Quelles sont les caractéristiques des interactions problématiques entre des utilisateurs et un conseiller virtuel ?

Irina Maslowski
EDF Lab Paris-Saclay, 91120 Palaiseau, France
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
irina.maslowski@edf.fr

RESUME

L'utilisation d'un conseiller virtuel pour la gestion de la relation client sur les sites des entreprises est une solution numérique de plus en plus adoptée. Le défi pour les entreprises est de mieux répondre aux attentes des clients en leur fournissant des interactions fluides entre le client et l'agent. Pour faire face à ce problème, cet article met l'accent sur la détection des problèmes d'interactions dans un corpus de tchat écrit entre un conseiller virtuel et ses utilisateurs. Il fournit une analyse de corpus en décrivant non seulement les spécificités linguistiques et les marqueurs d'opinion contenus dans le corpus du tchat humain-agent, mais aussi les indices linguistiques et dialogiques qui peuvent être pertinents pour caractériser une interaction problématique. Le modèle de règles proposé, utilisant les indices trouvés, est appliqué à un corpus avec des retours client négatifs et positifs pour révéler les tendances.

ABSTRACT

How to characterize problematic interactions between users and a web virtual advisor?
The use of virtual agents on website for the management of customer relationship is an increasing widespread phenomenon. The challenge for the companies is to better meet client's expectations by providing fluid interactions between the customer and the agent. In order to tackle this issue, this paper focuses on the detection of problematic interactions in a corpus of written chat between users and a virtual advisor by providing a corpus analysis dedicated to study not only the linguistic specificities and the opinion markers contained in the human-agent chat corpus, but also the relevant linguistic and dialogic cues that could be used to characterize a problematic interaction. The proposed model of rules using the found clues is applied to a corpus with negative and positive feedback to reveal the tendencies.

MOTS-CLES : problèmes d'interaction, langue tchatée, interaction humain-agent, agent virtuel, dialogue humain-machine

KEYWORDS: interaction problems, chat language, human-agent interaction, artificial conversational entity, human-machine dialogue

1 Introduction

Les interactions humain – agent sont un sujet de recherche majeur et complexe et l'utilisation d'un agent virtuel sur un site web pour la gestion de la relation client est en plein essor. La fluidité de la

conversation entre le client et l'agent virtuel est un défi pour les entreprises et la qualité de l'interaction peut influer sur la capacité d'une entreprise à rester concurrentielle sur le marché.

Nous abordons la question de la qualité de l'interaction dans cet article en nous concentrant sur la détection des interactions problématiques dans des dialogues écrits entre l'humain et l'agent. Auparavant, la détection des problèmes dans des interactions humain-agent a été étudiée dans des dialogues face-à-face par Turk (1996) à l'aide de l'information visuelle et par Walker (2000) en utilisant la prédiction automatique des situations problématiques. Certaines études mentionnent des situations susceptibles de nuire à la fluidité de l'interaction humain-agent, telles que les répétitions des énoncés agent (Bickmore et al., 2005) et les erreurs d'interprétation du sujet de discussion par l'agent.

La détection de problèmes d'interaction a aussi été étudiée dans le cadre d'interaction humain-robot. Ainsi, (Shiomi et al., 2008) a défini trois situations problématiques dans ce type d'interaction : l'absence de la réaction de la part de l'utilisateur suite à deux répliques de suite du robot, la poursuite de la conversation sans une commande particulière pour le robot et la répétition de la demande d'exécution d'une même tâche plus de trois fois. À notre connaissance, notre étude est la première à s'intéresser à la détection de problèmes d'interaction dans une interaction écrite humain-agent.

Notre objectif final est de fournir un système basé sur le Traitement Automatique des Langues Naturelles (TALN) pour la détection automatique des interactions problématiques. L'usage du TALN pour ce type de tâche peut être lié à la détection de *phénomènes reliés à l'opinion*[1] chez l'utilisateur dans une interaction humain-agent (Clavel, à paraitre, 2016). On considère par exemple le comportement verbal de l'agent comme une cible d'opinion et basé soit sur les règles linguistiques (Taboada et al., 2011), soit sur l'apprentissage automatique (Yang, Cardie, 2013), (Socher et al., 2013). Cet article propose une première étape dans cette direction, avec l'étude d'un corpus de tchat humain-agent. Commençant par une brève présentation du corpus dans la Section 2, cette étude tient compte non seulement des spécificités linguistiques (Section 3) et des marqueurs d'opinion (Section 4), mais aussi des indices linguistiques et dialogiques pertinents pour caractériser une interaction problématique (Section 5).

2 Corpus de tchat avec un agent virtuel

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* et enfin "corpus *Laura « utilisateur »*" pour la totalité des énoncés utilisateurs.

