Soit un chunk verbal $\text{CVN} \neq \text{CVN}_{op}$.

Si CVN contient une négation : $Pol(Opem) = \overline{Pol(\text{VN}).Pol(\text{AP}).Pol(\text{NP})}$
Si CVN ne contient pas de négation : $Pol(\text{OPEM}) = Pol(\text{VN}).Pol(\text{AP}).Pol(\text{NP})$

Nous traitons également des cas lorsque l'énoncé utilisateur contient une cible potentielle, précédée ou suivie par une négation, mais ne contient pas de chunks, contenant un terme de LIWC. La prise en compte de cette configuration permet à détecter, par exemple "pas de bonne réponse", "ça n'explique pas", "Non, relis!"

Gestion de l'anaphore

Nous ne gérons qu'un cas d'anaphore. Nous considérons les pronoms démonstratifs comme des cibles potentielles inconnues. Lorsque ce type de cible est suivi d'une cible potentielle exprimée par un verbe ou d'une OPEM, alors nous considérons que la cible potentielle inconnue a pour cible l'interaction, par exemple, "ça n'explique pas".

6.3 Approche non-supervisée : plongements lexicaux pour l'amélioration de la détection des répétitions et des reformulations utilisateur

Pour pouvoir détecter les énoncés utilisateur contenant peu ou pas de mots communs ou des énoncés avec des erreurs d'orthographe, il est nécessaire de pouvoir capter leur sens. Pour évaluer à quel point deux énoncés sont proches sémantiquement, nous analysons la possibilité d'utilisation des approches existantes au calcul de la distance sémantique. Ensuite nous présentons les tests des approches qui nous sont accessibles et finalement, l'approche retenue.

6.3.1 La similarité sémantique. Les approches testées

La similarité sémantique permet la détection des expressions avec un sens proche. La similarité sémantique peut être calculée en se basant sur le corpus ou sur une base de connaissance GOMAA et FAHMY [2013]. Malheureusement, nous n'avons pas d'ontologie métier à notre disposition. En revanche, nous disposons d'un corpus couvrant une année complète de dialogues entre les utilisateurs et la conseillère virtuelle Laura. Nous nous tournons donc vers les mesures de similarité sémantique basées sur notre corpus. Nous testons les modèles suivants :

1. une combinaison de l'approche "sac-de-mots" avec l'indexation sémantique latente (LSI) classiquement utilisée pour la recherche d'informations DEERWESTER et collab. [1990]. Nous utilisons la fonction doc2bow[10] de la librairie gensim pour Python afin de transformer les énoncés en sacs de mots.
   Nous testons ce modèle dans deux configurations : avec et sans la méthode de pondération Tf-Idf (Term Frequency - Inverse Document Frequency).
2. doc2vec LE et MIKOLOV [2014].
3. word2vec en combinaison avec une mesure de similarité géométrique entre deux vecteurs des énoncés obtenue par :
   (a) la somme des vecteurs de mots de l'énoncé et la mesure cosinus pour calculer la similarité
   (b) les mesures "heuristic-max" et "heuristic-avg" SONG et collab. [2016]
4. la distance de Jaccard intégrant les vecteurs de mots

Pour l'entraînement des modèles lors de ces tests nous utilisons 2 112 860 énoncés utilisateur (11 087 419 mots) extrait aléatoirement du corpus non-annoté. Nous avons effectués les pré-traitements suivants sur le corpus : la séparation des apostrophes des mots qui les suivent, l'homogénéisation de la casse, la suppression des chiffres, des hapaxes et des stop-mots les plus répandus, tels que les articles. Nous avons ainsi obtenu 8 888 049 mots.

Dans nos futurs travaux, puisque nos données contiennent beaucoup de variations orthographique, nous envisageons d'effectuer des tests avec des modèles de représentation vectorielle des mots, tels que fastText [BOJANOWSKI et collab., 2017] ou MIMICK [PINTER et collab., 2017], prévus pour gérer les mots hors vocabulaire et ne nécessitant pas des re-entraînements supplémentaires des modèles.

10. https ://radimrehurek.com/gensim/corpora/dictionary.html

LSI Nous avons fait l'entraînement de **doc2bow** (document comme un sac de mots) puis réduit les dimensions des vecteurs avec TF/IDF et finalement appliqué un modèle **LSI** (Latent Semantic Indexing) en demandant la clusterisation des énoncés sur 10 thématiques. Un nombre supérieur de thématiques nécessitait trop de mémoire. Nous avons également réalisé le même test sans la réduction des vecteurs.

La similarité de LSI va de 0 (documents similaires) à 1 (documents sans similarité). En choisissant des seuils de similarité plus élevés pour détecter les reformulations, on crée plus de faux positifs que de détections supplémentaires. Le meilleur résultat est obtenu avec un seuil de similarité de 0,01 et la réduction des dimensions des vecteurs avec TF-IDF. En utilisant ces paramètres, nous obtenons les résultats (par rapport à la version "2.0.0bc" [11] du système à la date du 28/03/2017) présentés dans le Tableau 6.7 page 104.

System	Precision, %	Recall, %	F-score, %	Accuracy, %
2.0.0bc	83,3	65,5	**73,3**	92,5
TF-IDF+LSI	76,4	70,3	73,2	91,9

TABLEAU 6.7 – Résultats obtenus sur le corpus de développement *DevCorpus* lors d'emplois de la similarité LSI pour la détection des PI

doc2vec Nous avons testé le modèle doc2vec LE et MIKOLOV [2014] avec la librairie Python Gensim [12] pour calculer la similarité entre deux énoncés utilisateur. La mesure cosinus allant de -1 à 1, avec 1 correspondant à deux énoncés similaires, est utilisée pour calculer l'angle entre des vecteurs obtenus à partir de deux énoncés utilisateur. L'entraînement de doc2vec avec des dimensions des vecteurs [13] différentes donne des résultats instables. En effet, pour un échantillon de phrases qui n'ont pas fait partie du corpus de l'entraînement, avec 1) deux énoncés identiques, 2) deux énoncés similaires et 3) deux énoncés différents, la similarité cosinus peut varier de négative jusqu'à positive (0,7). L'utilisation des permutations des énoncés de dictionnaires lors de chaque époque d'entraînement dégrade davantage les résultats (voir l'exemple des résultats obtenus dans l'Exemple 9).

Exemple 9 *Les exemples des scores obtenus pour les énoncés de test.*

1. *similaires : «je désire le total de mon compte année» et «total de mon compte annuel» :*
   dimension 20 : score de similarité : -0.10
   dimension 100 : score de similarité : 0.23
   dimension 100 avec shuffle : score de similarité : 0.11
2. *différents : «je désire le total de mon compte année» et «Le technicien est il passé chez moi hier ?» : dimension 20 : score de similarité : -0.44*
   dimension 100 : score de similarité : 0.28
   dimension 100 avec shuffle : score de similarité : 0.40
3. *identiques : «je désire le total de mon compte année» et «je désire le total de mon compte année» :*
   dimension 20 : score de similarité : 0.88
   dimension 100 : score de similarité : 0.79

11. La version du système à base de règles, contenant toutes les règles de la version finale du système mais la version des règles n'est pas encore définitive.
12. https://pypi.python.org/pypi/gensim
13. tous les autres paramètres sont des paramètres par défaut

dimension 100 avec shuffle : score de similarité : 0.26

Nous avons essayé de jouer avec les paramètres d'apprentissage, par exemple, en précisant les paramètres suivants :

```
model = gensim.models.doc2vec.Doc2Vec(window=10, size=100, min_count=2,
workers=2, alpha=0.025,steps=20, min_alpha=0.025, iter=55)
```

Mais nous n'avons pas obtenu d'amélioration des résultats. L'affichage d'énoncés de corpus d'entraînement le plus, le moins et moyennement similaires à un énoncé de test (voir l'Exemple 10) montre que l'énoncé le plus similaire peut apparaitre dans le même contexte que l'énoncé de test.

Exemple 10 *Documents similaires/dissimilaires pour le modèle Doc2Vec (dm/m,d100,n5,w5,mc2,s0.001,t2)*
L'énoncé de test : «mot de passe insoluble»
Les énoncés du corpus d'entraînement :
Le plus similaire : (26937, 0.6739590764045715) : «j ai 3 tentatives»
Médiane (108865, 0.020388584583997726) : «changement de locataires»
Le moins similaire (241566, -0.6399707794189453) : «communiquer chiffres service relevé confiance»

Lors de l'entraînement d'un modèle de word2vec qui sera décrit dans le paragraphe suivant, nous avons vu que l'augmentation du corpus permet de passer des mots de contexte (par exemple, pour le mot "compte", les mots les plus similaires sont "acceder" et "creer", lorsque le modèle est entrainé sur un petit corpus (499 941 mots)), vers des variantes orthographiques (voir l'Exemple 11) en tant que mots les plus similaires identifiés par la similarité cosinus. Il est possible qu'afin d'obtenir des reformulations des énoncés lors de recherche des énoncés similaires, le modèle doc2vec nécessite plus 2 112 860 énoncés utilisateur que nous lui avons fourni pour l'entraînement. Les auteurs du modèle LE et MIKOLOV [2014] soulignent que pour obtenir de meilleurs résultats, il faut utiliser la combinaison de deux modèles : PV-DM ("distributed memory") et PV-DBOW ("distributed bag of words"), en concaténant ensuite les vecteurs résultants. Nous avons utilisé uniquement le modèle *PV-DM* car les auteurs ont trouvé que ce modèle donne de meilleurs résultats que le modèle PV-DBOW. LAU et BALDWIN [2016] ont effectué des évaluations des deux modèles et ont trouvé que le modèle PV-DBOW donne de meilleurs résultats. Les auteurs soulignent que pour avoir de bonnes performances, le corpus d'entraînement doit être grand (jusqu'à 1 milliard de documents), il faut régler les hyper-paramètres et les documents doivent être longs. Dans notre cas un document est un énoncé utilisateur. Ce modèle semble donc ne pas être bien adapté pour notre tâche et corpus. Néanmoins, dans notre future travail nous envisageons d'effectuer plus de tests, notamment en adaptant les hyper-paramètres automatiquement, en utilisant des algorithmes telles que *Adam* [KINGMA et BA, 2014], une méthode d'optimisation stochastique, ou *Hyperopt* [BERGSTRA et collab., 2013], un algorithme pour l'optimisation distribué asynchrone des hyper-paramètres.

word2vec Nous avons vectorisé les mots des énoncés utilisateur en utilisant la bibliothèque standard de word2vec [MIKOLOV et collab., 2013a] pour python [14]. Les paramètres d'entraînement du modèle word2vec sont les suivants : size=100, cbow=0, verbose=False.

14. https://pypi.python.org/pypi/word2vec

Les résultats de l'entraînement sont satisfaisants puisque pour des mots de test, les 7 - 8 premières propositions de mots proches en sens sont majoritairement des variantes orthographiques (voir l'Exemple 11).

Exemple 11 *Résultat pour le mot "**compte**"*
[('comte', 0.868725677999538),
('compt', 0.8067718166129615),
('conte', 0.7844866744216233),
('copte', 0.7534716224064946),
('espace', 0.7409383738062161),
('cpte', 0.7382213393109656),
('compe', 0.7338182643458314),
('cmpte', 0.7263688871275367),
('considération', 0.7158105046431427),
('moncompte', 0.709802183918911)]

Pour calculer la similarité entre deux énoncés nous reprenons l'expérience de SONG et collab. [2016]. Nous testons trois formules de calcul de similarité sémantique entre deux vecteurs des énoncés.
La mesure **cosinus** classique :

$$cos(\boldsymbol{s}_1, \boldsymbol{s}_2) = \frac{\boldsymbol{s}_1^\top \cdot \boldsymbol{s}_2}{||\boldsymbol{s}_1|| \cdot ||\boldsymbol{s}_2||} \tag{6.2}$$

où s_1 et s_2 sont des vecteurs des énoncés S_1 et S_2. Les vecteurs de chaque énoncé sont obtenus en faisant la somme des vecteurs des mots de l'énoncé.

SONG et collab. [2016] proposent de complexifier la formule 6.2, ce qui augmenterait l'interaction entre les deux énoncés.
La mesure **"heuristic-max"** :

$$sim(\mathbf{S}_1, \mathbf{S}_2) = \frac{1}{n_1} \sum_{i=1}^{n_1} max_{j=0}^{n_2} \{cos(\boldsymbol{w}_i, \boldsymbol{v}_j)\} \tag{6.3}$$

où w_i et v_j sont des mots dans s_1 et s_2. n_1 et n_2 sont le nombre de mots dans S_1 et S_2. Pour chaque mot w_i dans s_1, nous cherchons un mot le plus lié sémantiquement dans s_2 (la partie max de la formule). La similarité entre les deux mots est déterminée par la mesure cosinus. La similarité entre deux énoncés est calculée en moyennant les scores de similarité des mots dans s_1.

La mesure **"heuristic-avg"** consiste à remplacer le "max" par la "moyenne" dans la formule. Pour la variante avec le calcul de **la moyenne (avg)**, nous avons calculé les résultats pour plusieurs seuils de 0,6 à 0,95. Le seuil de 0,9 présente de meilleurs résultats que d'autres seuils. Ce qui représente **7** détections **correctes** supplémentaires mais en même temps, **12 faux-positifs** supplémentaires. Les résultats obtenus pour le calcul du **max** sont très différents de la moyenne : le rappel augmente beaucoup mais la précision est plus basse que celle de "heuristic-avg" (voir Tableau 6.8).

Un certain nombre de recherches a été dédié à l'amélioration des mesures de distances linguistiques afin qu'elles puissent prendre en compte des synonymes, par exemple, les travaux de LU et collab. [2013]. Dans le même but d'amélioration de la mesure de distance de Jaccard, nous avons combiné le calcul de la distance de Jaccard avec le calcul de similarité cosinus entre deux mots vectorisés par word2vec, entraîné sur le corpus brut. Nous permettons à la distance de Jaccard de faire le calcul non seulement entre des mots identiques mais aussi entre les mots similaires.