Le corpus contient l'ensemble des interactions entre les utilisateurs et *l'agent virtuel LAURA* de janvier à novembre 2014. La conseillère virtuelle du site de l'entreprise EDF répond aux questions des utilisateurs sur la navigation sur le site EDF ou les services. Un dialogue se compose a minima d'une interaction qui à son tour contient un énoncé utilisateur et un énoncé agent. Un énoncé se compose au moins d'une phrase (non-vide); toutefois, il arrive qu'un énoncé soit vide en raison du mode de fonctionnement du système qui enregistrait une ligne vide dès que l'utilisateur changeait de page sur le site-web. Le corpus de l'entreprise EDF a été anonymisé.

[1] Un terme utilisé pour regrouper les termes comme sentiments, émotions et affect.

L'une des particularités de notre corpus est l'apparition d'énoncés utilisateur semi-automatisés. Ils représentent 23 % des énoncés utilisateur non-vides. 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 le lien en adéquation avec sa requête. Cette sélection sera considérée comme étant son propre énoncé. Dans ce cas d'interaction, les spécificités relatives au tchat sont réduites : fautes de frappes, erreurs orthographiques, émotions, etc. sont inexistantes.

La TABLE 1 donne les principaux chiffres qui nous permettent de caractériser le *corpus LAURA*. Le nombre d'énoncés client est légèrement supérieur à celui de l'agent. En effet, l'utilisateur commence pratiquement tous les dialogues excepté quand le client a oublié son mot de passe et l'*agent virtuel* LAURA le détectant, lui propose son aide. 84,5% des énoncés utilisateur contiennent du texte. 3% de ces énoncés non vides contiennent plus d'une phrase. Le nombre de participants correspond à celui de 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*.
Nous avons constaté que plusieurs énoncés utilisateur sont identiques au sein de la même conversation. Nous utilisons cette spécificité en tant qu'un indice de problèmes d'interactions.

Nombre de	**Total**	**Remarques**
Dialogues	1 813 934	En moyenne 164 903/mois
Enoncés utilisateur	6 046 695	En moyenne 3/dialogue, y compris les énoncés vides
Enoncés utilisateurs se répétant	249 258	4% des énoncés
Enoncés agent	6 045 099	En moyenne 3/dialogue
Enoncés agent contenant une URL	1 836 822	30% des énoncés agent

TABLE 1 : Description quantitative du corpus LAURA de janvier à novembre 2014

Pour le corpus de développement, nous avons choisi des interactions du mois de janvier, avril, juillet et octobre. Nous l'appellerons ici corpus *LauraDev*. 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.

Il est à noter que cette étude, réalisée a postériori, a pour objectif de fournir la base nécessaire pour la création d'un système pour la détection automatique des interactions problématiques en temps réel.

La section suivante portera sur les caractéristiques linguistiques du corpus qu'il est important de prendre en compte pour l'analyse automatique des dialogues.

3 Usage langagier des utilisateurs dans leur interaction avec l'agent virtuel

Les caractéristiques du langage des utilisateurs peuvent représenter un obstacle au traitement automatique des dialogues mais elles peuvent également constituer des indices permettant de détecter des problèmes d'interactions. Dans ce qui suit, nous comparons le langage des utilisateurs du tchat avec l'agent virtuel avec celui des utilisateurs des forums de discussion (des dialogues humain-humain) pour révéler des parallèles entre des tchats de finalité différente.

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

TABLE 2 : Caractéristiques des énoncés.

Le vocabulaire des utilisateurs est très focalisé car leurs phrases sont synthétiques : « mon chauffage est collectif que dois-je renseigner ? » (voir TABLE 2). Selon (Achille, 2005), les utilisateurs de forums de discussion ne font pas de phrases plus développées. Le nombre des formes distinctes est approximativement 10 fois plus restreint que dans le canal #18-25ans du corpus du français tchaté de (Achille, 2005). 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. (Achille, 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. »

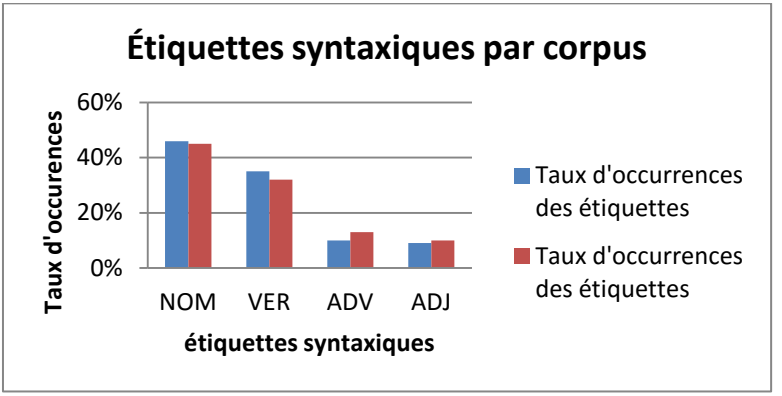

FIGURE 1 : La comparaison de représentativité de 4 types d'étiquettes syntaxiques dans la partie "utilisateur" du corpus LAURA et du corpus WebGRC

Néanmoins, pour s'assurer que les utilisateurs ne réduisent pas leurs énoncés à des mots clés, exempts de toute opinion, nous procédons à l'analyse morpho-syntaxique. Après avoir annoté le corpus avec le logiciel TreeTagger (Schmid, 1995), 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 al., 2012) sur le corpus WebGRC “brut”, corpus des forums clients de l’entreprise. Il s’avère que la distribution, entre ces deux corpus, des quatre types d’étiquettes syntaxiques choisies est comparable (voir FIGURE 1). 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 20%), 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 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? ». Il arrive d’ailleurs que certains clients tentent d’utiliser l’interface de Laura pour soumettre leurs e-mails. Ainsi la structure du texte nous parait favorable pour effectuer une recherche au sein de notre corpus de lexique émotionnel et affectif.

La détermination du type de langage des utilisateurs utilisé dans le corpus s’est imposée afin de prévoir les difficultés que nous rencontrerons lors de traitements plus complexes comme l’annotation automatique des termes exprimant des opinions et des phénomènes reliés aux opinions. Issu du tchat sur Internet, notre corpus porte des caractéristiques du langage français tchaté décrit par (Achille, 2005), telles que 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”. Certaines caractéristiques décrites ci-dessus font partie de l’*argot Internet* défini comme « un mélange de mots anglais, de sigles, d’onomatopées et d’abréviations, d’orthographe phonétique et d’expressions détournées. » (Calis, Candido, Champailler, et al., 2002). Nous utilisons la liste de 490 termes de Wiktionnaire[2] sur notre corpus. Le corpus *LauraDev* comporte 47% de formes différentes d’argot Internet présent dans cette liste, comme les abréviations « svp » pour « s’il-vous-plaît » et « bjr » pour « bonjour » ou « kwa », forme phonétique de « quoi ».

Cette analyse nous amène à réfléchir à la fonction communicative de l’argot Internet et à son utilisation potentielle pour un système de détection automatique des interactions problématiques. L’utilisateur utilise-t-il un langage peu strict, moins clair pour échanger avec l’*agent virtuel* LAURA sur des thématiques qui ne concernent pas le champ d’expertise de l’entreprise (“eh t sais ki?”) ; à moins que cet abus de langage n’indique que l’utilisateur soit pressé ou énervé? (“ et que sa sois fait au ^plus vite svp”, “oui mais persone rep”?) Les fautes de frappes transmettraient le même type d’information. 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.

Dans la section suivante nous étudierons des caractéristiques typographiques du langage tchaté ainsi que le lexique qui peuvent permettre de détecter des problèmes d’interactions.

4 Marqueurs d’opinions et de phénomènes reliés aux opinions

Les marqueurs d’Opinions et de Phénomènes Reliés aux Opinions (OPRO) peuvent nous permettre de détecter des opinions et de phénomènes reliés aux opinions négatives dont la cible (Martin,

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

White, 2005) est l'interaction. Par conséquent, nous nous intéressons, dans un premier temps, aux marqueurs spécifiques du web et, dans un second temps, aux marqueurs lexicaux.

4.1 Caractères échos, interjections et émoticônes.

Il existe des marqueurs qui permettent d'exprimer des OPRO dans les textes écrits sans pour autant les exprimer explicitement par des mots descriptifs. Cela peut être l'utilisation exagérée de majuscules (Yates, Orlikowski, 1993), d'onomatopées (Anis, 2006) ou la répétition de caractères (Panckurst, 2006) qui simulent la prononciation de sons, des sigles comme « lol » (version anglaise) ou « mdr » (version française) qui expriment le rire et sont utilisés en parallèle par les utilisateurs français (Lorenz & Michot, 2012) et de façon plus non-verbale : la ponctuation multiple (Dutrey et al., 2012) et les émoticônes (Marcoccia, Gauducheau, 2007), (Lorenz, Michot, 2012). Nous avons tout d'abord étudié l'usage de ces types de marqueurs potentiels d'OPRO, en complétant l'analyse de l'argot Internet présenté dans la partie précédente par l'analyse : (1) de la ponctuation multiple et des caractères échos, (2) des interjections, (3) des émoticônes. Nous nous appuyons sur la définition de (Cougnon, 2014) 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 pour la 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, Herring, 2010) et (Marcoccia, Gauducheau, 2007) et utilisons le dictionnaire fourni par la société DataGenetics[4]. Ce dictionnaire contient 2 242 émoticônes classés par ordre de fréquence décroissante.