System	Precision, %	Rappel, %	F-score, %	Accuracy, %
cosinus	64,2	76,3	69,7	89,7
heuristic-max	69,2	69,8	69,5	90,4
heuristic-avg	71,5	61,9	66,3	90,2

TABLEAU 6.8 – Résultats obtenus sur le corpus de développement *DevCorpus* en fonction de la mesure de similarité sémantique

On calcule la distance (d) de Jaccard améliorée selon la formule proposée par BRUNET [2003] :
Pour deux énoncés à comparer, nous constituons deux listes de mots a et b, chacune ne contenant pas de doublons. Soit ab, la liste de mots commun aux vocabulaires des listes a et b, alors

$$d = ((a - ab)/a) + ((b - ab)/b) \tag{6.4}$$

Le score résultant est compris entre 0 ($a = b$) et 2 ($ab = \emptyset$)

Soit u et v les mots des ensembles a et b respectivement. Soit u' et v' les vecteurs de mots u et v respectivement. Pour chaque vecteur u' de a nous calculons la distance cosinus avec tous les vecteurs v' de b. Pour chaque v' de b nous calculons la distance cosinus avec tous les vecteurs u' de a. Nous gardons les mots u et v, pour lesquels la distance entre leurs vecteurs u' et v' respectifs est minimale. Nous considérons ces mots comme des mots en commun entre les ensembles a et b. Ensuite nous calculons la distance de Jaccard améliorée entre les ensembles a et b. Nous avons testé la combinaison de plusieurs seuils. Le résultat est présenté dans la Figure 6.6

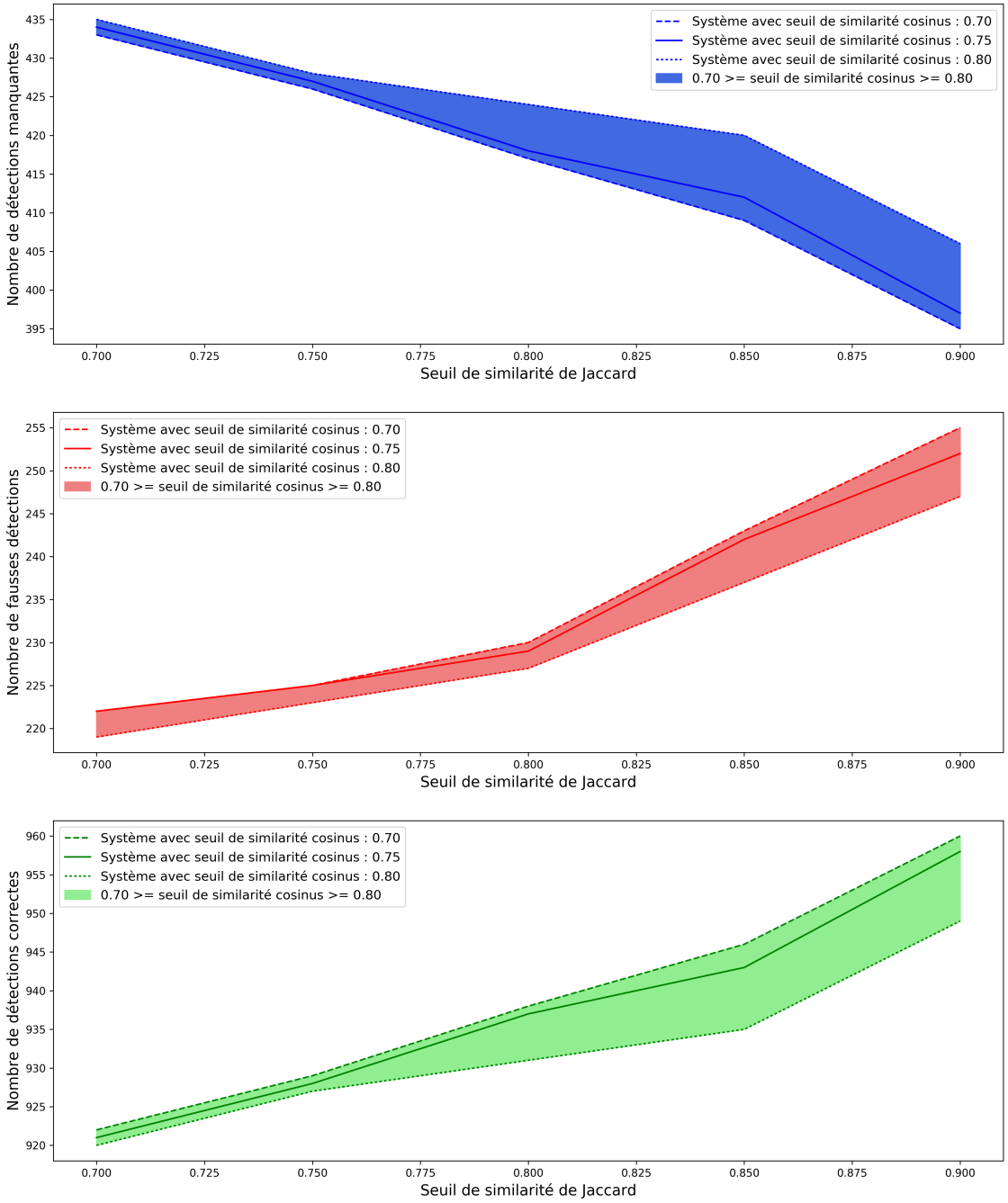
La Figure 6.6 montre que le meilleur résultat est obtenu avec le seuil de 0.80 pour la distance de Jaccard, et le seuil de 0.75 pour la distance de similarité cosinus entre deux vecteurs. Cette configuration permet d'obtenir 50 détections supplémentaires correctes et 51 faux-positifs supplémentaires, ce qui représente **1,02 fois plus de faux positifs** que de détections correctes.

Nous testons la même mesure de similarité en utilisant également le modèle de word2vec entraîné sur le corpus prétraité : séparation des apostrophes des mots qui suivent, homogénéisation de la casse, suppression des chiffres, suppression des hapaxes. Puisque le test précédent a montré que la combinaison des seuils la plus efficace était **"Jaccard 0.80 + Cosinus 0.75"**, nous avons essayé directement cette combinaison avec le modèle réentraîné. Le résultat s'est amélioré. Le nombre des détections correctes est 41, alors que le nombre des faux-positifs est 20. Ce qui représente un **ratio de 0,49**. Lorsque nous testons Jaccard 0.9 pour le même seuil de similarité cosinus, le ratio est moins bon : 0,74. Toutefois, la prévalence des faux positifs sur les détections correctes ne nous permet pas de retenir ce modèle.

La distance de Jaccard [JACCARD, 1901] et de la similarité sémantique prennent en compte le contexte global des mots dans un énoncé. Aucune de ces mesures ne prend en compte l'ordre de mots. ISLAM et INKPEN [2008] émettent l'hypothèse que l'ordre des mots est peu important dans la détermination de la similarité sémantique des textes courts. Nous pensons néanmoins qu'il serait intéressant d'étudier la prise en compte de l'ordre de mots, notamment pour les cas où nous aimerions de distinguer le thème et le rhème dans les énoncés utilisateur afin de différentier les répétitions de développement du dialogue.

Levenshtein sur mots Nous avons testé l'application de la distance de Levenshtein non pas sur les caractères mais sur les mots. Les résultats de F-mesure sur le corpus de dé-

FIGURE 6.6 – Distance de Jaccard combinée avec la distance cosinus Word2Vec



veloppement étaient inférieurs à 60,0%. En effet, lorsque le seuil de détection était très strict (< 3), la distance n'avait pas d'apport supplémentaire par rapport aux mesures déjà utilisés. L'augmentation de seuil menait à une forte hausse du nombre de faux positifs.

6.3.2 L'approche choisie

Puisque nous disposons d'un grand corpus, nous avons opté pour la distance sémantique géométrique. Le calcul de la distance sémantique entre deux énoncés se base sur la représentation des mots dans un espace vectoriel. Nous résumons ici l'approche choisie pour la détection des énoncés sémantiquement proches et la manière dont l'espace vectoriel a été créé.

Annotation des énoncés sémantiquement proches Nous utilisons le modèle Word2Vec de MIKOLOV et collab. [2013a] pour transformer les mots de notre corpus en vecteurs. Ensuite nous calculons la somme des vecteurs de mots pour chaque énoncé, selon la formule classique de mesure cosinus, avec laquelle, comme nous l'avons démontré, nous avons obtenu de meilleurs résultats. Les vecteurs de mots sont additionnés pour obtenir le vecteur de l'énoncé. La distance cosinus entre les vecteurs de deux énoncés avec un seuil de 0,85 détermine si deux énoncés sont similaires. Le seuil a été établi empiriquement. Si c'est le cas, alors nous annotons le deuxième énoncé comme un énoncé contenant un problème d'interaction.
Si $dcos(e_1, e_2) > 0,85$ alors l'énoncé est annoté PI

Création de l'espace vectoriel Pour entraîner un modèle word2vec sur notre corpus, nous avons suivi les étapes suivantes :

1. Nous utilisons une bibliothèque standard de word2vec pour python [15].
2. Nous effectuons l'entraînement sur un corpus de 2 112 860 énoncés utilisateur (11 087 419 mots). Une liste standard de stop-mots (les mots les plus fréquent tels que les articles) a été appliquée sur le corpus.
3. Les paramètres d'entraînement word2vec : size=100, cbow=0, verbose=False.

L'efficacité du modèle de word2vec est fonction de la taille du corpus d'apprentissage. En effet, lors de l'entraînement du modèle sur un corpus de 30 000 mots, les mots les plus proches des termes métiers (en terme de distance cosinus) étaient des termes décrivant le contexte du terme métier. Lorsque la taille du corpus a été augmentée à 50000 mots, des variantes orthographiques des termes métier sont apparues en tant que plus proches voisins. Le modèle skip-gram que nous avons choisi est décrit par MIKOLOV et collab. [2013a] comme un modèle prédisant les mots voisins à partir d'un terme donné.

Le modèle vectoriel permet ainsi d'aller vers la plus grande généralisation de la représentation de l'information dans les dialogues du corpus. Nous avons choisi délibérément de ne pas utiliser des modèles vectoriels pré-entraînés pour ne pas apporter du bruit lié à la polysémie, par exemple, des termes métier (comme le terme "puissance", dont le sens est réduit au sens de notre corpus).

15. https ://pypi.python.org/pypi/word2vec

6.4 Conclusion

Nous avons présenté dans ce chapitre l'architecture du système DAPI. La détection des problèmes d'interaction est organisée selon l'étendue du contexte pris en compte : le contexte d'un énoncé, d'une paire adjacente ou d'un dialogue. Cette détection repose sur une approche hybride qui d'un côté, comprend des règles linguistiques et de l'autre côté, la présentation sémantique des mots apprise de manière non-supervisée.

La détection des problèmes d'interaction utilise des résultats d'annotation effectués lors des prétraitements et de la désambiguïsation préliminaire.

Nous verrons comment les éléments introduits afin de détecter les différents types de problèmes d'interaction interagissent entre eux dans le chapitre 7. Dans le chapitre suivant nous présenterons et discuterons les résultats de nos expérimentations.

Chapitre 7

Évaluation et résultats

Sommaire

Dans le chapitre précédent, nous avons présenté le développement du système DAPI. Dans ce chapitre nous décrivons des tests pour évaluer l'apport de chaque module au score final du système (vo ir la section 7.2). Nous discutons des résultats obtenus dans la section 7.3 et concluons dans la section 7.5.

La section suivante décrit les annotations manuelles du corpus de référence.

7.1 Analyse quantitative des problèmes d'interaction dans le corpus de référence

Deux corpus (de développement et de référence) ont été annotés lors de deux campagnes d'annotation par un annotateur expert sémiologue ayant une excellente connaissance métier d'EDF. L'annotateur était guidé par le guide d'annotation. Nous avons décrit le procédé de constitution du guide d'annotation dans le Chapitre 5 page 67. Nous décrivons ici le corpus de référence. Il est constitué de 3 000 dialogues choisis aléatoirement et ne faisant pas partie du corpus de développement.

Nombre de	**Corpus de référence**
Dialogues	3 000
Paires Adjacentes (PA)	8 630
Dialogues contenant au moins un PI	845
Énoncés contenant un PI	1 349
Énoncés contenant un PI dans un dialogue problématique (en moyenne)	1.5

TABLEAU 7.1 – Informations statistiques du corpus de référence annoté manuellement

Comme le montre le Tableau 7.1, où un PA représente 1 énoncé système et 1 énoncé utilisateur, la part des dialogues contenant au moins un problème d'interaction est assez bas : 28% dans le corpus de référence. La part des énoncés problématiques est encore moindre : 16%. Il y a très peu de problèmes d'interaction explicites : 6%. 84% des dialogues contenant un PI, ne contiennent qu'un ou deux énoncés problématiques. La Figure 7.1 représente la répartition des dialogues contenant un PI en fonction du nombre d'énoncés annotés comme problématiques. Il est, de plus, important de détecter un PI dès son apparition : 24% des PI apparaissent dans le second énoncé utilisateur, 22% au milieu d'un dialogue et 27% dans le dernier énoncé utilisateur.

Il est rare que l'apparition d'un PI se manifeste autrement que par un problème d'interaction implicite. La Figure 7.2 montre qu'en effet le nombre de relations avec une cible-interaction explicite et avec une cible-interaction implicite est pratiquement équivalent (~3%). Lorsqu'un problème d'interaction apparaît une fois, dans 60,7% des cas il se poursuit encore au moins une fois.

Le peu d'expressions d'opinion/émotion de l'utilisateur envers l'interaction (4% de toutes les annotations) ne signifie pas que l'utilisateur exprime rarement ses opinions/émotions en général, car les étiquettes <relation> avec la cible "produits/services" représentent 28% de toutes les annotations manuelles.

La source est présente dans 76% des relations qui ont pour cible "produit/services" et seulement dans 26% des relations qui ont pour cible l'interaction. En effet, lorsque l'utilisateur exprime son insatisfaction des produits, il utilise souvent la structure "je" + la description d'un problème rencontré, par exemple, "j'ai un problème de..." ou "je n'arrive pas

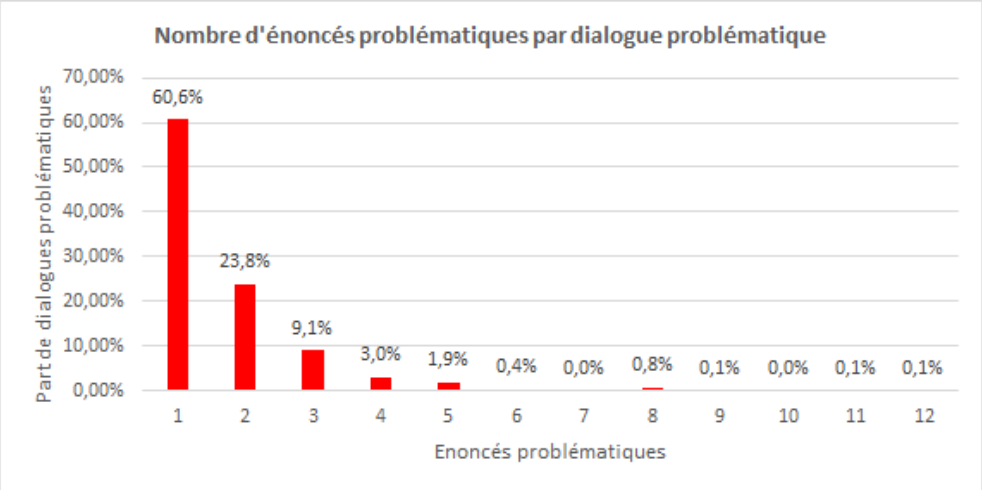


FIGURE 7.1 – Répartition des dialogues problématiques en fonction du nombre d'énoncés problématiques.

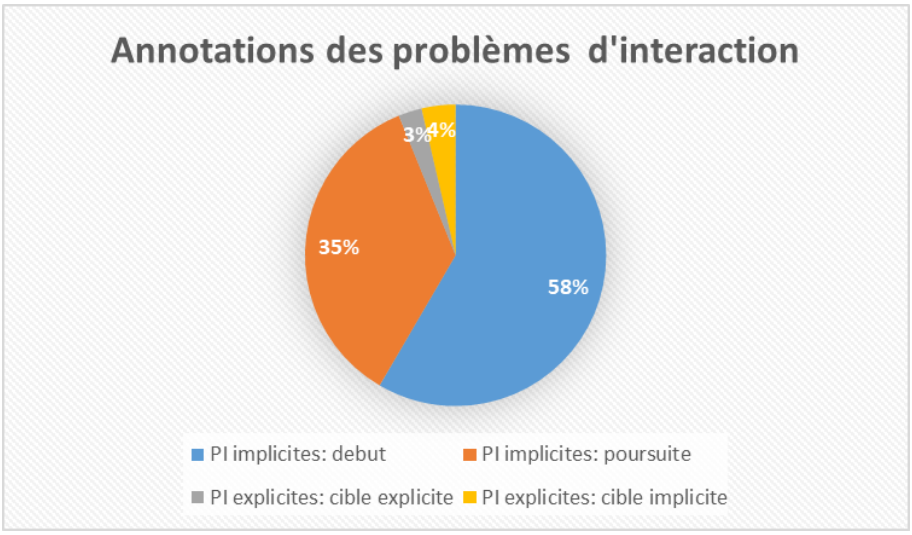


FIGURE 7.2 – Distribution des étiquettes de PI

à...", alors que lorsque son opinion concerne l'interaction, la structure de la phrase commence souvent par "vous" suivi d'une critique, par exemple "vous ne répondez pas à ma question". En revanche, sur les Figures 7.3 page 114 et 7.4 page 114, on voit qu'il y a beaucoup plus de relations avec une polarité négative et une cible "produits/services" que de relations avec une cible "interaction". En effet, le plus souvent, les utilisateurs cherchent de l'aide auprès de la conseillère virtuelle, d'où le nombre de relations de polarité négative et la cible "produits/services". Leurs remarques concernant l'interaction sont également liées aux performances de l'agent virtuel et à sa capacité à leur apporter des réponses attendues. Les relations avec une polarité positive et une cible "interaction" ne sont que 2% moins nombreuses, ce qui confirme que les utilisateurs mènent une conversation naturelle avec le chatbot et n'hésitent pas à le remercier et à exprimer leur appréciation.

Les annotations manuelles du corpus de référence et du corpus de développement montrent que les types d'émotions avec l'interaction pour cible sont différents de ceux avec une cible "produits/services" (voir la Figure 7.5 page 115 et la Figure 7.6 page 115).

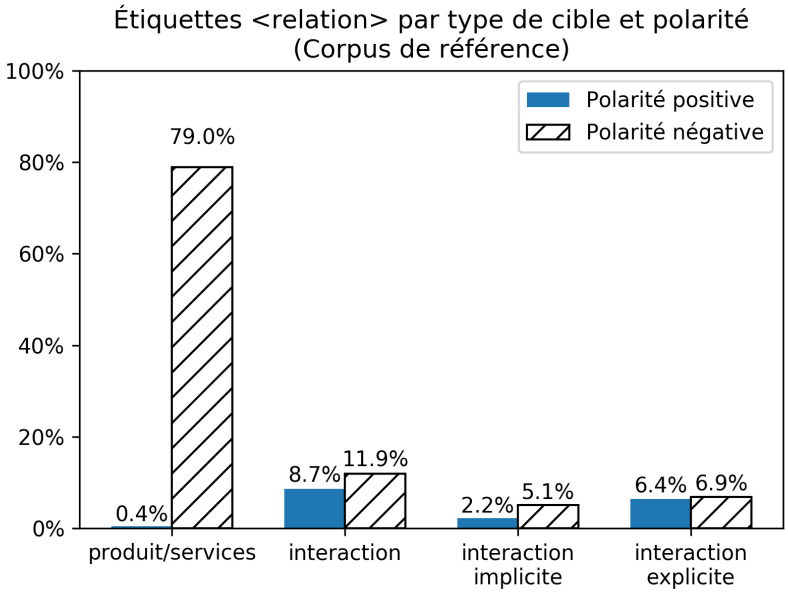


FIGURE 7.3 – Nombre d'étiquettes <relation> dans le corpus de référence en fonction de la polarité d'OPEM et du type de la cible. Les relations avec une cible-interaction explicite et avec une cible-interaction implicite sont des sous-catégories de relations avec une cible-interaction

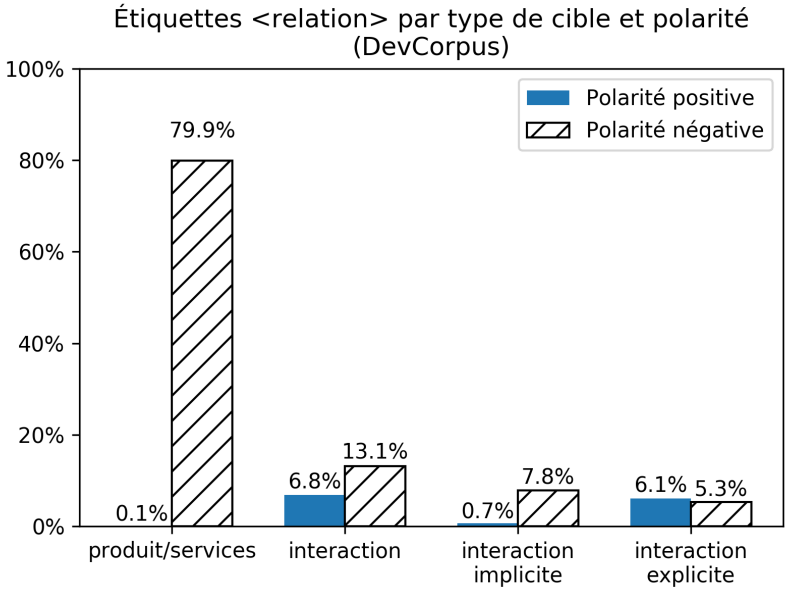


FIGURE 7.4 – Nombre d'étiquettes <relation> dans le corpus de développement en fonction de la polarité d'OPEM et du type de la cible. Les relations avec une cible-interaction explicite et avec une cible-interaction implicite sont des sous-catégories de relations avec une cible-interaction

L'émotion de type "surprise négative" semble ne pas être caractéristique des OPEM avec une cible "interaction". Inversement, les problèmes d'interaction provoquent davantage de colère chez les utilisateurs. Ainsi, les principales émotions négatives caractéristiques de l'interaction sont l'insatisfaction, la colère et le désarroi. CHOWDHURY et collab. [2016] s'appuient sur les mêmes émotions négatives lors de la construction de leur système de prédiction de la satisfaction de l'utilisateur dans les dialogues des centres d'appel, si nous considérons que l'émotion de frustration est équivalente à celle du désarroi.

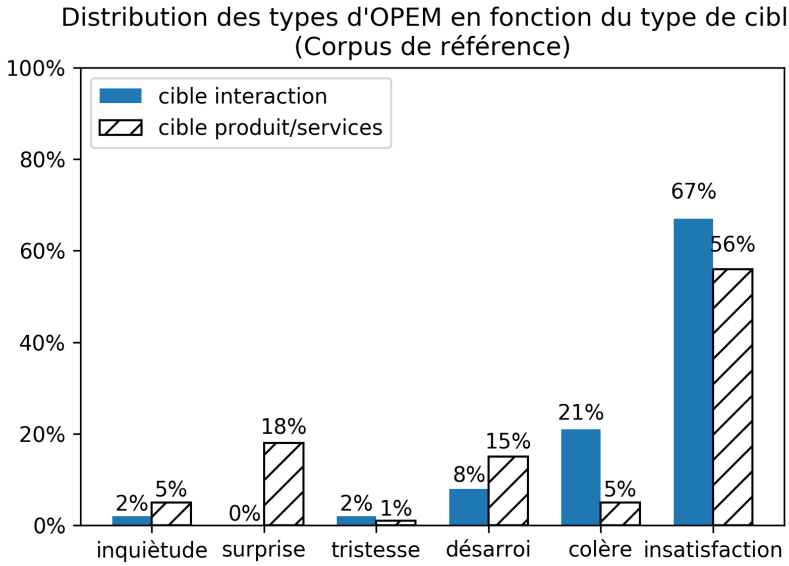


FIGURE 7.5 – Distribution des types d'OPEM dans le corpus de référence, présentées en fonction du type de cible

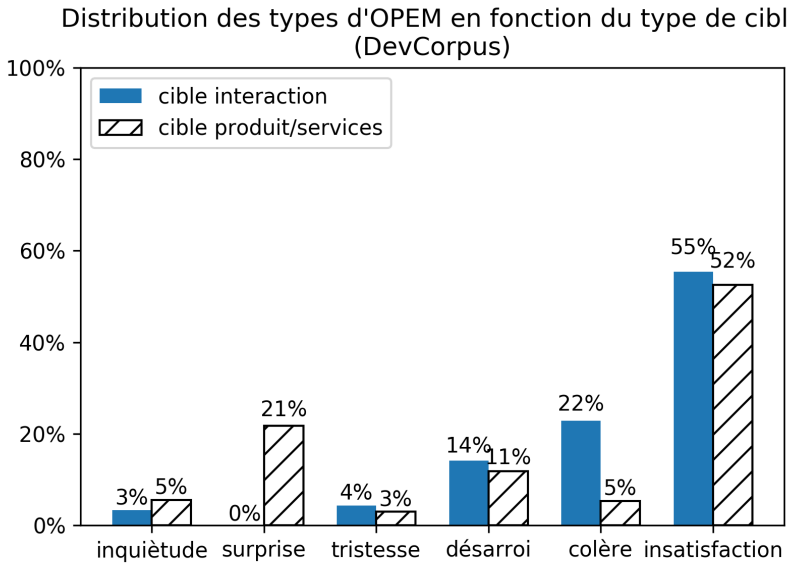


FIGURE 7.6 – Distribution des types d'OPEM dans le corpus de développement, présentées en fonction du type de cible

La forme des courbes sur la Figure 7.7 montre que l'expression des problèmes d'interaction est globalement construite avec des énoncés de la même longueur que l'utilisateur utilise habituellement dans son langage. Ce qui est logique, puisque les formes les plus répandues des PI sont les répétitions et les reformulations. La Figure 7.8 page 116 montre que les énoncés problématiques sont plutôt courts.

La longueur maximale des énoncés où il est possible de trouver une relation de polarité négative et la cible "interaction" est plus courte que celle des problèmes d'interaction implicites : le maximum de 46 mots à comparer avec un maximum de 106 mots. Il

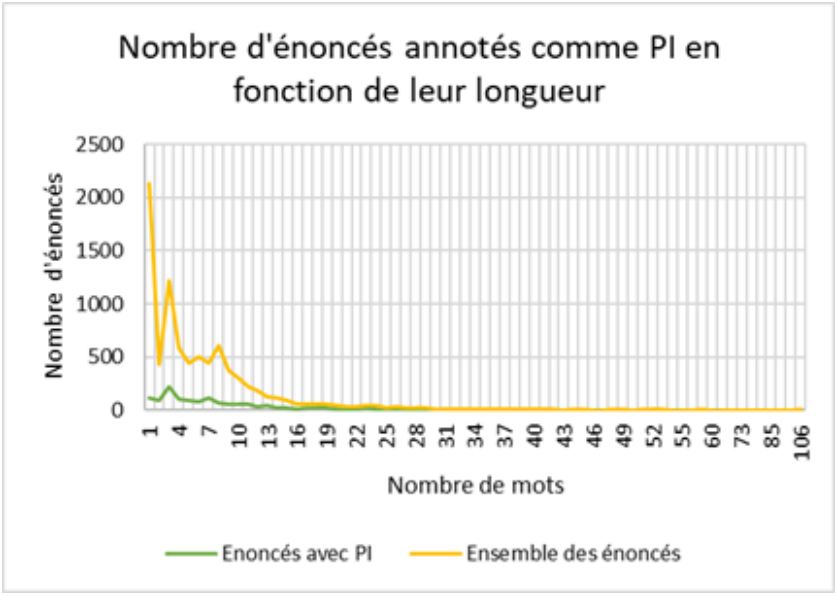


FIGURE 7.7 – Annotations des problèmes d'interaction dans les énoncés utilisateur en fonction de leur longueur en nombre de mots

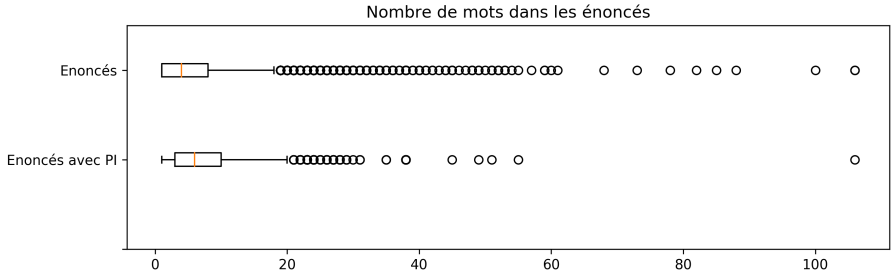


FIGURE 7.8 – Nombre de mots dans tous les énoncés (la figure du haut) et dans les énoncés contenant un PI (la figure en bas)

existe aussi des énoncés contenant une relation négative avec une cible "interaction" et ne contenant aucun mot, puisque leur contenu consiste en des signes de ponctuation. Comme le nombre d'énoncés contenant une relation de polarité négative et avec la cible "interaction" est très inférieur au nombre total des énoncés de tout type, nous présentons les courbes du nombre d'énoncés en fonction de leur longueur sur trois axes sur la Figure 7.9. La courbe des énoncés contenant une relation est très irrégulière. Sa forme est différente de celle de la courbe du nombre total des énoncés. L'utilisateur exprime la majorité de ses opinions/émotions négatives dans les énoncés d'une longueur comprise dans un intervalle entre 0 et 10 mots qui ne constitue pas une caractéristique spécifique à l'opinion/émotion. La majorité des énoncés utilisateur sont compris dans le même intervalle.

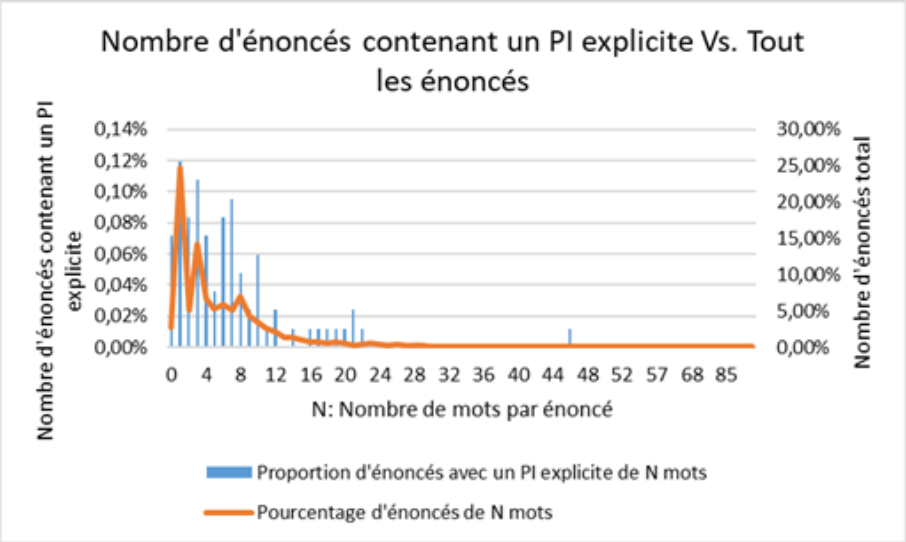


FIGURE 7.9 – Annotations des relations de polarité négative et avec la cible "interaction" en fonction de la longueur des énoncés en nombre de mots. Le nombre des énoncés est normalisé en fonction du nombre d'énoncés total.

7.2 Méthode et résultats de l'évaluation finale

L'évaluation de nouveaux systèmes est faite habituellement par comparaison avec des systèmes connus et facilement applicables à de nouvelles données ou en comparant à un résultat de l'état de l'art sur les même données, si ces dernières sont facilement accessibles pour la communauté. C'est le cas, par exemple, du corpus Let's Go [SCHMITT et collab., 2012], contenant des dialogues des utilisateurs avec un système proposant des informations sur les horaires des bus. La tâche du système DAPI est de détecter des énoncés utilisateur contenant un problème d'interaction. Nous n'avons pas connaissance d'un système qui puisse nous servir de référence pour la détection de PI dans un chat écrit en français avec un conseiller virtuel.

Pour établir des résultats pouvant servir de référence vis-à-vis de nos données, nous séparons la tâche de détection des PI en deux composantes : la détection des problèmes d'interaction implicites et explicites. Comme nous avons vu précédemment dans les Chapitres 3 page 29 et 6 page 81, les principaux indices des PI implicites sont les répétitions et les reformulations des utilisateurs. Les opinions et les émotions des utilisateurs sur l'interaction sont les principaux indices des PI explicites. Nous choisissons comme références pour chacune des composantes des méthodes connues et facilement reproductibles. Pour la détection des problèmes d'interaction implicites, nous appliquons la distance classique de Jaccard [JACCARD, 1901] pour détecter les répétitions et souligner l'intérêt des distances que nous utilisons pour la même tâche. Pour la détection des PI explicites, nous appliquons la classification Naïve Bayésienne, car cette dernière est couramment utilisée pour l'analyse des sentiments [SAAD, 2014].

Le seuil de 0,15 pour la distance de Jaccard est déterminé sur la base du compromis entre le rappel et la précision sur le DevCorpus. La validation croisée est appliquée à la classification Bayesenne sur le Corpus de test divisé en 10 échantillons. En tenant compte des faibles résultats des deux approches (voir Table 7.2), nous choisissons la configuration de base [1] de notre système (DAPI-1) comme référence.

1. des règles linguistiques uniquement

Afin d'évaluer les contributions du correcteur d'orthographe et de la représentation des plongements de mots, nous comparons quatre versions de notre système : **DAPI-1** système constitué uniquement de règles linguistiques, **DAPI-2** intégrant le correcteur d'orthographe, **DAPI-3** intégrant le calcul du score de similarité sémantique mais pas le correcteur d'orthographe et **DAPI-4** combinant à la fois le correcteur d'orthographe et le calcul du score de similarité sémantique. Les systèmes sont évalués sur le *corpus de référence* en calculant la précision (voir la Formule 7.2), le rappel (voir la Formule 7.1), la F-mesure (voir la Formule 7.3) VAN RIJSBERGEN [1979] et l'exactitude (Accuracy) pour la détection des PI au niveau de l'énoncé.

$$Rappel = \frac{\text{Vrai Positif}}{\text{Vrai Positif} + \text{Faux Négatif}} \tag{7.1}$$

$$Précision = \frac{\text{Vrai Positif}}{\text{Vrai Positif} + \text{Faux Positif}} \tag{7.2}$$

$$F_1 = 2 * \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{7.3}$$

$$Exactitude(Accuracy) = \frac{\text{Vrai Positif} + \text{Vrai Négatif}}{\text{Vrai Positif} + \text{Vrai Négatif} + \text{Faux Positif} + \text{Faux Négatif}} \tag{7.4}$$

Les résultats de détection des PI sont affichés dans la **Table 7.2**. L'utilisation de plongements de mots (DAPI-3 et DAPI-4) améliore sensiblement les performances du système. DAPI-3 obtient le meilleur score pour la F-mesure. Cependant, DAPI-4 permet un rappel plus élevé, ce qui est important dans notre contexte (il est important de détecter le maximum de PI existants). Il est à noter que nous avons également expérimenté l'entraînement de word2vec sur le corpus prétraité avec le correcteur orthographique mais les résultats du calcul du score de similarité sémantique ont chuté. En effet, le correcteur orthographique apporte du bruit dans le texte. Ce bruit ne permet plus d'"apprendre" la sémantique des expressions.