La FIGURE 2 présente la répartition de ces différents types de marqueurs dans notre corpus. Notre corpus comporte en moyenne 0,04 marqueur par énoncé non vide. Les marqueurs les plus représentés (68%) sont les ponctuations multiples mais l'usage de ces marqueurs reste cependant marginal car seul 1 % des énoncés non vides comprend une ponctuation multiple.

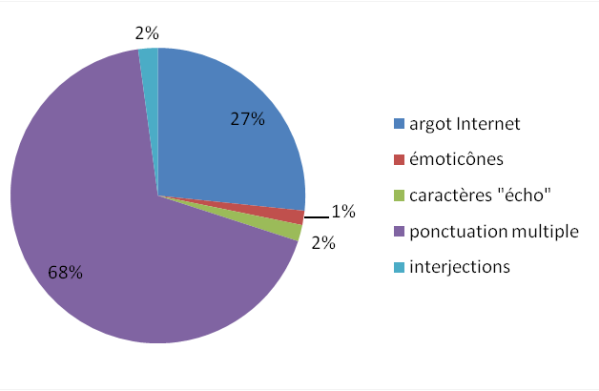

FIGURE 2 : Répartition des principaux phénomènes lexicaux parmi ceux détectés.

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

[4] http://www.datagenetics.com/blog/october52012/index.html

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 « commment » (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 », « pfffffffffffffffffffffffffffffffffffffffff », etc.), dont l’intensité nous renseigne sur le degré émotionnel que l’utilisateur souhaite communiquer, phénomène 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’OPRO 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 non vide 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 phénomènes reliés aux opinions de l’utilisateur. Les émoticônes “standard” 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, Carpenter, 2005).

En nous reposant sur la typologie de la relation entre les émoticônes et le contenu verbal de (Marcoccia, 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 : information redondante (« ok merci beaucoup et bonne journée Laura :-) »); aide à la compréhension (« Bonjour, j'aimerai savoir comment payer ma facture sur internet :) ») ; atténuation (« Laura !!! fait un effort ma poulette !!!! :) »); ironie («c 'était pourtant clair :-) » ; familiarité: « vous avez qulle age ? :) »). 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 al., 2008).

4% d’énoncés utilisateur contiennent au moins un marqueur, y compris l’argot Internet. 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 avons recours à ces phénomènes lors de la conception de nos règles décrites dans la Section 5.

4.2 Lexique d’émotion dans le corpus

Les OPRO envers l’interaction peuvent aussi être exprimées expressément dans le texte rédigé par l’utilisateur. Pour pouvoir les détecter nous avons choisi d’utiliser un dictionnaire d’OPRO. Parmi ceux disponibles en langue française, nous avons choisi d’utiliser la version française du dictionnaire pour LIWC (Piolat, Booth, Chung, et al., 2011) 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 al., 1988), et des dictionnaires de mots communs. Les mots sont divisés en 64 catégories liées d’une manière ou d’une autre aux «

processus psychologiques ». Ce sont non seulement les catégories qui contiennent 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, Chung, Ireland, et al., 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.

En combinant le lexique d'OPRO avec quelques règles basiques de reversement de polarité, nous avons annoté 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 *LauraDev* par le lexique en dépit des ambigüités non-résolues. Ainsi, environ 25% de dialogues du corpus contiennent une annotation *Emo**: 15% d'*EmoPos* et 10% d'*EmoNeg*. Ce résultat est prometteur pour l'utilisation du dictionnaire LIWC pour détecter les OPRO. Il est néanmoins nécessaire de prendre en compte le contexte afin de désambigüiser les termes d'émotions, par exemple à l'aide d'un lexique de modifieurs de polarité (Taboada et al., 2011).

La Section 5 décrit l'utilisation de certains marqueurs d'OPRO pour tenter de détecter des interactions problématiques.

5 Les caractéristiques des interactions problématiques

Le but de cette section est d'étudier les indices qui peuvent être pertinents pour la détection des interactions problématiques et que nous avons identifiés en observant le corpus *LauraDev*. Nous proposons un premier modèle d'indices linguistiques à travers des règles conçues pour leur utilisation dans le module JAPE (Cunningham et al., 2000) du logiciel GATE. Nous proposons trois séries de règles : les règles basées sur le contexte conversationnel, sur les signes de répétitions faites par l'utilisateur et les règles basées sur le contexte des phrases.