System	Précision	Rappel	F-mesure	Exactitude (Accuracy)
Naïve Bayes	25.9	14.6	18.6	90.1
Jaccard	55.5	38.6	45.6	79.2
DAPI-1	72.4	63.6	67.7	90.1
DAPI-2	72.0	65.4	68.5	90.2
DAPI-3	72.0	77.0	74.4	91.4
DAPI-4	71.1	**77.8**	74.3	91.3

TABLEAU 7.2 – Résultats en % de la détection des PI dans le corpus de référence.

7.3 Discussion

Les énoncés détectés en tant que similaires lors de l'utilisation de la similarité sémantique peuvent être caractérisés de la façon suivante : les répétitions des utilisateurs avec un contexte fortement mal orthographié (les règles basées sur les distances linguistiques detectent les cas les plus simples de répétition) ; reformulations contenant des mots avec

la même racine (ex. le mot "payer" dans l'énoncé utilisateur " je ne trouve pas ma facture pour la *payer* en ligne" et "paiement" dans "je ne veux pas le télépaiement je veux le *paiement* par carte bleu", le score de similarité est 0.877) et les reformulations contenant au moins une expression en commun (ex. l'expression "je souhaite" dans l'énoncé utilisateur suivant : "bonjour, *je souhaite* voir le récapitulatif de mes prélèvement/ "*je souhaite* savoir combien je suis relevé par mois", le score de similarité est 0.869).

Les règles linguistiques basées sur la répétition des concepts commerciaux détectent également les reformulations. Ce sont des reformulations contenant des termes métier ayant une racine commune (ex."pourquoi paie t on d'avance l'abonnement"/ "paiement abonnement d'avance", où les mots avec la racine commune sont "payer" et "payement") L'utilisation conjointe des deux approches pour la détection de la reformulation de l'utilisateur en tant que marque de problèmes d'interaction contribue à la robustesse du système pour faire face aux défis du corpus «in-the-wild». Cependant, nos deux approches pour la détection de la reformulation de l'utilisateur (la répétition du concept métier et la similarité sémantique) créent encore beaucoup de faux positifs difficiles à gérer.

Les cas de faux positifs lors de la détection des reformulations sont les suivants :

— clarification de sa demande par l'utilisateur, par exemple "rendez vous"/ "date de rendez vous"

— questions différentes sur la même thématique : "pourkoi mon compte est il associe avec celui dexma mere?"/ "Pourquoi mon compte est il associe avec celui de mon fils???"

— l'évaluation de discussion : "je souhaite consulter le contrat [CODE]"/ "le contrat [CODE] n'est pas proposé"

Pour les détecter des reformulations contenant une nouvelle information, il faudrait prendre en compte le thème et le rhème de l'énoncé.

En ce qui concerne le modèle conjoint des spécificités du langage de chat et de l'historique de dialogue, ils contribuent, par exemple, à détecter une irritation de l'utilisateur vis-à-vis de l'interaction avec le chatbot (règle combinant plusieurs ponctuations et termes métier). En particulier, l'indice de ponctuation multiple joue un rôle important dans la détection (78,5% des détections correctes faites avec les règles exploitant les spécificités du langage de tchat, sont faites en tenant compte de l'indice de ponctuation multiple).

Un certain nombre des problèmes d'interactions sont liés à la configuration du chatbot et à l'incompréhension de son fonctionnement par l'utilisateur (voir l'Exemple 12). Dans cet exemple, l'utilisateur ne comprend pas que l'agent ne peut traiter les réponses "oui" et "non", et attend que l'utilisateur soit clique sur un des liens proposés, soit reformule sa demande. Nous ne détectons pas ces cas car ils ne sont pas génériques.

Exemple 12 *Exemple d'incompréhension de l'utilisation de l'agent*
Agent : En savoir plus sur [...] (l'énoncé de l'agent représente un lien cliquable)
User : non

En outre, les faux négatifs se traduisent par des reformulations des énoncés dont le thème semble être plus rare dans le corpus. Pour le confirmer une étude supplémentaire sera nécessaire. Les erreurs d'orthographe, l'utilisation des abréviations en parallèle avec des formes complètes de certains termes et parfois une confusion des termes par l'utilisateur empêchent également un certain nombre de détections.

7.4 Recherche d'indices supplémentaires des problèmes d'interaction pour une ouverture des perspectives

Dans le cadre du tchat avec la conseillère virtuelle, les utilisateurs sont amenés à rédiger un texte de longueur variable. CUISINIER et collab. [2010] et FARTOUKH et collab. [2012] ont démontré l'influence des émotions sur les processus cognitifs lors de la rédaction des textes. Selon WENGELIN et collab. [2009], la production écrite est une activité impliquant des processus cognitifs complexes. Puisque les processus cognitifs et le processus moteur lors de la génération de textes sont liés [WALLOT et GRABOWSKI, 2013], nous avons émis l'hypothèse que le temps de rédaction d'un message par un utilisateur peut être révélateur d'une émotion négative liée à un PI.

Le système d'agent conversationnel enregistre le temps de début d'une paire adjacente (PA) qui commence par un énoncé client. Nous allons donc calculer la différence de temps entre deux énoncés utilisateur (e(U)) : $e_1(U) - e_0(U)$. Nous catégorisons les différences de temps en différents groupes :

— entre deux énoncés non-problématiques ;

— entre un énoncé non-problématique et un énoncé avec un problème d'interaction implicite ;

— entre un énoncé non-problématique et un énoncé avec un problème d'interaction explicite ;

— entre un énoncé avec un problème d'interaction implicite et un énoncé non-problématique ;

— entre un énoncé avec un problème d'interaction explicite et un énoncé non-problématique ;

La synthèse des résultats est présente dans le tableau 7.3.

Type de 1-er énoncé	Type de 2-nd énoncé	Min	1 quartile	Médiane	Moyenne	3 quartile	Max
sans PI	sans PI	00 :00	00 :09	00 :16	00 :37	00 :35	09 :59
sans PI	PI implicite	00 :00	00 :20	00 :39	01 :09	01 :16	10 :22
sans PI	PI explicite	00 :06	00 :19	00 :35	00 :58	01 :08	06 :30
PI implicite	sans PI	00 :02	00 :07	00 :11	00 :22	00 :20	08 :08
PI explicite	sans PI	00 :02	00 :09	00 :15	00 :15	00 :23	02 :34

TABLEAU 7.3 – La synthèse des résultats de la différence de temps entre des énoncés utilisateur selon le type d'énoncé : problématique ou non (minutes : secondes).

Comme nous pouvons l'observer dans le tableau 7.3, le temps de réponse médian lorsque l'on est confronté à un problème d'interaction, est deux fois plus long que le temps de réponse dans une situation normale. Similairement, lorsqu'un problème d'interaction est résolu, le temps moyen de réponse est un peu plus court que dans d'autres combinaisons des énoncés. Nous pouvons supposer que les cas de réponses rapides sont liés à des liens proposés par l'agent et les répétitions utilisateur et que lorsque l'utilisateur a besoin de reformuler sa question, cela prend davantage de temps, mais des analyses supplémentaires sont nécessaires pour le confirmer.

Un graphique par type de PI permet de mieux appréhender la distribution des durées des réponses des utilisateurs. La Figure 7.10 représente la densité des temps de réponse entre les énoncés non-problématiques des utilisateurs. L'ordonnée de la figure représente la durée d'une réponse en secondes. Les lignes droites en bas des figures font partie d'un diagramme en boîte et représentent le premier quartile, la médiane et le 3-ème quartile.

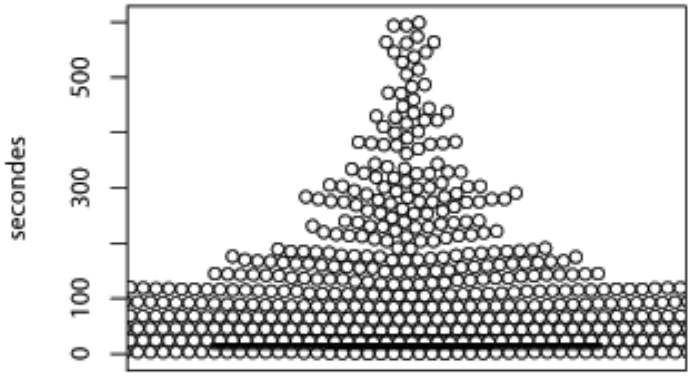


FIGURE 7.10 – Le temps de réponse entre les énoncés non-problématiques.

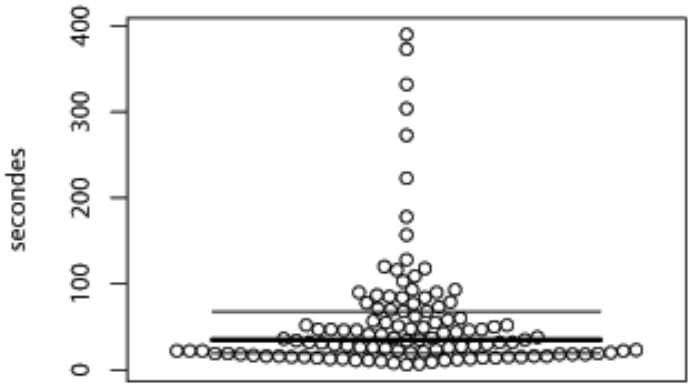


FIGURE 7.11 – Le temps de réponse avant le problème d'interaction explicite.

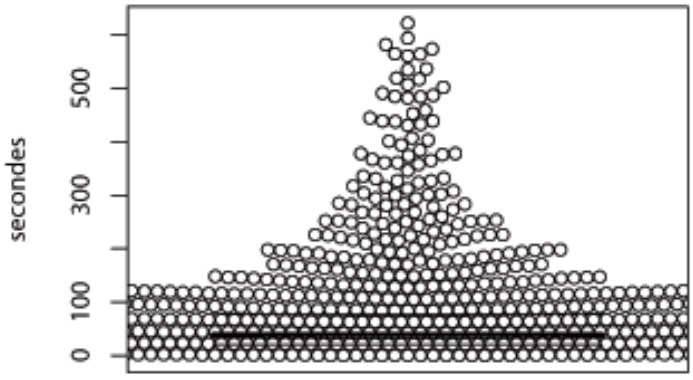


FIGURE 7.12 – Le temps de réponse avant le problème d'interaction implicite.

La Figure 7.11 illustre la densité des temps de réponse entre n'importe quel énoncé utilisateur et un énoncé contenant un PI explicite. La Figure 7.12 représente la densité des temps de réponse lors d'un PI implicite. Un rond symbolise une occurence de temps de réponse. Les deux figures illustrant le temps de réponse avant un PI, ont la même forme. Nous procédons à l'analyse comparative de la figure représentant le temps de réponse entre des énoncés sans PI et la figure représentant le temps de réponse entre un énoncé quelconque d'utilisateur et un énoncé avec un PI implicite.

Les figures 7.10 et 7.12 ont toutes deux une longue traîne de valeurs extrêmes. Toutefois, dans le cas des énoncés non-problématiques, l'étendue des valeurs moyennes est plus resserrée : la majorité des individus ont un temps de réponse entre 9 et 35 secondes. Nous avons alors émis l'hypothèse que le ratio entre la durée d'une réponse et la durée totale du dialogue est propre à chaque utilisateur et permettrait de mieux distinguer les PI des énoncés non-problématiques.

Pour des temps de réponse utilisateur de 5 à 30 secondes et des ratios compris entre 20% et 140%, il est possible de détecter un grand nombre d'énoncés problématiques. Malheureusement, le nombre de faux-positifs est alors pratiquement équivalent au nombre d'annotations correctes. Nous avons également essayé de combiner les seuils de durée avec le seuil de proportion. La seule combinaison qui permettrait d'augmenter un peu le F-score (de + 0,01) est la combinaison 10% + 5sec. Ce seuil permettrait de détecter 100 énoncés problématiques supplémentaires mais multiplierait par 3 le nombre de faux positifs. Le temps de réponse après un PI est trop court pour permettre de créer une règle afin de le différentier des temps de réponses hors PI.

Nous concluons que notre hypothèse ne s'est pas confirmée : le temps de réponse n'est a priori pas un indice exploitable pour la conception de règles de détection de problèmes d'interaction. Néanmoins il peut représenter un indice valide pour un système d'apprentissage automatique.

7.5 Conclusion

Dans ce chapitre nous avons évalué l'efficacité du système DAPI, en fonction des combinaisons de ces composants :

— le correcteur d'orthographe ;

— la mesure de la similarité sémantique.

Les choix effectués pour configurer le système DAPI sont fondés sur les résultats de ces évaluations. Nos tests des composants et des mesures de la similarité textuelle ne se veulent pas exhaustifs et laissent une marge pour le futur travail.

Les résultats obtenus démontrent néanmoins l'intérêt de l'hybridation de l'approche à base de règles linguistiques prenant en compte les spécificités du langage Internet avec l'apprentissage non supervisé de la représentation du sens des mots. Le système DAPI utilise l'association des caractéristiques linguistiques et sémantiques du corpus afin de détecter les problèmes d'interaction entre l'utilisateur et l'agent virtuel sans recours à un grand corpus annoté manuellement. Les évaluations des versions du système DAPI nous amènent vers la conclusion que les défis de l'interaction "in-the-wild" peuvent être résolus d'une façon suffisamment efficace par des solutions qui varient en fonction des besoins. Dans le contexte métier nous privilégions la combinaison du correcteur d'orthographe avec la "compréhension" du contexte par la méthode à base de représentation vectorielle des mots.

Nos résultats montrent également que les approches utilisant l'apprentissage de modèles de représentation vectorielle de mots à base de réseaux neuronaux sont plus efficaces que les approches classiques d'indexation sémantique latente. La mesure cosinus de la similarité sémantique montre également de meilleurs résultats que des mesures plus complexes.

Nous avons également proposé une ouverture en analysant un indice potentiel supplémentaire de détection des problèmes d'interaction.

Quatrième partie

Partie 4 : Conclusion et perspectives

Chapitre 8

Conclusion et perspectives

Sommaire

Cette thèse en Traitement Automatique des Langues (TAL) s'inscrit dans le cadre du développement des méthodes de détection des problèmes d'interaction dans un tchat en temps réel. Les principaux défis à relever lors du traitement des textes du tchat sont le contexte fortement bruité des données "in-the-wild" d'une part et les spécificités du langage utilisateur d'autre part. Nous définissons un problème d'interaction en fonction de l'objectif final qui est l'amélioration de la relation client. Nous orientons donc notre perception d'un problème d'interaction vers la prise en compte de l'opinion de l'utilisateur. Cette prise de vue constitue un défi supplémentaire car le système de conversation écrite ne fournit pas l'information audio que les systèmes vocaux ont sur l'état émotionnel de l'interlocuteur humain.

Dans les sections suivantes, nous dressons un état des lieux de la détection des problèmes d'interaction et décrivons notre apport en proposant des éléments permettant de le valider. Ensuite nous proposons les perspectives à court, moyen et long terme, qui découlent de nos recherches.

8.1 État des lieux de la détection des problèmes d'interaction

Notre problématique englobe deux aspects principaux, souvent étudiés séparément dans la littérature : la détection des problèmes d'interaction et l'analyse des opinions et des phénomènes reliés aux opinions. Le second aspect, à son tour, est étudié d'une manière différente dans la communauté de fouille de texte par rapport à la communauté agent, comme nous l'avons vu dans l'état-de-l'art. De plus, les travaux sur la détection des opinions et des phénomènes reliés aux opinions montrent que les typologies des phénomènes recherchés doivent être adaptées au domaine. Par conséquent, nous proposons une typologie des problèmes d'interaction qui reflète la nature du phénomène du point de vue de l'opinion de l'utilisateur sur l'interaction. Nous distinguons les problèmes d'interaction implicites et explicites.

Nous modélisons l'opinion et les phénomènes reliés aux opinions en combinant les approches existantes en choisissant les éléments les mieux adaptés au texte des dialogues écrits. L'opinion de l'utilisateur est représentée par une relation constituée du triplet : source - opinion - cible. L'opinion négative envers l'interaction représente les problèmes d'interaction explicites. Les problèmes d'interactions implicites se traduisent en pratique surtout par des répétitions et des reformulations de l'utilisateur.

8.2 Approche

Afin de répondre à notre problématique, nous proposons un système hybride DAPI. Son *originalité* est qu'il se compose d'une approche à base de règles linguistiques développées à la main et d'une approche non-supervisée d'apprentissage des représentations sémantiques des mots.

L'approche à base de règles permet d'étudier davantage le phénomène des problèmes d'interaction dans un tchat humain-agent en français. De plus, elle permet de pallier au manque de ressources telles qu'un corpus annoté manuellement de volume suffisant pour l'apprentissage supervisé. Ainsi, les règles annotent les relations en utilisant les spécificités du langage Internet et des dictionnaires d'émotion (LIWC) et de concepts, tel que le concept de l'interaction. Les distances linguistiques sont utilisées pour la détection des

répétitions et des reformulations de l'utilisateur. Elles permettent, en partie, de résoudre le problème des fautes d'orthographe.

⟹ *La détection des problèmes d'interaction s'appuie, entre autres, sur les spécificités du français sur Internet, ne dépendant pas d'un domaine.*

L'approche non-supervisée d'apprentissage des représentations sémantiques des mots permet d'effectuer le calcul de similarité sémantique entre les énoncés utilisateur. La détection des énoncés sémantiquement similaires sert à étoffer la détection des répétitions et reformulations de l'utilisateur. Elle apporte au système l'information sémantique manquante dans l'approche à base de règles.

⟹ *L'approche non-supervisée de l'apprentissage des représentations sémantiques des mots bénéficie de la disponibilité d'un grand corpus non-annoté.*

8.3 Validation

La validation des résultats d'un système demande soit de les comparer avec une référence existante si disponible, soit d'établir une référence, si elle n'existe pas pour la langue, la tâche ou le type de corpus à l'étude, ce qui est notre cas. En premier lieu, nous avons organisé une campagne d'annotation pour obtenir un corpus de référence.

Corpus annoté Nous avons fait annoter le corpus non seulement en problèmes d'interaction mais aussi en opinions et phénomènes reliés aux opinions de l'utilisateur dont la cible est un produit ou un service. Les annotations des opinions sont fines. L'annotateur devait indiquer la polarité positive ou négative de l'opinion, son intensité faible ou élevée, ainsi qu'une des émotions négatives proposées, lorsque cela était possible.

⟹ *Notre corpus est riche en annotations. Les annotations disponibles permettent l'étude des opinions et des phénomènes reliés aux opinions au niveau fin et avec différents types de cible.*

En second lieu, nous avons cherché à établir des résultats de référence pour notre tâche.

Résultats de référence Puisque la nature des problèmes d'interaction est complexe, nous avons établi deux références : une en utilisant la distance de Jaccard, une méthode classique pour la détection des répétitions et la seconde en appliquant une classification Naïve Bayes, une méthode de base pour la classification des opinions et des phénomènes reliés aux opinions. Ces deux approches obtenant de faibles résultats, nous choisissons le module à base de règles de notre système en tant que système de référence. Il possède une bonne précision (72,4%) mais le rappel est plus bas (63,6%).

L'approche hybride montre une bonne amélioration des résultats de détection des problèmes d'interaction par rapport aux résultats de référence. Le test de différents composants du système confirme que le correcteur d'orthographe apporte beaucoup de bruit mais en même temps permet d'augmenter le rappel (77,8%).

⟹ *Le système DAPI montre de bons résultats de détection malgré le contexte complexe de l'analyse automatique du texte, lié aux données récoltées "in-the-wild".*

8.4 Perspectives de recherche

Les applications de tchat avec un agent virtuel sont en plein essor. Il est primordial de continuer l'investigation des moyens pour les évaluer de façon accessible afin de permettre de les faire ensuite évoluer, les rendre plus "humaines". Puisque la majorité des travaux en détection des problèmes d'interaction sont produits pour les systèmes vocaux non-francophones, les travaux de cette thèse ouvrent la voie pour de nombreuses pistes de recherche.

8.4.1 Perspectives à court terme

Passage à l'échelle L'utilisation industrielle du système DAPI demande la réduction du temps de l'analyse des énoncés utilisateur le plus possible. Cela demandera la révision de la chaîne afin d'optimiser les règles et une solution de parallélisation des calculs.

Système hybride Il existe de nombreuses méthodes d'apprentissage non-supervisé des représentations. Celles que nous avons testées ne prennent pas en compte la structure syntaxique de la phrase. Nous aimerions étudier d'autres types d'hybridation entre l'approche à base de règles et l'approche à base d'apprentissage non-supervisé des représentations afin d'améliorer la précision. Il serait intéressant d'effectuer des tests des modèles des représentations des mots, tels que fastText [BOJANOWSKI et collab., 2017] ou MIMICK [PINTER et collab., 2017] permettant à gerer les mots hors le vocabulaire. Nous aimerions également améliorer les hyper-paramètres de l'apprentissage non-supervisé des représentations vectorielles des mots de façon automatique. Nous étudierons en détails l'apport possible de la pénalité affine gap.

Evaluation de la capacité de détection des types d'émotions. Le dictionnaire LIWC contient une classification des émotions correspondant à une partie des types d'émotions que nous avons définis dans la typologie des problèmes d'interaction. Malgré l'absence dans le présent système de règles permettant de gérer la détection des types d'émotions. L'évaluation de la détection à base de dictionnaire est nécessaire pour permettre une analyse plus fine de ses capacités. Puisque le corpus de référence contient déjà les informations sur les types d'émotion, l'évaluation ne demandera pas une constitution de ressources supplémentaires.

Enrichissement de ressources existantes Le lexique d'évaluation Blogoscopie est construit sur la base des blogs. Nous nous attendions à y trouver d'avantage d'expressions d'opinions et de phénomènes reliés aux opinions spécifiques au langage web. Nous avons effectué une étude préliminaire de l'intérêt de l'utilisation du lexique Blogoscopie qui a donné des résultats prometteurs. Nous avons également vu qu'une étude plus approfondie est nécessaire afin d'évaluer le réel intérêt de compléter le lexique du dictionnaire LIWC par le lexique de Blogoscopie. La fusion des deux lexiques devra permettre d'améliorer la détection des problèmes d'interaction explicites.

8.4.2 Perspectives à moyen terme

Extension à un système de détection à grain fin La détection des opinions et des phénomènes reliés aux opinions comprend un niveau plus détaillé d'intensité et des types d'émotions.

L'intensité Le système DAPI détecte l'opinion et les phénomènes reliés aux opinions de polarité négative. La détection plus fine de l'intensité de l'opinion permettra la mise en place des actions en fonction de l'"urgence" d'un problème détecté.

Les types d'émotions L'évaluation de la capacité de la détection des types d'émotions prévue en tant que perspective à court terme, permettra d'établir une stratégie pour l'amélioration de la détection des types d'émotions. Des règles et probablement de nouveaux dictionnaires seront nécessaires. Une détection des types d'émotions permettra de choisir une stratégie d'interaction de l'agent virtuel en fonction de l'émotion détectée.

Test de généralisation du système L'entreprise EDF a récemment constitué d'autres corpus métier de tchat humain-agent. Le système DAPI est prévu pour des tchats métier car il fait usage des termes métier. Il serait également intéressant d'étudier quels composants du système sont généralisables sur les tchats généralistes. Pour pouvoir tester le niveau de généralisation du système DAPI sur d'autres corpus de tchat humain-agent, leur annotation manuelle est nécessaire. C'est la raison pour laquelle nous plaçons ce test important non pas en tant que perspective à court terme mais plutôt à moyen terme.

Un traitement de désambiguïsation des énoncés similaires Lors de l'analyse des faux positifs lors de l'annotation des problèmes d'interaction implicites, nous avons vu que l'un des défis qu'il nous reste à relever est de résoudre le problème de distinction entre une répétition de l'utilisateur et une précision ou un nouveau sous-thème. Ces problématiques sont régulièrement adressées lors des campagnes SemEval et y font l'objet de systèmes dédiés. Par conséquent, ce point demandera une étude détaillée.

8.4.3 Perspectives à long terme

Etude des stratégies de comportement empathique Notre but à long terme de détection des problèmes d'interaction est de pouvoir proposer des stratégies de comportement empathique pour l'agent virtuel en fonction du type de problème d'interaction détecté. De nos jours, la communauté humain-agent dispose des modèles des agents virtuels empathiques. L'étude à long terme consistera à croiser les modèles existants de comportement emphatique et les types des problèmes d'interaction détectés afin de vérifier leur compatibilité. Cette étude demandera des tests en laboratoire et grandeur nature.

Système à base d'apprentissage supervisé Nous avons vu qu'il existe des indices de problèmes d'interaction tels que le temps de réponse qui ne peut pas être utilisé dans le système à base de règles mais en revanche peut servir de descripteur pour un système d'apprentissage supervisé. Il nous parait donc intéressant d'étudier la possibilité d'hybridation du système DAPI avec une approche à base d'apprentissage supervisé ou semi-supervisé. Le système DAPI peut dans ce cas être utilisé pour créer un corpus d'apprentissage.

Proposition d'une métrique de satisfaction de l'utilisateur Avec le système DAPI nous tentons de détecter l'insatisfaction de l'utilisateur par rapport à l'interaction. C'est une analyse non-intrusive par rapport aux questionnaires classiques de satisfaction client. De plus, il permet de détecter d'avantage d'opinions. Les utilisateurs répondent rarement aux questionnaires et comme nous l'avons vu, leurs réponses sont peu corrélées avec la présence des problèmes d'interaction. Les indicateurs automatiques de satisfaction client

existent, tels qu'un clic sur un lien proposé par l'agent virtuel. Il serait intéressant d'étudier la possibilité d'établir plus qu'un indicateur, une métrique automatique de la satisfaction de l'utilisateur.

Liste des figures

Liste des tableaux

Bibliographie

ALLAUZEN, A. et H. BONNEAU-MAYNARD. 2008, «Training and evaluation of pos taggers on the french multitag corpus.», dans *LREC*. 94

ANG, J., R. DHILLON, A. KRUPSKI, E. SHRIBERG et A. STOLCKE. 2002, «Prosody-based automatic detection of annoyance and frustration in human-computer dialog.», dans *INTERSPEECH*. 12, 38

ANGIANI, G., L. FERRARI, T. FONTANINI, P. FORNACCIARI, E. IOTTI, F. MAGLIANI et S. MANICARDI. 2016, «A comparison between preprocessing techniques for sentiment analysis in twitter.», dans *KDWeb*. 44

ANIS, J. 2003, «Communication électronique scripturale et formes langagières», *Actes des Quatrièmes rencontres Réseaux humains/Réseaux technologiques*, vol. 31. 60, 63

APPEL, O., F. CHICLANA, J. CARTER et H. FUJITA. 2016, «A hybrid approach to the sentiment analysis problem at the sentence level», *Knowledge-Based Systems*, vol. 108, p. 110–124. 43

ARTSTEIN, R., S. GANDHE, J. GERTEN, A. LEUSKI et D. TRAUM. 2009, «Semi-formal evaluation of conversational characters», dans *Languages : From Formal to Natural*, Springer, p. 22–35. 84

BACCIANELLA, S., A. ESULI et F. SEBASTIANI. 2010, «Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining.», dans *LREC*, vol. 10, p. 2200–2204. 43, 95

BANSE, R. et K. R. SCHERER. 1996, «Acoustic profiles in vocal emotion expression.», *Journal of personality and social psychology*, vol. 70, nº 3, p. 614. 19, 25

BAO, Y., C. QUAN, L. WANG et F. REN. 2014, «The role of pre-processing in twitter sentiment analysis», dans *International Conference on Intelligent Computing*, Springer, p. 615–624. 44

BARKHUYSEN, P., E. KRAHMER et M. SWERTS. 2005, «Problem detection in human–machine interactions based on facial expressions of users», *Speech communication*, vol. 45, nº 3, p. 343–359. 30

BARRETT, L. F. 2017, «Categories and their role in the science of emotion», *Psychological Inquiry*, vol. 28, nº 1, p. 20–26. 21

BARTNECK, C. 2002, «Integrating the occ model of emotions in embodied characters», dans *Workshop on Virtual Conversational Characters*, p. 39–48. 21

BARTNECK, C., M. J. LYONS et M. SAERBECK. 2017, «The relationship between emotion models and artificial intelligence», *arXiv preprint arXiv :1706.09554*. 21

BASARI, A. S. H., B. HUSSIN, I. G. P. ANANTA et J. ZENIARJA. 2013, «Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization», *Procedia Engineering*, vol. 53, p. 453–462. 42, 43

BEAVER, I. et C. FREEMAN. 2016, «Detection of user escalation in human-computer interactions.», dans *INTERSPEECH*, p. 2075–2079. 12, 31, 44, 98

BECHET, F., G. RICCARDI et D. Z. HAKKANI-TÜR. 2004, «Mining spoken dialogue corpora for system evaluation and modelin.», dans *EMNLP*, p. 134–141. 33

BENAMARA, F., C. GROUIN, J. KAROUI et V. M. I. ROBBA. 2017, «Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du défi fouille de textes deft2017», dans *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 1. 40

BERGSTRA, J., D. YAMINS et D. D. COX. 2013, «Making a science of model search : Hyperparameter optimization in hundreds of dimensions for vision architectures», . 105

BERNSEN, N. O., H. DYBKJAER et L. DYBKJAER. 1996, «Principles for the design of cooperative spoken human-machine dialogue», dans *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, IEEE, p. 729–732. 14

BLEI, D. M., A. Y. NG et M. I. JORDAN. 2003, «Latent dirichlet allocation», *Journal of machine Learning research*, vol. 3, nº Jan, p. 993–1022. 42

BOJANOWSKI, P., E. GRAVE, A. JOULIN et T. MIKOLOV. 2017, «Enriching word vectors with subword information», *Transactions of the Association of Computational Linguistics*, vol. 5, nº 1, p. 135–146. 103, 128

VAN DEN BOSCH, A., E. KRAHMER et M. SWERTS. 2001, «Detecting problematic turns in human-machine interactions : Rule-induction versus memory-based learning approaches», dans *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, p. 82–89. 16, 30, 31, 33, 35

BRADLEY, M. M. et P. J. LANG. 1999, «Affective norms for english words (anew) : Instruction manual and affective ratings», cahier de recherche, Technical report C-1, the center for research in psychophysiology, University of Florida. 34

BRODY, S. et N. ELHADAD. 2010, «An unsupervised aspect-sentiment model for online reviews», dans *Human publisher Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 804–812. 42

BRUN, C., D. N. POPA et C. ROUX. 2015, «Un système hybride pour l'analyse de sentiments associés aux aspects», . 42

BRUNET, E. 2003, «Peut-on mesurer la distance entre deux textes ?», *Corpus*, , nº 2. 88, 99, 107

BUCK, R. 1984, «The communication of emotion.», . 22

BUCK, R. 1999, «The biological affects : a typology.», *Psychological review*, vol. 106, nº 2, p. 301. 19

CAILLIAU, F. et A. CAVET. 2010, «Analyse des sentiments et transcription automatique : modélisation du déroulement de conversations téléphoniques», *Revue Traitement Automatique des Langues*, vol. 51, nº 3, p. 131–154. 3, 34, 38, 82

CAMBRIA, E. et A. HUSSAIN. 2015, *Sentic computing : a common-sense-based framework for concept-level sentiment analysis*, vol. 1, Springer. 43

CAMBRIA, E., B. SCHULLER, Y. XIA et C. HAVASI. 2013, «New avenues in opinion mining and sentiment analysis», *IEEE Intelligent Systems*, vol. 28, nº 2, p. 15–21. 41

CHAI, J. Y., C. ZHANG et T. BALDWIN. 2006, «Towards conversational qa : automatic identification of problematic situations and user intent», dans *Proceedings of the COLING/ACL on Main conference poster sessions*, Association for Computational Linguistics, p. 57–64. 16, 31, 32