Les sous-corpus. Trois sous-corpus sont créés pour pouvoir analyser le comportement de ces règles sur les données. Le premier contient 103 791dialogues évalués comme négatifs à travers le *feedback*[5] des utilisateurs. Le second comporte 107 796 dialogues évalués en tant que positifs. Le troisième compte 62 373 dialogues avec une mention « failed » dans les métadonnées et ne contenant aucun dialogue avec *feedback*. La métadonnée « failed » est obtenue automatiquement soit dans les cas où l'agent n'a trouvé aucune information correspondant à la question de l'utilisateur, soit lorsque l'utilisateur n'a pas cliqué sur un des liens proposés par l'agent. Même si ces annotations ne peuvent pas être considérées comme une référence fiable pour une évaluation des

[5] Un avis d'un utilisateur

règles développées, elles nous permettent d'analyser des tendances dans le comportement des règles et de mieux identifier les caractéristiques pertinentes des interactions problématiques.

5.1 Les règles basées sur le contexte conversationnel

Les règles reposant sur le contexte conversationnel tiennent compte de deux énoncés consécutifs de l'utilisateur afin de détecter une répétition complète ou partielle dans les énoncés utilisateur. La distance entre deux énoncés utilisateur est mesurée avec les distances de Levenshtein (seuil 4) et Jaccard (seuil 0,3). Les seuils ont été choisis de manière expérimentale afin d'obtenir le meilleur compromis entre les éléments suffisamment proches pour être considérés comme des répétitions et les faux-positifs (ils détectent 1,6 fois plus d'interactions dans les sous-corpus « failed » et « négatif », que dans le sous-corpus « positif »).
Le calcul de la distance de Levenshtein est intéressant pour la détection des énoncés qui diffèrent en quelques lettres, alors que la distance de Jaccard permet de détecter des énoncés qui se distinguent l'un de l'autre en quelques mots. Les exemples suivants illustrent l'intérêt des deux types des distances : *"atestation" /"n° attestation"* (jaccard="1.0" levenshtein="4"); *"ou se trouve mon estimation sur le site Y"*/*"ou trouvé mon estimation sur le site Y"* (jaccard="0.3" levenshtein="4").

5.2 Les règles basées sur les signes de répétitions faites par l'utilisateur

La seconde série de règles repose sur l'analyse des signes de répétitions faits par l'utilisateur et des marqueurs d'émotions (voir la Section 4.1). Suite aux résultats de l'analyse linguistique du corpus, la ponctuation multiple parait être le marqueur d'émotion le plus pertinent et le moins ambigu à utiliser dans les règles (voir la FIGURE 1). Puisque la thématique des dialogues que nous étudions est fortement liée à celle de l'entreprise, un autre élément important de ce type de règles sont des termes métier tels que « Alerte Relevé Confiance », « Espace client », « facture électronique » ou encore « la puissance d'un compteur ». Nous considérons que les termes métiers, sans être identiques d'un énoncé à l'autre, couplés avec des marqueurs d'émotions, pris en compte dès le second énoncé de l'utilisateur, peuvent indiquer des problèmes d'interactions.

Cela est confirmé par la réalisation d'un apprentissage supervisé avec une classification naïve bayésienne de deux classes « failed » et « autre » sur un échantillon de 1 000 dialogues : 8 des 15 marqueurs les plus discriminants pour la classe « failed » sont des termes métiers tels que : « resilier », « domicile », « justificatif ».

L'apparition d'une émotion initialement absente dans un contexte métier peut être un témoin d'une réaction de l'utilisateur au déroulement de l'interaction. Les observations du corpus indiquent que ce type de configuration de marqueurs est le plus souvent l'expression d'une opinion négative envers l'interaction (on y trouve l'insistance et l'agacement de l'utilisateur, la répétition de questions précédemment posées...).
Nous utilisons les trois émoticônes négatifs les plus fréquents (deux variantes de l'émoticône « triste » :(et :-(, et « embêté » :/) ainsi qu'un autre marqueur fort des émotions négatives : les insultes. Ainsi, l'annotation des dialogues avec ce type de règles se déroule en trois étapes : 1) l'annotation des termes métier à l'aide d'un dictionnaire de termes métier, fourni par l'entreprise, et du dictionnaire LIWC pour les insultes ; 2) l'annotation de la ponctuation multiple en utilisant une règle qui prend en compte des signes d'exclamation, interrogation et leurs combinaisons se répétant au moins deux fois, des points et points-virgules se répétant au moins quatre fois ; 3) l'annotation d'un problème d'interaction dans une partie d'énoncé ou dans un énoncé entier.

Nous avons conçu quatre règles (voir le TABLE 4) qui annotent des problèmes d'interaction en s'appuyant sur les indications décrites ci-dessus. Dans la description de ces règles (R), nous utiliserons n pour l'énoncé courant, *TM* pour un terme métier, P_m pour la ponctuation multiple, Q_u pour une question d'utilisateur, $Emot_n$ pour un émoticône exprimant une émotion négative et A pour l'annotation obtenue. La TABLE 3 définit le formalisme que nous avons créé pour décrire les règles.

A(n)	Présence de l'annotation A dans l'énoncé n	A(m,n)	Présence de l'annotation entre les énoncés m et n inclus
A(n) OU B(n)	Présence de l'annotation A à l'énoncé n ou de l'annotation B à l'énoncé n	A(n) ET B(n)	Présence de l'annotation A à l'énoncé n et de l'annotation B à l'énoncé n
A(n) <ET B(n)	Présence de l'annotation A à l'énoncé n et de l'annotation B à l'énoncé n avec A précédent B	$(A \subset B)(n)$	L'annotation A existe dans l'énoncé n et contient l'annotation B

TABLE 3 : Définition du formalisme adopté.

Règle	**Détail**	**Exemple**
Interaction Problem 1 soit R_1	$R_1 = A_1 = TM$ (n-1, n) < ET Pm(n)	« *commen fair pour relever le conteur de gaz ??* »
Interaction Problem 2 soit R_2	$R_2 = (TM$ (n-1) < ET A_1 (n)) OU (TM ((1, n-1) < ET P_m (n)), A_2 = n	*Utilisateur: "comment faire pour releverle comteur de gaz pas de cle pour ouvrir" Agent: (…) Utilisateur: "commen fair pour relever le conteur de gaz ??"*
Interaction Problem 3 soit R_3	$R_3 = (A_1$ (n-1) < ET insulte (N)) OU (($Q_u \subset TM$)(1, n-1) < ET insulte (n), A_3 = n	*Utilisateur: « commen fair pour relever le conteur de gaz ?? » Agent: (…) Utilisateur:* insulte *Utilisateur « Comment accéder à mon deuxième contrat ? »*, Agent : (…) *Utilisateur :* insulte
Interaction Problem 4 soit R_4	$R_4 = TM$ (1, n-1) < ET $Emot_n$ (n), A_4 = n	*Utilisateur « mon paiement en ligne semble ne pas avoir été pris en compte », Agent : (…) Utilisateur : « plutôt moyen comme service... :-(»*

TABLE 4 : Règles pour la détection de problèmes d'interaction.

Les règles *Interaction Problem 2 et Interaction Problem 3* utilisent l'annotation effectuée par la règle *Interaction Problem 1* pour détecter des cas plus complexes. Ces deux règles consistent chacune en deux niveaux d'analyse des énoncés pour ne pas être trop restrictives.
Dans toutes les règles *Interaction Problem1 – 4,* les termes métier doivent nous permettre d'éviter l'annotation des dialogues « hors sujet ». La détection de marqueurs de sentiments se fait à partir du deuxième énoncé de l'utilisateur (sauf pour la règle *Interaction Problem1)* pour ne pas prendre en compte les humeurs qui ne sont pas liées à l'interaction avec l'agent.

En ce qui concerne les autres spécificités du langage des utilisateurs que nous avons présenté plus haut, il est intéressant d'observer leur distribution entre les trois sous-corpus. Les dialogues avec une métadonnée « failed » contiennent beaucoup plus d'argot Internet que les dialogues avec les feedbacks positif et négatif (9% de dialogues dans« failed » contre 0,4% dans « feedback négatif » et 0,3% dans « feedback positif »). Le phénomène « écho » est moins répandu que l'argot Internet mais la même tendance peut être observée : 0,8% de dialogues du sous-corpus « failed » en contiennent, alors que seulement 0,04% de dialogues dans le sous-corpus « feedback négatif » et 0,03% de celui de « feedback positif ». Il est fort probable que ces particularités du langage utilisateur représentent un obstacle conséquent pour la compréhension de l'objet de la requête par l'*agent Laura* et entrainent l'incohérence de ses réponses. Ceci expliquerait leur nombre dans le sous-corpus « failed ». Nous étudierons la manière dont l'argot Internet, les phénomènes « écho » et les interjections peuvent nous aider à détecter un problème d'interaction dans notre futur travail.

5.3 Les règles basées sur le contexte des phrases

Les règles basées sur le contexte d'une phrase ont pour but d'identifier l'opinion d'un utilisateur sur l'interaction en tant que relation entre une opinion et sa cible (Martin, White, 2005). Premièrement, nous nous concentrons sur la modélisation de la cible d'opinion qui peut être exprimée par l'utilisateur avec des expressions de type « votre réponse », comme dans l'exemple suivant : « *j'attends votre réponse....* ». Deuxièmement, nous modélisons des phrases des utilisateurs qui expriment un reproche, telles que « pourquoi vous me dites/pourquoi tu me dis », « vous me répondez/tu me réponds ». Les règles développées peuvent être exprimées par des expressions régulières : {pourquoi}({v(ou)?s}|{tu})({token})?{dire}, qui annote, par exemple, « *pourquoi vous me dite que mon adresse mail est deja utilisé alors que je fait seulement m'inscrire »*) et ({v(ou)?s}|{tu})({token})?{r[ée]pond.*} (« *vous ne repondez pas à ma question* »). Ces règles prennent en compte quelques variations orthographiques les plus fréquentes.
Les dialogues qui comportent des phrases du type « Merci pour/de votre réponse », « je vous remercie de votre réponse » ou encore « en attente de votre réponse », ne sont pas annotés grâce à la règle suivante {attente|merci|remercier}({token})?{votre}{r[éè]ponse}, où les mots « attente », « merci », « remercier », « votre » sont pris en tant que des lemmes.

5.4 Les résultats d'annotations

La règle *Interaction Problem1* n'est pas assez discriminante pour la détection de problèmes d'interactions car elle reflète plutôt l'irritation générale d'un utilisateur, puisque elle n'est pas limitée en énoncés utilisateur et peut annoter dès le premier énoncé d'un dialogue. Nous l'utilisons comme un élément de base pour d'autres règles. Néanmoins, il est intéressant de noter que les fichiers avec des dialogues contenants la métadonnée « failed » contiennent plus d'annotation en « *Interaction Problem1 »* (4%) que les fichiers avec feedback négatif (2,6%) et feedback positif (1%). Nous supposons qu'il est plus compliqué pour un agent virtuel d'interpréter un discours

rempli d'OPRO et de satisfaire une demande utilisateur.
Les règles *Interaction Problem 2 et 3* discriminent assez bien les différents sous-corpus : *Interaction Problem 2* est détectée dans 2% de dialogues dans le sous-corpus « failed », 1,5% dans « feedback négatif » et seulement 0,5% dans « feedback positif » ; *Interaction Problem 3* détecte des problèmes assez rares avec 0,04% dans les dialogues « failed », 0,4% dans « feedback négatif » et 0,1% dans le « feedback positif ».
La règle *Interaction Problem 4* nous semble ne pas avoir beaucoup d'intérêt pour la détection, car elle ne trouve que 9 occurrences au total, tous corpus confondu. Les dialogues identifiés sont le plus souvent « hors-sujet ».

Les cas où l'utilisateur exprime son opinion expressément dans une phrase sont rares. Ainsi, les règles avec un reproche couvrent 0,4% des dialogues avec une métadonnée « failed », 0,2% des dialogues avec un feedback négatif et 0,1% de dialogues avec un feedback positif.
Les règles basées sur le calcul du seuil des distances de Levenshtein et Jaccard permettent la détection de 7% de dialogues avec un feedback négatif et de 7% de dialogue avec une métadonnée « failed », et de 4% des dialogues avec un feedback positif.

L'union de toutes les règles nous permet d'identifier 10% des dialogues dans le sous-corpus avec feedback négatif et 6% des dialogues contenus dans le sous-corpus avec feedback positif. Ce résultat est encourageant. Il pourra être confirmé et précisé après une campagne d'annotation manuelle d'un corpus et d'évaluation du système.

Conclusion

Dans cet article nous avons proposé une étude de corpus du tchat écrit entre une conseillère virtuelle d'un site web commercial et un utilisateur. Nous avons proposé des marqueurs d'opinion et de phénomènes reliés aux opinions et une première approche pour la détection des interactions problématiques dans les dialogues humain-agent. Afin de couvrir un champ plus large de cas détectés, trois types d'indices linguistiques sont exposés : les indices basées sur le contexte conversationnel, sur les signes de répétitions faits par l'utilisateur, et ceux, basées sur le contexte des phrases.

Nous avons présenté une analyse de tendances d'efficacité des indices proposés obtenues grâce à l'application des règles linguistiques à des sous-corpus avec le retour client négatif et positif. Ainsi, des indices, tels que la répétition de phrases, la ponctuation multiple et la présence d'insultes semblent être performants, alors que les émoticônes ne représentent pas d'intérêt en tant qu'indice.

L'originalité de ce travail est la contextualisation des marqueurs d'opinion et de phénomènes reliés aux opinions et l'utilisation des termes métier pour filtrer les conversations « hors-sujet » en prenant en compte plusieurs tours de paroles.

Dans nos futurs travaux, grâce à cette étude, nous choisirons pour notre système des règles linguistiques les plus performantes. Nous étudierons la possibilité d'élargir le nombre d'indices linguistiques utilisés pour nos règles : l'argot, les caractères « écho », les interjections sont des indices à prendre en considération. Ensuite, nous procéderons à l'annotation manuelle d'une partie de notre corpus pour l'évaluation de notre système. Enfin, nous prévoyons d'intégrer les résultats d'annotation à base de règles linguistiques dans un système d'apprentissage automatique.

Références

ANIS J. (2006). Communication électronique scriptural et formes langagières, *http://edel.univ-poitiers.fr/rhrt/document.php?id=547#documents*, consulté le 17/10/2015

ACHILE F. (2005). Constitution d'un corpus de français tchaté. *RECITAL*, Dourdan, France

BICKMORE T W., PICARD R W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no 2, p. 293-327.

CALIS L., CANDIDO G., CHAMPAILLER S., ET AL.(1998). *Anthropologie de la société digitale.*

CLAVEL C., CALLEJAS Z. (à paraitre, 2016) Sentiment analysis: from opinion mining to human-agent interaction. *Affective Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1.

COUGNON L.-A.(2015) Langage et SMS: Une étude internationale des pratiques actuelles. *Presses universitaires de Louvain*.

CUNNINGHAM H., MAYNARD D., TABLAN V. (2000). JAPE: a Java Annotation Patterns Engine (Second Edition). *Technical report* CS—00—10, University of Sheffield, Department of Computer Science.

DE ANGELI A., CARPENTER R. (2005). Stupid computer! Abuse and social identities. In : *Proceedings of the INTERACT 2005 workshop* Abuse: The darker side of Human-Computer Interaction.

DRESNER E., HERRING S. C. (2010).Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication theory*, vol. 20, no 3, p. 249-268.

DUTREY C., PERADOTTO A., CLAVEL C. (2012). Analyse de forums de discussion pour la relation clients: du Text Mining au Web Content Mining. *Actes JADT*.

FERRARI S., MATHET Y., CHARNOIS T., ET AL. (2008). Analyse d'opinion: discours évaluatif et classification de documents. *Actes de l'atelier FODOP*, VOL. 8, P. 23-36.

LORENZ P., MICHOT N. (2012). Le lexique du chat sur Internet: étude comparative français-espagnol-polonais. In : *SHS Web of Conferences. EDP Sciences*. p. 939-954.

MARCOCCIA M., GAUDUCHEAU N. (2007). L'analyse du rôle des smileys en production et en réception: un retour sur la question de l'oralité des écrits numériques. *Glottopol, revue sociologique en ligne*.Récupéré de *http://www. univ-rouen. fr/dyalang/glottopol/numero_10. html# sommaire* [29 October 2007].

MARTIN J. R., WHITE P. R. R. (2005). *The Language of Evaluation: Appraisal in English*. Palgrave, London.

PANCKURST R. (2006). Le discours Électronique médié : bilan et perspectives [a], dans A.Piolat (dir.). *Lire, Ècrire, communiquer et apprendre avec Internet*, Marseille, Editions Solal, pp. 345-365.

PENNEBAKER J. W., CHUNG C K., IRELAND M., ET AL. (2007) *The development and psychometric properties of LIWC2007.*

PIOLAT A., BOOTH R. J., CHUNG C K., ET AL. (*2011).* La version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. *Psychologie française*, vol. 56, no 3, p. 145-159.

PIOLAT A., BANNOUR R. (*2009).* EMOTAIX: un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année psychologique*, vol. 109, no 04, p. 655-698. (pour EMOTAIX)

SHIOMI M., SAKAMOTO D., KANDA T., ET AL. (2008). A semi-autonomous communication robot: a field trial at a train station. In : *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction. ACM*. p. 303-310.

SCHMID H. (1995). *Treetagger a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, vol. 43, p. 28.

SOCHER R., PERELYGIN A., WU J. Y., ET AL. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *In : Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. p. 1642.

TABOADA M., BROOKE J., TOFILOSKI M., ET AL. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, , vol. 37, no 2, p. 267-307.

TURK M. (1996). Visual interaction with lifelike characters. In : Automatic Face and Gesture Recognition, 1996., *Proceedings of the Second International Conference on. IEEE*. p. 368-373.

WALKER M., LANGKILDE I., WRIGHT J., ET AL. (2000). Learning to predict problematic situations in a spoken dialogue system: experiments with how may I help you? In : *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. Association for Computational Linguistics*. p. 210-217.

WATSON D., CLARK L. A., TELLEGEN A. (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.

YANG B., CARDIE C. (2013). Joint Inference for Fine-grained Opinion Extraction. *In ACL* (1).. p. 1640-1649

YATES J., ORLIKOWSKI W J., ET AL. (1993). *Knee-jerk anti-loopism and other e-mail phenomena: Oral, written, and electronic patterns in computer-mediated communication*. Alfred P. Sloan School of Management, Massachusetts Institute of Technology.