CHOWDHURY, S. A., E. A. STEPANOV et G. RICCARDI. 2016, «Predicting user satisfaction from turn-taking in spoken conversations.», dans *INTERSPEECH*, p. 2910–2914. 114

CLAVEL, C., G. ADDA, F. CAILLIAU, M. GARNIER-RIZET, A. CAVET, G. CHAPUIS, S. COURCINOUS, C. DANESI, A.-L. DAQUO, M. DELDOSSI et collab.. 2013, «Spontaneous speech and opinion detection : mining call-centre transcripts», *Language resources and evaluation*, vol. 47, nº 4, p. 1089–1125. 3, 82

CLAVEL, C. et Z. CALLEJAS. 2016, «Sentiment analysis : from opinion mining to human-agent interaction», *IEEE Transactions on Affective Computing*, vol. 7, nº 1, p. 74–93. 17, 38, 68

COHEN, J. 1960, «A coefficient of agreement for nominal scales», *Educational and psychological measurement*, vol. 20, nº 1, p. 37–46. 100

COHEN, W. W. 1995, «Fast effective rule induction», dans *Proceedings of the twelfth international conference on machine learning*, p. 115–123. 35

COHEN, W. W. 1996, «Learning trees and rules with set-valued features», dans *AAAI/IAAI, Vol. 1*, p. 709–716. 31, 35

COLTHEART, M. 1981, «The mrc psycholinguistic database», *The Quarterly Journal of Experimental Psychology*, vol. 33, nº 4, p. 497–505. 33

CONATI, C. et X. ZHOU. 2002, «Modeling students' emotions from cognitive appraisal in educational games», dans *Intelligent Tutoring Systems : 6th International Conference, ITS 2002, Biarritz, France and San Sebastian, Spain, June 2-7, 2002. Proceedings*, Springer, p. 944. 21

COUGNON, L.-A. 2015, *Langage et SMS : Une étude internationale des pratiques actuelles*, vol. 8, Presses universitaires de Louvain. 60

COWIE, R. et R. R. CORNELIUS. 2003, «Describing the emotional states that are expressed in speech», *Speech communication*, vol. 40, nº 1, p. 5–32. 19, 20, 24, 131

CUISINIER, F., C. SANGUIN-BRUCKERT, J.-P. BRUCKERT et C. CLAVEL. 2010, «Les émotions affectent-elles les performances orthographiques en dictée?», *L'Année psychologique*, vol. 110, nº 01, p. 3–48. 120

CUNNINGHAM, H., D. MAYNARD, K. BONTCHEVA, V. TABLAN, N. ASWANI, I. ROBERTS, G. GORRELL, A. FUNK, A. ROBERTS, D. DAMLJANOVIC, T. HEITZ, M. A. GREENWOOD, H. SAGGION, J. PETRAK, Y. LI et W. PETERS. 2011, *Text Processing with GATE (Version 6)*, ISBN 978-0956599315. URL http://tinyurl.com/gatebook. 72, 83, 90

CUNNINGHAM, H., D. MAYNARD et V. TABLAN. 2000, «Jape-a java annotation patterns engine , department of computer science, university of sheffield», . 90, 94

DAMNATI, G., A. GUERRAZ et D. CHARLET. 2015, «Entre écrit et oral? analyse comparée de conversations de type tchat et de conversations téléphoniques dans un centre de contact client», . 54, 55, 131

DAMNATI, G., A. GUERRAZ et D. CHARLET. 2016, «Web chat conversations from contact centers : a descriptive study.», dans *LREC*. 52, 54, 55

DE ANGELI, A. et R. CARPENTER. 2005, «Stupid computer! abuse and social identities», dans *Proc. INTERACT 2005 workshop Abuse : The darker side of Human-Computer Interaction*, p. 19–25. 62

DEERWESTER, S., S. T. DUMAIS, G. W. FURNAS, T. K. LANDAUER et R. HARSHMAN. 1990, «Indexing by latent semantic analysis», *Journal of the American society for information science*, vol. 41, nº 6, p. 391. 103

DENIS, P. et B. SAGOT. 2009, «Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort», dans *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, vol. 1. 94

DERBAIX, C. et M. T. PHAM. 1989, «Pour un développement des mesures de l'affectif en marketing : synthèse des prérequis», *Recherche et applications en marketing*, vol. 4, nº 4, p. 71–87. 22

DING, X. et B. LIU. 2007, «The utility of linguistic rules in opinion mining», dans *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 811–812. 39

DRESNER, E. et S. C. HERRING. 2010, «Functions of the nonverbal in cmc : Emoticons and illocutionary force», *Communication theory*, vol. 20, nº 3, p. 249–268. 60

DUTREY, C. 2011, «Spécificités structurelles et rédactionnelles des corpus issus du web : du text mining au web mining», . 52

DUTREY, C., A. PERADOTTO et C. CLAVEL. 2012, «Analyse de forums de discussion pour la relation clients : du text mining au web content mining», *Actes JADT*. 52, 55, 59, 60, 63

DYBKJÆR, L., N. O. BERNSEN et H. DYBKJÆR. 1996, «Grice incorporated : cooperativity in spoken dialogue», dans *Proceedings of the 16th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, p. 328–333. 14, 15

EFRAIM, O. et F. MOREAU. 2016, «Détecter le besoin d'information dans des requêtes d'usagers d'agents virtuels : sélection de données pertinentes», dans *23ème Conférence sur le Traitement Automatique des Langues Naturelles*. 52, 53

EISENSTEIN, J. 2013, «What to do about bad language on the internet.», dans *HLT-NAACL*, p. 359–369. 66

EKMAN, P. 1999, «Basic emotions», *Handbook of cognition and emotion*, vol. 16, p. 301–320. 19, 25

EKMAN, P., W. V. FRIESEN et P. ELLSWORTH. 1972, *Emotion in the Human Face : Guide-lines for Research and an Integration of Findings : Guidelines for Research and an Integration of Findings*, Pergamon. 18

EL ASRI, L., H. KHOUZAIMI, R. LAROCHE et O. PIETQUIN. 2014, «Ordinal regression for interaction quality prediction», dans *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, p. 3221–3225. 33, 35

ENGELBRECH, K.-P., F. GÖDDE, F. HARTARD, H. KETABDAR et S. MÖLLER. 2009, «Modeling user satisfaction with hidden markov model», dans *Proceedings of the SIGDIAL 2009 conference : the 10th annual meeting of the special interest group on discourse and dialogue*, Association for Computational Linguistics, p. 170–177. 35

FALAISE, A. 2005, «Constitution d'un corpus de français tchaté», dans *RECITAL*, Dourdan, France. 52, 54, 58, 63

FARTOUKH, M., L. CHANQUOY et A. PIOLAT. 2012, «Effects of emotion on writing processes in children», *Written Communication*, vol. 29, n° 4, p. 391–411. 120

FERRARI, S., Y. MATHET, T. CHARNOIS et D. LEGALLOIS. 2008, «Analyse d'opinion : discours évaluatif et classification de documents», *Actes de l'atelier FODOP*, vol. 8, p. 23–36. 62

FORT, K., A. NAZARENKO et S. ROSSET. 2012, «Modeling the complexity of manual annotation tasks : a grid of analysis», dans *International Conference on Computational Linguistics*, p. 895–910. 74

FRIJDA, N. H. 1986, «The emotions : Studies in emotion and social interaction», *Paris : Maison de Sciences de l'Homme*. 21

FUNAKOSHI, K., R. HIGASHINAKA, M. INABA, Y. KOBAYASHI, S. SUGAWARA, K. TAKANASHI, H. OTSUKA, H. KOISO et M. BONO. 2016, «On dialogue breakdown : Annotation and detection», *wochat*. 35

FUNK, A., Y. LI, H. SAGGION, K. BONTCHEVA et C. LEIBOLD. 2008, «Opinion analysis for business intelligence applications», dans *Proceedings of the first international workshop on Ontology-supported business intelligence*, ACM, p. 3. 83

GALLEY, M., K. MCKEOWN, J. HIRSCHBERG et E. SHRIBERG. 2004, «Identifying agreement and disagreement in conversational speech : Use of bayesian networks to model pragmatic dependencies», dans *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, p. 669. 40

GANCHEV, K., J. GILLENWATER, B. TASKAR et collab.. 2010, «Posterior regularization for structured latent variable models», *Journal of Machine Learning Research*, vol. 11, nº Jul, p. 2001–2049. 41

GARDNER, M. P. 1985, «Mood states and consumer behavior : A critical review», *Journal of Consumer research*, vol. 12, nº 3, p. 281–300. 22

GEORGILADAKIS, S., G. ATHANASOPOULOU, R. MEENA, J. LOPES, A. CHORIANOPOULOU, E. PALOGIANNIDI, E. IOSIF, G. SKANTZE et A. POTAMIANOS. 2016, «Root cause analysis of miscommunication hotspots in spoken dialogue systems.», dans *INTERSPEECH*, p. 1156–1160. 12, 30, 31, 32, 33, 34, 35, 38

GHIASSI, M., J. SKINNER et D. ZIMBRA. 2013, «Twitter brand sentiment analysis : A hybrid system using n-gram analysis and dynamic artificial neural network», *Expert Systems with applications*, vol. 40, nº 16, p. 6266–6282. 40

GIANOLA, L. 2014, «Dénition d'un modèle d'émotions pour la catégorisation de textes narratifs courts», *mémoire de Master 2 T̈raitement Automatique des Langues, spécialité Ingénierie Multilingue¨*. 71

GINDL, S., A. WEICHSELBRAUN et A. SCHARL. 2013, «Rule-based opinion target and aspect extraction to acquire affective knowledge», dans *Proceedings of the 22nd International Conference on World Wide Web*, ACM, p. 557–564. 39

GOMAA, W. H. et A. A. FAHMY. 2013, «A survey of text similarity approaches», *International Journal of Computer Applications*, vol. 68, nº 13. 103

GORIN, A. L., G. RICCARDI et J. H. WRIGHT. 1997, «How may i help you ?», *Speech communication*, vol. 23, nº 1-2, p. 113–127. 33

GRATCH, J. et S. MARSELLA. 2005, «Evaluating a computational model of emotion», *Autonomous Agents and Multi-Agent Systems*, vol. 11, nº 1, p. 23–43. 21

GRICE, H. P., P. COLE, J. MORGAN et collab.. 1975, «Logic and conversation», *1975*, p. 41–58. 14

HADDI, E., X. LIU et Y. SHI. 2013, «The role of text pre-processing in sentiment analysis», *Procedia Computer Science*, vol. 17, p. 26–32. 44

HARA, S., N. KITAOKA et K. TAKEDA. 2010, «Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system.», dans *LREC*. 31, 33

HARTIKAINEN, M., E.-P. SALONEN et M. TURUNEN. 2004, «Subjective evaluation of spoken dialogue systems using ser vqual method.», dans *INTERSPEECH*. 30

HASTIE, H. W., R. PRASAD et M. WALKER. 2002, «What's the trouble : automatically identifying problematic dialogues in darpa communicator dialogue systems», dans *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, p. 384–391. 12, 16, 30, 33

HAVILAND, J. M. et M. LEWIS. 1993, *Handbook of emotions*, Guilford Press. 19

HEMMATIAN, F. et M. K. SOHRABI. 2017, «A survey on classification techniques for opinion mining and sentiment analysis», *Artificial Intelligence Review*, p. 1–51. 42

HIGASHINAKA, R., K. FUNAKOSHI, M. ARAKI, H. TSUKAHARA, Y. KOBAYASHI et M. MIZUKAMI. 2015a, «Towards taxonomy of errors in chat-oriented dialogue systems.», dans *SIGDIAL Conference*, p. 87–95. 15

HIGASHINAKA, R., K. FUNAKOSHI, M. INABA, Y. TSUNOMORI, T. TAKAHASHI et N. KAJI. 2017, «Overview of dialogue breakdown detection challenge 3», *Proceedings of Dialog System Technology Challenge*, vol. 6. 34, 36, 37

HIGASHINAKA, R., K. FUNAKOSHI, Y. KOBAYASHI et M. INABA. 2016, «The dialogue breakdown detection challenge : Task description, datasets, and evaluation metrics.», dans *LREC*. 35, 36

HIGASHINAKA, R., K. FUNAKOSHI, M. MIZUKAMI, H. TSUKAHARA, Y. KOBAYASHI et M. ARAKI. 2015b, «Analyzing dialogue breakdowns in chat-oriented dialogue systems», *errare2015*. 38

HIGASHINAKA, R., Y. MINAMI, K. DOHSAKA et T. MEGURO. 2010a, «Issues in predicting user satisfaction transitions in dialogues : Individual differences, evaluation criteria, and prediction models», dans *Spoken Dialogue Systems for Ambient Environments*, Springer, p. 48–60. 33, 35

HIGASHINAKA, R., Y. MINAMI, K. DOHSAKA et T. MEGURO. 2010b, «Modeling user satisfaction transitions in dialogues from overall ratings», dans *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Association for Computational Linguistics, p. 18–27. 35

HIRSCHBERG, J., D. LITMAN et M. SWERTS. 1999, «Prosodic cues to recognition errors», dans *In Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Citeseer. 30

HIRST, G., S. MCROY, P. HEEMAN, P. EDMONDS et D. HORTON. 1994, «Repairing conversational misunderstandings and non-understandings», *Speech communication*, vol. 15, n° 3-4, p. 213–229. 13, 14

HOLBROOK, M. B. 1986, «Emotion in the consumption experience : toward a new model of the human consumer», *The role of affect in consumer behavior : Emerging theories and applications*, vol. 6, n° 23, p. 17–52. 22

HONE, K. S. et R. GRAHAM. 2000, «Towards a tool for the subjective assessment of speech system interfaces (sassi)», *Natural publisher Engineering*, vol. 6, n° 3-4, p. 287–303. 30

HORII, T. et M. ARAKI. 2015, «A breakdown detection method based on taxonomy of errors in chat-oriented dialogue», dans *JSAI Technical Report (SIG-SLUD-75-B502)*, in Japanese, p. 33–36. 35

HORII, T., H. MORI et M. ARAKI. 2017, «Breakdown detector for chat-oriented dialogue», dans *Dialogues with Social Robots*, Springer, p. 119–127. 15

HUMMEL, R. A. et S. W. ZUCKER. 1983, «On the foundations of relaxation labeling processes», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , n° 3, p. 267–287. 42

HUTTO, C. J. et E. GILBERT. 2014, «Vader : A parsimonious rule-based model for sentiment analysis of social media text», dans *Eighth international AAAI conference on weblogs and social media*. 39, 40

IKI, T. et A. SAITO. 2017, «End-to-end character-level dialogue breakdown detection with external memory models», *Proceedings of Dialog System Technology Challenge*, vol. 6. 37

INABA, M. et K. TAKAHASHI. 2015, «Dialogue breakdown detection using long short-term memory recurrent neural network.», dans *JSAI Technical Report (SIG-SLUD-75-B502)*, in Japanese, p. 57–60. 35

ISLAM, A. et D. INKPEN. 2008, «Semantic text similarity using corpus-based word similarity and string similarity», *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, nº 2, p. 10. 107

IZARD, C. E. 1977, «Human emotions.», . 22

IZARD, C. E. et S. BUECHLER. 1979, «Emotions in personality and psychopathology», dans *Emotions in personality and psychopathology*, Springer, p. 445–472. 22

JACCARD, P. 1901, «Distribution de la flore alpine dans le bassin des drouces et dans quelques regions voisines.», vol. 37(140), p. 241–272. 107, 117

JACCARD, P. 1908, «Nouvelles researches sur la distribution florale», *Bull Soc Vaud Sci Nat*, vol. 44, p. 223–270. 88

KATO, S. et T. SAKAI. 2017, «Rsl17bd at dbdc3 : Computing utterance similarities based on term frequency and word embedding vectors», dans *Proceedings of DSTC6. http ://workshop. colips. org/dstc6/papers/track3_paper13_kato. pdf*. 34, 37, 44

KEMPER, T. D. 1987, «How many emotions are there? wedding the social and the autonomic components», *American journal of Sociology*, vol. 93, nº 2, p. 263–289. 22

KENNEDY, J. et R. EBERHART. 1995, «Particle swarm optimization», dans *Proceedings of IEEE International Conference on Neural Networks IV*, vol. 1000. 43

KIEU, B. T. et S. B. PHAM. 2010, «Sentiment analysis for vietnamese», dans *Knowledge and Systems Engineering (KSE), 2010 Second International Conference on*, IEEE, p. 152–157. 83

KIM, S.-M. et E. HOVY. 2004, «Determining the sentiment of opinions», dans *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, p. 1367. 38, 41

KIM, W. 2007, «Online call quality monitoring for automating agent-based call centers», dans *Eighth Annual Conference of the International Speech Communication Association*. 33

KINGMA, D. P. et J. BA. 2014, «Adam : A method for stochastic optimization», *ICLR 2015*. 105

KOBAYASHI, S., Y. UNNO et M. FUKUDA. 2015, «Multi-task learning of recurrent neural network for detecting breakdowns of dialog and publisher modeling», dans *JSAI Technical Report (SIG-SLUD-75-B502)*, in Japanese, p. 41–46. 35

KRAHMER, E., M. SWERTS, M. THEUNE et M. WEEGELS. 1999, «Problem spotting in human-machine interaction», . 30, 31

KUMAR, V. R. et K. RAGHUVEER. 2013, «Dependency driven semantic approach to product features extraction and summarization using customer reviews», dans *Advances in computing and information technology*, Springer, p. 225–238. 39

KUSNER, M., Y. SUN, N. KOLKIN et K. WEINBERGER. 2015, «From word embeddings to document distances», dans *International Conference on Machine Learning*, p. 957–966. 34

KUZNICK, L., A.-L. GUÉNET, A. PERADOTTO et C. CLAVEL. 2010, «L'apport des concepts métiers pour la classification des questions ouvertes d'enquête», *les actes de TALN*. 3

LAGARDE, D. et A. PERADOTTO. 2013, «Guide d'annotation concepts metier edf», *documentation interne EDF R&D*. 71

LANDRAGIN, F. et L. ROMARY. 2004, «Dialogue history modelling for multimodal human-computer interaction», dans *Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)*, Universitat Pompeu Fabra, p. 41–48. 33

LANGKILDE, I., M. WALKER, J. WRIGHT, A. GORIN et D. LITMAN. 1999, «Automatic prediction of problematic human-computer dialogues in how may i help you», dans *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRUU99*, p. 1–4. 12, 13, 30, 35

LANGLET, C. et C. CLAVEL. 2015, «Improving social relationships in face-to-face human-agent interactions : when the agent wants to know user's likes and dislikes», dans *Proceedings of annual meeting of the association for computational linguistics, ACL 2015*. 71, 82, 94, 96

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, p. 116–124. 39

LARIVEY, M. 2002, «La puissance des émotions : Comment distinguer les vraies des fausses., de l'homme ed», *Québec : Les éditions de l'Homme*. 19, 20, 24, 25

LAROS, F. J. et J.-B. E. STEENKAMP. 2005, «Emotions in consumer behavior : a hierarchical approach», *Journal of business Research*, vol. 58, nº 10, p. 1437–1445. 22, 23

LAU, J. H. et T. BALDWIN. 2016, «An empirical evaluation of doc2vec with practical insights into document embedding generation», dans *Proceedings of the 1st Workshop on Representation Learning for NLP*, p. 78–86. 105

LAVALLEY, R., C. CLAVEL et P. BELLOT. 2010, «Extraction probabiliste de chaînes de mots relatives à une opinion», *Traitement Automatique des Langues*, vol. 51, p. 101–130. 3

LAZARUS, R. S. 1966, «Psychological stress and the coping process.», . 21

LAZARUS, R. S. 1999, «Stress and emotion : A new synthesis.», . 19

LE, Q. et T. MIKOLOV. 2014, «Distributed representations of sentences and documents», dans *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, p. 1188–1196. 103, 104, 105

LEHRER, A. 1974, «Semantic fields and lexical structure.», *North Holland, Amsterdam and New York*. 39

LENDREVIE, J. et J. LÉVY. 2014, *Mercator 11e édition : Tout le marketing à l'ère numérique*, Dunod, 527 p.. 24

LENDVAI, P., A. VAN DEN BOSCH, E. KRAHMER et M. SWERTS. 2002a, «Multi-feature error detection in spoken dialogue systems», *publisher and Computers*, vol. 45, nº 1, p. 163–178. 31, 35

LENDVAI, P., A. VAN DEN BOSCH, E. KRAHMER et M. SWERTS. 2002b, «Improving machine-learned detection of miscommunications in human-machine dialogues through informed data splitting», dans *Proceedings of the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics*, p. 1–15. 33

LEVENSHTEIN, V. I. 1966, «Binary codes capable of correcting deletions, insertions, and reversals», dans *Soviet physics doklady*, vol. 10, p. 707–710. 88, 99

LEWIS, M., J. M. HAVILAND-JONES et L. F. BARRETT. 2010, *Handbook of emotions*, Guilford Press. 25

LIN, D. et collab.. 1998, «An information-theoretic definition of similarity.», dans *Icml*, vol. 98, p. 296–304. 32

LISCOMBE, J., G. RICCARDI et D. Z. HAKKANI-TÜR. 2005, «Using context to improve emotion detection in spoken dialog systems.», dans *Interspeech*, p. 1845–1848. 12, 32, 33, 34, 38

LIU, B. 2010, «Sentiment analysis and subjectivity.», *Handbook of natural publisher processing*, vol. 2, p. 627–666. 39

LIU, B. 2012, «Sentiment analysis and opinion mining», *Synthesis lectures on human publisher technologies*, vol. 5, nº 1, p. 1–167. 38

LIU, Z. et Z. G. PAN. 2005, «An emotion model of 3d virtual characters in intelligent virtual environment», dans *International Conference on Affective Computing and Intelligent Interaction*, Springer, p. 629–636. 21

LOPES, J. 2017, «How generic can dialogue breakdown detection be? the kth entry to dbdc3», *Proceedings of Dialog System Technology Challenge*, vol. 6. 33, 34, 37

LORENZ, P. et N. MICHOT. 2012, «Le lexique du chat sur internet : étude comparative français-espagnol-polonais», dans *SHS Web of Conferences*, vol. 1, EDP Sciences, p. 939–954. 60

LORIA, S., P. KEEN, M. HONNIBAL, R. YANKOVSKY, D. KARESH, E. DEMPSEY et collab.. 2014, «Textblob : simplified text processing», . 44

LU, J., C. LIN, W. WANG, C. LI et H. WANG. 2013, «String similarity measures and joins with synonyms», dans *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, ACM, p. 373–384. 106

LUQUE, J., C. SEGURA, A. SÁNCHEZ, M. UMBERT et L. A. GALINDO. 2017, «The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls», *Proc. Interspeech 2017*, p. 2346–2350. 31

LYNN, T., J. FOSTER, M. DRAS et L. TOUNSI. 2014, «Cross-lingual transfer parsing for low-resourced languages : An irish case study», dans *Proceedings of the First Celtic Language Technology Workshop*, p. 41–49. 64

MACHLEIT, K. A. et S. A. EROGLU. 2000, «Describing and measuring emotional response to shopping experience», *Journal of Business Research*, vol. 49, nº 2, p. 101–111. 22

MAHARANI, W., D. H. WIDYANTORO et M. L. KHODRA. 2015, «Aspect extraction in customer reviews using syntactic pattern», *Procedia Computer Science*, vol. 59, p. 244–253. 39

MAHRER, R. 2017, *Phonographie : La représentation écrite de l'oral en français*, vol. 3, Walter de Gruyter GmbH & Co KG. 63

MARCOCCIA, M. 2000, «La représentation du nonverbal dans la communication écrite médiatisée par ordinateur», *Communication et organisation*, , nº 18. 52

MARCOCCIA, M. et N. GAUDUCHEAU. 2007, «L'analyse du rôle des smileys en production et en réception : un retour sur la question de l'oralité des écrits numériques», *Glottopol*, vol. 10, p. 38–55. 60, 62

MARTIN, J. R. et P. R. WHITE. 2005, «The language of evaluation», *Appraisal in English. Basingstoke & New York : Pal grave Macmillan*. 18, 21, 23, 24, 26, 45, 60, 68, 82, 91, 131

MARTÍN-VALDIVIA, M.-T., E. MARTÍNEZ-CÁMARA, J.-M. PEREA-ORTEGA et L. A. UREÑA-LÓPEZ. 2013, «Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches», *Expert Systems with Applications*, vol. 40, nº 10, p. 3934–3942. 43

MARTINOVSKY, B. et D. TRAUM. 2006, «The error is the clue : Breakdown in human-machine interaction», cahier de recherche, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES. 12

MASLOWSKI, I. 2016, «Quelles sont les caractéristiques des interactions problématiques entre des utilisateurs et un conseiller virtuel?», *PARIS Inalco du 4 au 8 juillet 2016*, p. 94. 57

MAYNARD, D., J. PETRAK, A. FUNK et L. DERCZYNSKI. December 2016, «D3.1 multilingual content processing methods», *deliverable*. 92

MCGILL, M. 1979, «An evaluation of factors affecting document ranking by information retrieval systems.», . 87

MEENA, R., J. LOPES, G. SKANTZE et J. GUSTAFSON. 2015, «Automatic detection of miscommunication in spoken dialogue systems.», dans *SIGDIAL Conference*, p. 354–363. 16, 31, 32, 35, 88

MIKOLOV, T., K. CHEN, G. CORRADO et J. DEAN. 2013a, «Efficient estimation of word representations in vector space», *arXiv preprint arXiv :1301.3781*. 36, 105, 109

MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. S. CORRADO et J. DEAN. 2013b, «Distributed representations of words and phrases and their compositionality», dans *Advances in neural information processing systems*, p. 3111–3119. 36

MILLER, G. 1998, *WordNet : An electronic lexical database*, MIT press. 41

MIZUKAMI, M., K. SUGIYAMA, G. NEUBIG, K. YOSHINO, S. SAKTI et S. NAKAMURA. 2015, «Construction of rnn-based dialogue breakdown detector.», dans *JSAI Technical Report (SIG-SLUD-75-B502)*, in Japanese, p. 47–50. 35

MOILANEN, K. et S. PULMAN. 2007, «Sentiment composition», dans *Proceedings of RANLP*, vol. 7, p. 378–382. 39, 40

MÖLLER, S., K.-P. ENGELBRECHT et A. OULASVIRTA. 2007, «Analysis of communication failures for spoken dialogue systems», dans *Eighth Annual Conference of the International Speech Communication Association*. 12, 13, 14

MORAES, R., J. F. VALIATI et W. P. G. NETO. 2013, «Document-level sentiment classification : An empirical comparison between svm and ann», *Expert Systems with Applications*, vol. 40, nº 2, p. 621–633. 38

MOREAU, E. 2001, «Le «chat» et les relations humaines», *Anthropologie de la société digitale*, vol. 1, p. 117. 63

MUKHERJEE, A. et B. LIU. 2012, «Aspect extraction through semi-supervised modeling», dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, Association for Computational Linguistics, p. 339–348. 41

MULHOLLAND, E., P. MC KEVITT, T. LUNNEY et K.-M. SCHNEIDER. 2016, «Analysing emotional sentiment in people's youtube channel comments», dans *International Conference on ArtsIT, Interactivity & Game Creation*, Springer, p. 181–188. 83

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. 17, 18, 23

NASR, A., G. DAMNATI, A. GUERRAZ et F. BECHET. 2016, «Syntactic parsing of chat language in contact center conversation corpus», dans *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 175. 63

NIGAM, K., A. K. MCCALLUM, S. THRUN et T. MITCHELL. 2000, «Text classification from labeled and unlabeled documents using em», *Machine learning*, vol. 39, nº 2-3, p. 103–134. 41

ORTONY, A., G. L. CLORE et A. COLLINS. 1990, *The cognitive structure of emotions*, Cambridge university press. 20

OSGOOD, C. E., W. H. MAY et M. S. MIRON. 1975, *Cross-cultural universals of affective meaning*, University of Illinois Press. 18

OSHERENKO, A. et E. ANDRÉ. 2009, «Differentiated semantic analysis in lexical affect sensing», dans *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, Ieee, p. 1–6. 39

PALOGIANNIDI, E., E. LOSIF, P. KOUTSAKIS et A. POTAMIANOS. 2015, «Valence, arousal and dominance estimation for english, german, greek, portuguese and spanish lexica using semantic models», . 34, 38

PANCKHURST, R. 2006, «Le discours électronique médié : bilan et perspectives.», . 60

PANG, B. et L. LEE. 2008, «Opinion mining and sentiment analysis», *Foundations and trends in information retrieval*, vol. 2, nº 1-2, p. 1–135. 23

PARK, C., K. KIM et S. KIM. 2017, «Attention-based dialog embedding for dialog breakdown detection», *Proceedings of Dialog System Technology Challenge*, vol. 6. 37

PAROUBEK, P. 2016, «Critères pour l'annotation active de microblogs», *PARIS Inalco du 4 au 8 juillet 2016*, p. 7. 72

PAROUBEK, P., J.-B. BERTHELIN, S. EL AYARI, C. GROUIN, T. HEITZ, M. HURAULT-PLANTET, M. JARDINO, Z. KHALIS et M. LASTES. 2007, «Résultats de l'édition 2007 du défi fouille de textes», *Actes de l'atelier de clôture du 3eme DEfi Fouille de Textes*, p. 9–17. 40

PENNEBAKER, J., C. CHUNG, M. IRELAND, A. GONZALES et R. BOOTH. 2007, «The development and psychometric properties of liwc2007 : Liwc. net», . 62

PENNEBAKER, J. W., M. E. FRANCIS et R. J. BOOTH. 2001, «Linguistic inquiry and word count : Liwc 2001», *Mahway : Lawrence Erlbaum Associates*, vol. 71, nº 2001, p. 2001. 39

PENNEBAKER, J. W. et A. GRAYBEAL. 2001, «Patterns of natural publisher use : Disclosure, personality, and social integration», *Current Directions in Psychological Science*, vol. 10, nº 3, p. 90–93. 39

PIETERS, R. G. et W. F. VAN RAAIJ. 1988, «Functions and management of affect : Applications to economic behavior», *Journal of Economic Psychology*, vol. 9, nº 2, p. 251–282. 22

PINTER, Y., R. GUTHRIE et J. EISENSTEIN. 2017, «Mimicking word embeddings using subword rnns», dans *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 102–112. 103, 128

PIOLAT, A., R. BOOTH, C. CHUNG, M. DAVIDS et J. PENNEBAKER. 2011a, «The french dictionary for liwc : Modalities of construction and examples of use| la version franaise du dictionnaire pour le liwc : modalités de construction et exemples d'utilisation», . 91

PIOLAT, A., R. J. BOOTH, C. K. CHUNG, M. DAVIDS et J. W. PENNEBAKER. 2011b, «La version française du dictionnaire pour le liwc : modalités de construction et exemples d'utilisation», *Psychologie française*, vol. 56, nº 3, p. 145–159. 62, 95

PLUTCHIK, R. 1980, *Emotion : A psychoevolutionary synthesis*, Harpercollins College Division. 18, 19, 22, 25, 131

POPESCU, A.-M. et O. ETZIONI. 2007, «Extracting product features and opinions from reviews», dans *Natural publisher processing and text mining*, Springer, p. 9–28. 42

PORIA, S., A. GELBUKH, A. HUSSAIN, N. HOWARD, D. DAS et S. BANDYOPADHYAY. 2013, «Enhanced senticnet with affective labels for concept-based opinion mining», *IEEE Intelligent Systems*, vol. 28, nº 2, p. 31–38. 41

PRAGST, L., S. ULTES et W. MINKER. 2017, «Recurrent neural network interaction quality estimation», dans *Dialogues with Social Robots*, Springer, p. 381–393. 35

RAMÍREZ-TINOCO, F. J., G. ALOR-HERNÁNDEZ, J. L. SÁNCHEZ-CERVANTES, B. A. OLIVARES-ZEPAHUA et L. RODRÍGUEZ-MAZAHUA. 2017, «A brief review on the use of sentiment analysis approaches in social networks», dans *International Conference on Software Process Improvement*, Springer, p. 263–273. 40

RAMOS, F. Y. 2005, «Attitudes and emotions through written text : the case of textual deformation in internet chat rooms.», *PragmalingÃ1/4Ãstica*, , nº 13, p. 147–173. 63, 66

RATCLIFF, J. W. et D. E. METZENER. 1988, «Pattern-matching-the gestalt approach», *Dr Dobbs Journal*, vol. 13, nº 7, p. 46. 93

RICHINS, M. L. 1997, «Measuring emotions in the consumption experience», *Journal of consumer research*, vol. 24, nº 2, p. 127–146. 23

ROUILLARD, J. et J. CAELEN. 1998, «Etude du dialogue hommemachine en langue naturelle sur le web pour une recherche documentaire», dans *Deuxième colloque international sur l'apprentissage personne-système, Caps*, vol. 98. 52, 53

ROUVIER, M. et P.-M. BOUSQUET. 2017, «Lia@ deft'2017 : Multi-view ensemble of convolutional neural network», dans *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 13. 40

ROY, S., R. MARIAPPAN, S. DANDAPAT, S. SRIVASTAVA, S. GALHOTRA et B. PEDDAMUTHU. 2016, «Qart : A system for real-time holistic quality assurance for contact center dialogues.», dans *AAAI*, p. 3768–3775. 34, 38

RUSSELL, J. 1980, «A circumplex model of affect.», *Journal of Personality and Social Psychology.*, vol. 39, p. 1161–1178. 18, 19, 25, 131

SAAD, F. 2014, «Baseline evaluation : an empirical study of the performance of machine learning algorithms in short snippet sentiment analysis», dans *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, ACM, p. 6. 117

SADOUN, D. 2016, «Création semi-automatique d'un corpus annoté pour l'analyse d'opinions», dans *SHS Web of Conferences*, vol. 27, EDP Sciences. 74

SALTON, G. 1971, «The smart retrieval system—experiments in automatic document processing», . 31

SAMHA, A. K. 2016, «Aspect-based opinion mining using dependency relations», vol. 4, p. 113–123. 39

SCHAPIRE, R. E. et Y. SINGER. 2000, «Boostexter : A boosting-based system for text categorization», *Machine learning*, vol. 39, nº 2-3, p. 135–168. 38, 41

SCHERER, K. R. 2005, «What are emotions? and how can they be measured?», *Social science information*, vol. 44, nº 4, p. 695–729. 18

SCHERER, K. R., A. SCHORR et T. JOHNSTONE. 2001, *Appraisal processes in emotion : Theory, methods, research*, Oxford University Press. 21

SCHMID, H. 1994, «Probabilistic part-of-speech tagging using decision trees», dans *Proceedings of International Conference on New Methods in Language Processing*, Association for Computational Linguistics, Manchester, UK. 58, 59, 63, 94

SCHMID, H. 1995, «Treetagger| a language independent part-of-speech tagger», *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, vol. 43, p. 28. 83

SCHMITT, A., B. SCHATZ et W. MINKER. 2011, «Modeling and predicting quality in spoken human-computer interaction», dans *Proceedings of the SIGDIAL 2011 Conference*, Association for Computational Linguistics, p. 173–184. 16, 33, 34, 35

SCHMITT, A., S. ULTES et W. MINKER. 2012, «A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system.», dans *LREC*, p. 3369–3373. 117

SCHULLER, B., J.-G. GANASCIA et L. DEVILLERS. 2016, «Multimodal sentiment analysis in the wild : Ethical considerations on data collection, annotation, and exploitation», dans *Actes du Workshop on Ethics In Corpus Collection, Annotation & Application (ETHICA2), LREC, Portoroz, Slovénie*. 4

SEGURA, C., D. BALCELLS, M. UMBERT, J. ARIAS et J. LUQUE. 2016, «Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls», dans *Advances in Speech and publisher Technologies for Iberian publishers : Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3*, Springer, p. 255–265. 38

SHAIKH, M. A. M., H. PRENDINGER et M. ISHIZUKA. 2009, «A linguistic interpretation of the occ emotion model for affect sensing from text», *Affective information processing*, p. 45–73. 21

SHRIBERG, E., E. WADE et P. PRICE. 1992, «Human-machine problem solving using spoken language systems (sls) : Factors affecting performance and user satisfaction», dans *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, p. 49–54. 60

SILVERVARG, A. et A. JÖNSSON. 2011, «Subjective and objective evaluation of conversational agents in learning environments for young teenagers», dans *Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. 84

SINGH, T. et M. KUMARI. 2016, «Role of text pre-processing in twitter sentiment analysis», *Procedia Computer Science*, vol. 89, p. 549–554. 44

SMITH, T. F. et M. S. WATERMAN. 1981, «Comparison of biosequences», *Advances in applied mathematics*, vol. 2, n° 4, p. 482–489. 99

SOCHER, R., E. H. HUANG, J. PENNIN, C. D. MANNING et A. Y. NG. 2011, «Dynamic pooling and unfolding recursive autoencoders for paraphrase detection», dans *Advances in Neural Information Processing Systems*, p. 801–809. 33

SOCHER, R., A. PERELYGIN, J. WU, J. CHUANG, C. D. MANNING, A. NG et C. POTTS. 2013, «Recursive deep models for semantic compositionality over a sentiment treebank», dans *Proceedings of the 2013 conference on empirical methods in natural publisher processing*, p. 1631–1642. 41

SONG, Y., L. MOU, R. YAN, L. YI, Z. ZHU, X. HU et M. ZHANG. 2016, «Dialogue session segmentation by embedding-enhanced texttiling», *Interspeech 2016*, p. 2706–2710. 103, 106

SUGIYAMA, H. 2015, «Chat-oriented dialogue breakdown detection based on combination of various data.», dans *JSAI Technical Report (SIG-SLUD-75-B502)*, in Japanese, p. 51–56. 35

SUGIYAMA, H. 2017, «Dialogue breakdown detection based on estimating appropriateness of topic transition», *Dialog System Technology Challenges (DSTC6)*. 33, 34, 37

SUIGNARD, P. 2010, «Naviquest : un outil pour naviguer dans une base de questions posées à un agent conversationnel», dans *Workshop sur les agents conversationnels Animés, Lille*. 3, 82

SUIGNARD, P., F. CAILLIAU et A. CAVET. 2012, «La longueur des tours de parole comme critère de sélection de conversations dans un centre d'appels (turn-taking length as criterion to select call center conversations)[in french]», dans *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 551–558. 30

SUN, S., C. LUO et J. CHEN. 2017, «A review of natural language processing techniques for opinion mining systems», *Information Fusion*, vol. 36, p. 10–25. 41

TABOADA, M., J. BROOKE, M. TOFILOSKI, K. VOLL et M. STEDE. 2011, «Lexicon-based methods for sentiment analysis», *Computational linguistics*, vol. 37, n° 2, p. 267–307. 39

TAKAYAMA, J., E. NOMOTO et Y. ARASE. 2017, «Dialogue breakdown detection considering annotation biases», *Diaog System Technology Challenges*, vol. 6. 37

TANG, H., C. B. P. LEE et K. K. CHOONG. 2017, «Consumer decision support systems for novice buyers–a design science approach», *Information Systems Frontiers*, vol. 19, n° 4, p. 881–897. 83

TANIGUCHI, R. et Y. KANO. 2015, «Construction of automatic detector for dialogue breakdowns based on rules with keywords extraction», dans *JSAI Technical Report (SIG-SLUD-75-B502)*, in Japanese, p. 37–40. 35

THOUËSNY, S. 2009, «Increasing the reliability of a part-of-speech tagging tool for use with learner language», dans *Presentation given at the Automatic Analysis of Learner Language (AALL'09) workshop on automatic analysis of learner language : from a better understanding of annotation needs to the development and standardization of annotation schemes*. 64

TOUTANOVA, K., D. KLEIN, C. D. MANNING et Y. SINGER. 2003, «Feature-rich part-of-speech tagging with a cyclic dependency network», dans *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, p. 173–180. 84, 94

TROCHE, J., S. J. CRUTCH et J. REILLY. 2017, «Defining a conceptual topography of word concreteness : Clustering properties of emotion, sensation, and magnitude among 750 english words», *Frontiers in psychology*, vol. 8, p. 1787. 33

TSAGKALIDOU, K., V. KOUTSONIKOLA, A. VAKALI et K. KAFETSIOS. 2011, «Emotional aware clustering on micro-blogging sources», *Affective Computing and Intelligent Interaction*, p. 387–396. 42

TURK, M. 1996, «Visual interaction with lifelike characters», dans *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, IEEE, p. 368–373. 30

TURNEY, P. D. 2002, «Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews», dans *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 417–424. 42

VAN RIJSBERGEN, C. 1979, «Information retrieval. dept. of computer science, university of glasgow», *URL : citeseer. ist. psu. edu/vanrijsbergen79information. html*, vol. 14. 118

VERNIER, M. et L. MONCEAUX. 2010, «Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques», *Traitement automatique des langues*, vol. 51, nº 1, p. 125–149. 34

VERNIER, M., L. MONCEAUX et B. DAILLE. 2010, «Learning subjectivity phrases missing from resources through a large set of semantic tests», dans *The 7th International Conference on Language Resources and Evaluation (LREC'10)*, p. 1335–1341. 95

VINYALS, O. et Q. LE. 2015, «A neural conversational model», *Proc. ICML Deep Learning Workshop*. 36

WALKER, M. A., I. LANGKILDE-GEARY, H. WRIGHT HASTIE, J. WRIGHT et A. GORIN. 2002, «Automatically training a problematic dialogue predictor for a spoken dialogue system», *Journal of Artificial Intelligence Research*, vol. 16, p. 293–319. 2, 12, 13, 16, 30, 35, 82

WALKER, M. A., D. J. LITMAN, C. A. KAMM et A. ABELLA. 1997, «Paradise : A framework for evaluating spoken dialogue agents», dans *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 271–280. 84

WALLOT, S. et J. GRABOWSKI. 2013, «Typewriting dynamics : What distinguishes simple from complex writing tasks ?», *Ecological Psychology*, vol. 25, nº 3, p. 267–280. 120

WATSON, D., L. A. CLARK et A. TELLEGEN. 1988, «Development and validation of brief measures of positive and negative affect : the panas scales.», *Journal of personality and social psychology*, vol. 54, nº 6, p. 1063. 62

WENGELIN, Å., M. TORRANCE, K. HOLMQVIST, S. SIMPSON, D. GALBRAITH, V. JOHANSSON et R. JOHANSSON. 2009, «Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production», *Behavior research methods*, vol. 41, nº 2, p. 337–351. 120

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, p. 165–210. 21, 26

WIERZBICKA, A. 1999a, «Emotional universals», *Language Design : Journal of Theoretical and Experimental Linguistics*, , nº 2, p. 23–69. 20, 25

WIERZBICKA, A. 1999b, *Emotions across languages and cultures : Diversity and universals*, Cambridge University Press. 17

WILSON, T., J. WIEBE et P. HOFFMANN. 2005, «Recognizing contextual polarity in phrase-level sentiment analysis», dans *Proceedings of the conference on human publisher technology and empirical methods in natural publisher processing*, Association for Computational Linguistics, p. 347–354. 40

WILSON, T., J. WIEBE et P. HOFFMANN. 2009, «Recognizing contextual polarity : An exploration of features for phrase-level sentiment analysis», *Computational linguistics*, vol. 35, nº 3, p. 399–433. 41

XIANG, Y., Y. ZHANG, X. ZHOU, X. WANG et Y. QIN. 2014, «Problematic situation analysis and automatic recognition for chi-nese online conversational system», *Proc. CLP*, p. 43–51. 13, 15, 16, 31, 32, 34, 35, 38, 44, 45

XIE, Z. et G. LING. 2017, «Dialogue breakdown detection using hierarchical bi-directional lstms», *Proceedings of Dialog System Technology Challenge*, vol. 6. 37

YANG, B. et C. CARDIE. 2014, «Context-aware learning for sentence-level sentiment analysis with posterior regularization.», dans *ACL (1)*, p. 325–335. 41

YATES, J., W. J. ORLIKOWSKI et collab.. 1993, «Knee-jerk anti-loopism and other e-mail phenomena : Oral, written, and electronic patterns in computer-mediated communication», . 60

YESSENALINA, A., Y. YUE et C. CARDIE. 2010, «Multi-level structured models for document-level sentiment classification», dans *Proceedings of the 2010 Conference on Empirical Methods in Natural publisher Processing*, Association for Computational Linguistics, p. 1046–1056. 38

ZHAI, Z., B. LIU, H. XU et P. JIA. 2011, «Clustering product features for opinion mining», dans *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, p. 347–354. 41

ZHANG, L. 2012, *Analyse automatique d'opinion : problématique de l'intensité et de la négation pour l'application à un corpus journalistique*, thèse de doctorat, Université de Caen. 39

ZHANG, L., B. LIU, S. H. LIM et E. O'BRIEN-STRAIN. 2010, «Extracting and ranking product features in opinion documents», dans *Proceedings of the 23rd international conference on computational linguistics : Posters*, Association for Computational Linguistics, p. 1462–1470. 39

De l'analyse d'opinions à la détection des problèmes d'interactions humain-machine : application à la gestion de la relation client

Irina POLTAVCHENKO
ÉPOUSE MASLOWSKI

RESUMÉ :

Motivée par le gain en popularité des chatbots prenant le rôle de conseillers sur les sites Web des entreprises, cette thèse s'attaque au problème de la détection des problèmes d'interaction entre un conseiller virtuel et ses utilisateurs sous l'angle de l'analyse des opinions et des émotions dans les textes. Cette thèse s'est déroulée dans le cadre d'une application concrète pour l'entreprise EDF et s'est appuyée sur le corpus du chatbot d'EDF. Ce corpus regroupe des expressions spontanées et riches, collectées dans les conditions écologiques (parfois appelées « in-the-wild »), difficiles à analyser de façon automatique, et encore peu étudiées.

Nous proposons une typologie des problèmes d'interaction et faisons annoter une partie du corpus selon cette typologie, annotation dont une partie servira à l'évaluation du système. Le système de Détection Automatique des Problèmes d'Interaction (DAPI) développé lors de cette thèse est un système hybride qui allie l'approche symbolique et l'apprentissage non supervisé de représentation sémantique par plongements lexicaux (word embeddings). Le système DAPI a pour vocation d'être directement connecté au chatbot et de détecter des problèmes d'interaction en ligne, dès la réception d'un énoncé utilisateur. L'originalité de la méthode proposée repose sur : i) la prise en compte de l'historique du dialogue ; ii) la modélisation des problèmes d'interaction en tant qu'expressions des opinions et des phénomènes reliés aux opinions spontanées de l'utilisateur vis-à-vis de l'interaction ; iii) l'intégration des spécificités du langage web et « in-the-wild » comme des indices linguistiques pour les règles linguistiques ; iv) recours aux plongements lexicaux de mots (word2vec) appris sur le grand corpus du chatbot non étiqueté afin de modéliser des similarités sémantiques. Les résultats obtenus sont très encourageants compte tenu de la complexité des données : F-score = 74,3%.

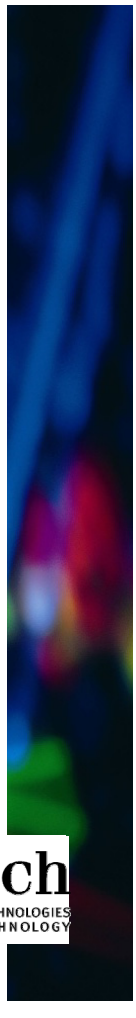
MOTS-CLEFS : chatbot, problèmes d'interactions, analyse d'opinion, interaction humain - machine, interactions écrites, dialogue

ABSTRACT :

This PHD thesis is motivated by the growing popularity of chatbots acting as advisors on corporate websites. This research addresses the detection of the interaction problems between a virtual advisor and its users from the angle of opinion and emotion analysis in the texts. The present study takes place in the concrete application context of a French energy supplier EDF, using EDF chatbot corpus. This corpus gathers spontaneous and rich expressions, collected in "in-the-wild" conditions, difficult to analyze automatically, and still little studied.

We propose a typology of interaction problems and annotate a part of the corpus according to this typology. A part of created annotation is used to evaluate the system. The system named DAPI (automatic detection of interaction problems) developed during this thesis is a hybrid system that combines the symbolic approach and the unsupervised learning of semantic representation (word embeddings). The purpose of the DAPI system is to be directly connected to the chatbot and to detect online interaction problems as soon as a user statement is received. The originality of the proposed method is based on : i) taking into account the history of the dialogue ; ii) the modeling of interaction problems as the expressions of user spontaneous opinion or emotion towards the interaction ; iii) the integration of the web-chat and in-the-wild language specificities as linguistic cues for linguistic rules ; iv) use of lexical word embedding (word2vec) learned on the large untagged chatbot corpus to model semantic similarities. The results obtained are very encouraging considering the complexity of the data : F-score = 74.3%.

KEY-WORDS : Chatbot dialog, Interaction problem, Opinion mining, Human-computer interaction, Written interactions



TELECOM ParisTech

EDiTE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY