Comparaison d'approches de classification automatique des actes de dialogue dans un corpus de conversations écrites en ligne sur différentes modalités

Soufian Salim Nicolas Hernandez Emmanuel Morin
LINA UMR 6241, Université de Nantes, 2 rue de la houssinière, 44322 Nantes Cedex 03
{prenom}.{nom}@univ-nantes.fr

RÉSUMÉ

L'analyse des conversations écrites porteuses de demandes d'assistance est un enjeu important pour le développement de nouvelles technologies liées au support client. Dans cet article, nous nous intéressons à l'analyse d'un même type d'échange sur un canal différent : les conversations se déroulant sur les plate-formes d'entraide entre utilisateurs. Nous comparons des approches de classification supervisées sur trois modalités des CMR [1] différentes à même thématique : des courriels, forums et chats issus de la communauté Ubuntu. Le système emploie une taxonomie fine basée sur le schéma DIT++. D'autres expériences sont détaillées, et nous rapportons les résultats obtenus avec différentes approches et différents traits sur les différentes parties de notre corpus multimodal.

ABSTRACT

A comparison of automatic dialog act recognition approaches in a multimodal corpus of online written conversations

The analysis of online written conversations bearing requests for assistance is a major challenge for the developement of novel customer support technologies. In this paper, we focus on the analysis of a very similar type of communication: conversations that take place on community-driven user assistance platforms. We compare supervised classification approaches on three distinct CMC modalities on thematically similar data: emails, forums and chats gathered from the Ubuntu community. Additionnal experiments are detailed, and we reported our results with different approaches and feature sets on the different parts of our multimodal corpus.

MOTS-CLÉS : Analyse du discours, Conversation, Acte de dialogue, Multimodalité, CMR.

KEYWORDS: Discourse analysis, Conversation, Dialogue act, Multimodality, CMC.

1 Introduction

Les capacités d'interaction entre internautes se sont spectaculairement accrues au cours des dernières décennies, et avec elles les plate-formes d'échange participatives. Ces plate-formes, qui se déclinent en différentes modalités - salons de chat, listes de diffusion, forums de discussions - sont souvent utilisées par les particuliers comme par les professionnels pour demander et offrir de l'aide. Ce type d'échanges, les conversations écrites en ligne orientées vers la résolution de problèmes, représentent un champ de recherche encore peu exploré par la communauté scientifique. Cette forme de com-

1. Communications Médiées par les Réseaux

munication peut néanmoins être exploitée, et est très similaire aux conversations entre agents et utilisateurs des services de GRC [2] des entreprises qui assurent une présence en ligne. C'est pourquoi les conversations porteuses de demandes d'assistance intéressent déjà les industriels, qui s'efforcent de développer des systèmes adaptés à la gestion de ce type d'échanges. Développer un système capable de les analyser automatiquement représente ainsi non seulement un enjeu industriel important, mais permettrait également d'améliorer les plate-formes d'échanges collaboratives qui sont quotidiennement sollicitées par des millions d'utilisateurs. Cependant, les techniques d'identification de la structure des conversations n'ont pas été développées autour des conversations écrites en ligne. Il serait intéressant de déterminer l'efficacité de ces techniques sur des données extraites de conversations orientées vers la résolution de problèmes en fonction de leur modalité. Cette problématique est à placer dans le cadre de la recherche en communication médiée par les réseaux (CMR), et nous confronte à ses problématiques propres : données en français de différents registres et à forte variation orthographique, faibles performances des outils linguistiques, absence d'informations prosodiques, structures de conversations non-linéaires, *etc*.

Dans la littérature, les interactions entre humains sont typiquement modélisées en termes d'actes de dialogue. Les actes de dialogue sont les héritiers des actes de langage, qui désignent les actes que l'on fait en parlant. Dans la théorie d'Austin (1975), tout énoncé peut être analysé à trois niveaux. D'abord, au niveau locutoire : il s'agit de sa forme de surface, *i.e.* de la signification de l'énoncé et des mots qui le composent, d'un point de vue phonétique, syntaxique et sémantique. Puis au niveau de l'acte illocutoire, porteur de l'intention rhétorique du locuteur. Et enfin, au niveau de l'acte perlocutoire, qui s'intéresse aux conséquences directes et pragmatiques de l'énoncé (par exemple, faire accomplir une action à quelqu'un). C'est l'analyse des énoncés en termes d'intention communicative, c'est-à-dire de leur niveau illocutoire, qui a mené à la popularisation de cette théorie, notamment au travers des travaux de Searle (1969) pour qui tout acte est illocutoire. Aujourd'hui, les actes de dialogue - ou « actes de la conversation » (Traum & Hinkelman, 1992) - désignent typiquement les types de fonctions remplies par les énoncés dans un discours. Ainsi, dans ce travail, nous nous focalisons sur la reconnaissance des actes de dialogue présents dans les énoncés de notre corpus. L'approche la plus commune et efficace pour la tâche de classification d'énoncés en termes d'actes de dialogue est d'employer des techniques supervisées d'apprentissage machine (Tavafi *et al.*, 2013). C'est également l'approche que nous retenons. Les données d'apprentissage ne manquent pas : dans le cas des forums et des listes de diffusion, les conversations sont généralement perpétuellement sauvegardées pour permettre aux utilisateurs de faire des recherches dans les archives. Ce n'est pas toujours le cas pour les messages transmis dans les salons de chat, mais ils peuvent eux aussi faire l'objet d'un archivage automatique et de nombreux outils existent à cette fin.

Cet article a pour objectif de comparer les approches communes pour la classification des actes de dialogue dans un corpus de conversations écrites en ligne orientées vers la résolution de problèmes décliné en différentes modalités. Dans un premier temps, nous examinerons l'état de l'art et évoquerons les travaux similaires. Ensuite, nous justifierons notre choix de taxonomie pour la modélisation des conversations. Nous détaillerons ensuite l'approche que nous adoptons pour notre tâche de reconnaissance automatique, avant d'exposer nos méthodes d'évaluation. Enfin, nous présenterons les expériences effectuées et discuterons des résultats obtenus.

2. Gestion Relation Client

2 Travaux similaires

La plupart des travaux existant dans le domaine de la détection des actes de dialogue s'intéressent à l'étude de conversations orales. C'est notamment le cas de nombreux travaux fondateurs en matière d'analyse du dialogue, qui reposent souvent sur le même corpus, TRAINS (Traum & Hinkelman, 1992; Poesio & Traum, 1997; Core & Allen, 1997; Bunt, 2009). C'est dû à l'objectif de ces chercheurs, à savoir développer des systèmes de dialogues dans lesquels le système prend la place d'un agent humain pour assister l'opérateur dans sa tâche. Le caractère textuel des conversations médiées par les réseaux nous prive ainsi d'importantes informations prosodiques, qui ont été démontrées utiles pour la classification d'actes de dialogue (Ang *et al.*, 2005; Fernandez & Picard, 2002). Les conversations qu'ils étudient et modélisent sont également souvent bi-participants, tandis que nous nous intéressons principalement à des conversations multi-participants, ce qui accroît la complexité de la tâche. Enfin, le rôle des utilisateurs est souvent pris en compte par le système pour effectuer ses choix, or cette information n'est souvent pas disponible ou non-applicable à nos conversations (*e.g.*, un système de *e-learning* bénéficie grandement de savoir si l'énoncé vient d'un élève ou d'un professeur puisque l'un est supposé poser des questions tandis que l'autre est chargé d'y répondre et de donner des instructions). Enfin, l'écrasante majorité des travaux portent sur la reconnaissance des actes de dialogue dans des corpus en anglais, mais très rarement en français.

Plusieurs travaux ont déjà porté sur des tâches de classification d'actes de dialogue dans des corpus de conversations en ligne. C'est l'approche supervisée qui est employée dans l'immense majorité des cas. Tavafi *et al.* (2013) étudient les travaux effectués dans le domaine et testent un panorama de techniques de classification employées dans la littérature. Le travail soutient l'idée que les actes de dialogue doivent être étiquetés séquentiellement pour une meilleure performance. Leur conclusion est que le modèle SMV-HMM est plus performant que les autres techniques testées, à savoir l'utilisation de champs conditionnels markoviens (CRF) et l'utilisation d'un SVM multiclasse. Cependant, l'approche séquentielle ne peut pas nécessairement être appliquée aux chats puisque plusieurs conversations peuvent se produire simultanément sur le même canal, produisant des conversations entremêlées dont les messages ne peuvent pas être immédiatement identifiés. Ils testent notamment leurs techniques sur des corpus de réunions (MRDA) et de conversations téléphoniques (SWBD), mais aussi de courriels (BC3) et de forums (CNET). Leur travail n'étudie pas le cas des chats. Seul le corpus CNET correspond bien à notre tâche, puisqu'il contient des échanges visant à résoudre des problèmes utilisateurs. L'expérience présentée a cependant l'inconvénient de ne pas rapporter des résultats très satisfaisants (58 % micro-précision, 17 % macro-précision). En outre les classes employées, qui sont assez restreintes, sont principalement composées de déclinaisons de « question » et de « réponse ». Elles reposent sur une définition assez limitée de ce qu'est un acte de dialogue pour pouvoir être appliquée à leur ensemble varié de corpus. Leur travail présente l'intérêt de comparer la classification des actes de dialogue au travers différentes modalités, mais le fait d'utiliser une taxonomie différente pour chacune d'entre elles limite fortement les conclusions que l'on peut tirer des résultats rapportés.

Cohen *et al.* (2004) s'intéressent à la classification de courriels en termes d'actes de discours. Leur taxonomie est constituée d'actes composés d'un nom et d'un verbe (*e.g.* REQUEST MEETING), et n'est pas basée sur la théorie des actes de dialogue. Par ailleurs, ils ne cherchent pas à classifier les énoncés mais les messages entiers, ce qui éloigne également leur travail de notre tâche. Lampert *et al.* (2006) en revanche s'intéressent à la classification des énoncés dans les courriels, et basent leur taxonomie sur les VRM (*Verbal Response Mode*), qui trouvent leur source dans la théorie des actes du langage d'Austin. Leurs résultats montrent que leur approche basée sur un classifieur SVM est crédible. Cependant les classes de la taxonomie VRM (DISCLOSURE, EDIFICATION, ADVISEMENT,

CONFIRMATION, QUESTION, ACKNOWLEDGEMENT, INTERPRETATION et REFLECTION) ne sont pas suffisantes pour permettre une analyse fine des conversations. Elles ne permettent par exemple pas de faire la différence entre les fonctions commissives et les fonctions expressives (*e.g.* « je vais me coucher » et « j'aime l'opéra ») ni de distinguer une demande d'action d'une demande d'information, ni de capturer les énoncés de politesse, *etc*.

En ce qui concerne les forums, Qadir & Riloff (2011) cherchent à classifier les actes de discours dans les messages. Ils distinguent le texte contenant ces actes de ce qu'ils nomment « texte expositif », qu'ils définissent comme le discours comportant uniquement de l'information factuelle. Leur taxonomie est composée de quatre classes de Searle (1976) : les commissifs, les directifs, les expressifs et les représentatifs. Ils ignorent les déclaratifs, trop rares dans leur corpus, et les représentatifs, qu'ils ne semblent pas considérer comme des actes de langage, contrairement à la théorie qui leur attribue un caractère illocutoire (à savoir, l'intention de faire connaître leur contenu sémantique). Malgré cette taxonomie assez limitée, leur travail rapporte des résultats encourageants, d'autant plus que les traits utilisés pour leur classifieur - un SVM - ne demandent pas d'analyse linguistique et se prêtent bien aux phrases peu grammaticales que l'on peut rencontrer dans les corpus de CMR.

Les chats sont également fréquemment utilisés pour la résolution collaborative de problèmes, et se distinguent deux autres modalités étudiées par leur caractère synchrone. Ha *et al.* (2013) s'y intéressent dans le cadre du développement de systèmes de dialogue : ils cherchent à la fois à identifier le type d'acte accompli par l'utilisateur et à choisir le type d'acte que le système doit produire. Leur approche utilise un classifieur d'entropie maximale (*MaxEnt*). Leur corpus est constitué de situations de tutorat plus que de situations d'aide à la résolution de problèmes, ce qui est différent même si les deux ont des similarités. Cette distinction se ressent dans leur taxonomie, qui inclut des classes particulièrement présentes dans les dialogues de tutorat (*e.g.* HINT, POSITIVE FEEDBACK). Leurs résultats dépassent l'état de l'art pour la tâche de classification, et ils parviennent également à prédire le timing des interventions du tuteur avec une bonne précision. Ivanovic (2005) cherche à classifier les énoncés des messages chats et se base sur une taxonomie dérivée du schéma d'annotation DAMSL (Core & Allen, 1997). Il parvient à une précision de 80 % en combinant un classifieur naïf de Bayes et un modèle de n-grammes. Kim *et al.* utilisent la même taxonomie qu'Ivanovic et l'appliquent à un corpus de conversations biparticipants (Kim *et al.*, 2010) et multiparticipants (Kim *et al.*, 2012). Un classifieur basé sur les CRF obtient des précisions extrêmement élevées (plus de 97 %) pour les deux tâches en combinant des traits lexicaux, structurels et relationnels.

3 Taxonomie

Tandis que jusque dans les années 1990 la théorie des actes de langage se préoccupait principalement de l'analyse d'énoncés isolés, elle a plus tard incorporé les notions de contexte et de terrain d'entente (Traum & Hinkelman, 1992), qui représentent le fait que l'information doit être synchronisée entre les participants pour que la conversation puisse aller de l'avant. Un nombre important de schémas d'annotation se sont développés autour de ce concept, tels que DAMSL (Core & Allen, 1997) et DIT++ (Bunt, 2009). DAMSL est un standard *de facto* en analyse du dialogue, grâce à ses fondements théoriques (les actes sont annotés en tant qu'opérations de mise à jour du contexte informationnel des participants), à sa généricité (des classes de haut niveau permettent l'annotation de conversations de différentes natures) et à sa multi-dimensionnalité (chaque énoncé peut être annoté avec différentes étiquettes). Cependant, les dimensions employées sont peu discutées et manquent de signification

conceptuelle. Le schéma DIT++ est construit sur les mêmes bases que DAMSL mais cherche à combler ses faiblesses, c'est pourquoi nous l'avons choisi comme base pour notre taxonomie.

3.1 Taxonomie des actes de dialogue adaptés aux CMC

DIT++ propose d'annoter les actes de dialogue en termes de contenu sémantique et de fonction communicative. Chaque énoncé peut contenir plusieurs actes de dialogue. La taxonomie propose dix dimensions définies comme des classes indépendantes de comportements conversationnels (*e.g.* TIME MANAGEMENT, SOCIAL OBLIGATIONS MANAGEMENT). Chaque acte de dialogue appartient à l'une de ces dimensions et est défini plus précisément au travers de sa fonction communicative. Les fonctions communicatives cherchent à capturer l'intention rhétorique de l'énoncé (*e.g.* THANKING, REQUEST FOR INFORMATION). DIT++ propose un large panel de fonctions communicatives, certaines étant générales et pouvant s'appliquer en combinaison avec n'importe quelle dimension, d'autres étant spécifiques à une dimension particulière (*e.g.* GREETING). Par ailleurs, DIT++ propose un nombre important de qualifieurs utilisés pour représenter plus précisément les énoncés en terme de sentiment, de partialité, de certitude ou de conditionnalité.

Le schéma DIT++ a été développé dans le cadre de l'étude de dialogues oraux. Bien que nous acceptions son cadre conceptuel sans réserves, nous avons dû modifier la taxonomie pour l'adapter aux spécificités propres des conversations en ligne et pour en retirer les éléments non applicables aux énoncés écrits.

3.2 Dimensions sémantiques et fonctions communicatives

Cette section détaille les classes employées. Les dimensions retenues sont indiquées en table 1.

Dimension	Description
Domain/Activity	se rapporte à la tâche qui est l'objet de la conversation
Contact Management	établit ou maintient la communication
Communication Management	prépare ou modifie une contribution au dialogue
Discourse Management	structure thématiquement la conversation
Social Obligations Management	participe à la gestion sociale du dialogue
Extra Discourse	actes textuels ne relevant pas du discours
Evaluation	indique qu'un acte précédent a été évalué et exécuté
Attention Perception Interpretation	a trait à la perception, la compréhension ou l'interprétation
Psychological State	informe sur l'état mental et psychologique du locuteur
Time Management	affecte la gestion du temps

TABLE 1 – Dimensions sémantiques retenues

Le schéma d'annotation propose deux types de fonctions communicatives : les fonctions génériques (*general-purpose functions*), qui peuvent se combiner avec n'importe quelle dimension (*e.g. answer*, *requestForInformation*), et les fonctions spécifiques (*dimension-specific functions*), qui ne peuvent être appliquées qu'en conjonction avec une dimension particulière (*e.g. apologizing*, *shift topic*). La table 2 détaille les fonctions que nous avons retenues.

Fonction	Dimension	Exemple
Inform	-	« Le lecteur fonctionne bien »
Request for Action	-	« Redémarre ton pc avant de tester »
Request for Information	-	« Tu utilises quelle version ? »
Request for Directive	-	« Qu'est-ce que je dois faire ? »
Commit	-	« J'essaye ça tout de suite »
Answer	-	« Ça me paraît un peu compliqué »
Answer Positively	-	« Oui ok »
Answer Negatively	-	« Non pas du tout »
Valediction	Social Obligations Management	« Ciao »
Apologizing	Social Obligations Management	« Arf désolé »
Final Self Introduction	Social Obligations Management	« Benou »
Greetings	Social Obligations Management	« Hello ! »
Self Introduction	Social Obligations Management	« Moi c'est Anaïs »
Anticipate Thanking	Social Obligations Management	« Merci d'avance ! »
Thanking	Social Obligations Management	« Merci beaucoup ! »
Summarize	Discourse Management	« En gros il faut changer le pilote »
Reintroduce Topic	Discourse Management	« Et du coup pour le driver »
Report Speech	Discourse Management	« Il a dit "non" »
Conclude	Discourse Management	« Donc forcément ça marchera pas »
Close Topic	Discourse Management	« Ok, ça c'est réglé »
Introduce Topic	Discourse Management	« J'ai aussi un problème avec la carte »
Shift Topic	Discourse Management	« Pour en revenir au driver »
Announce	Discourse Management	« Je vais vous expliquer »
Other Extra	Extra Discourse	« <cite> Shadok a écrit : </cite> »
Boilerplate	Extra Discourse	« −− »
Correct	Communication Management	« Oracle* »
Pause	Time Management	« Hum »
Stall	Time Management	« Euuuh alors... »
Resume	Time Management	« Ok on peut reprendre »

TABLE 2 – Fonctions communicatives retenues

4 Méthode

Nous appréhendons le problème de la modélisation des conversations comme une tâche de classification des énoncés en termes d'actes de dialogue basée sur un apprentissage statistique supervisé.

Cette section détaille l'implémentation de notre approche, les traits choisis pour caractériser les énoncés, notre méthode d'évaluation, et enfin le corpus utilisé.

4.1 Approche et implémentation

Nous avons choisi de développer un système basé sur un classifieur SVM multiclasses. La première raison qui a motivé notre choix, c'est qu'un CRF ou un HMM-SVM ne permettrait pas de comparer nos trois modalités, puisque les conversations chats ne peuvent pas facilement être reconstituées pour permettre leur étiquetage séquentiel, et nous cherchons à effectuer une comparaison homogène sur différentes modalités. De plus, étant donné que les classes de notre taxonomie sont fortement déséquilibrées, nous voulions éviter une approche qui soit dépendante de l'observation *a priori* des annotations. Enfin, l'efficacité des SVMs a été démontrée pour la reconnaissance d'actes de dialogue (Lampert *et al.*, 2006; Qadir & Riloff, 2011). Sauf mention contraire, ce classifieur est utilisé dans la

plupart des expériences rapportées. L'implémentation utilise la librairie *liblinear*[3](Fan *et al.*, 2008).

4.2 Traits

Nous cherchons à tester l'efficacité de traits classiquement utilisés en reconnaissance des actes de dialogue sur un corpus de modalités des CMR. L'efficacité des traits lexicaux tels que les *n*-grammes, et en particulier les unigrammes, a été établie dans le cadre de la tâche de classification des actes de dialogue (Kim *et al.*, 2010; Sun & Morency, 2012; Ferschke *et al.*, 2012; Ravi & Kim, 2007; Carvalho & Cohen, 2005). Nous avons donc choisi de faire des unigrammes et des bigrammes les traits principaux du classifieur. Différentes expériences ont été effectuées avec des combinaisons d'unigrammes, de bigrammes, de trigrammes et de quadrigrammes, mais c'est la combinaison des deux premiers qui a produit les meilleurs résultats. Les autres traits que nous avons testé sont communément utilisés pour des tâches similaires, et incluent les racines des mots, leurs lemmes, leurs étiquettes morpho-syntaxiques, ainsi que des informations contextuelles (*e.g.* la position de l'énoncé dans le message (Ferschke *et al.*, 2012), l'auteur de l'énoncé (Sun & Morency, 2012), la taille de l'énoncé (Ferschke *et al.*, 2012; Lampert *et al.*, 2006)...).

5 Évaluation

5.1 Corpus

À l'instar de Uthus & Aha (2013) et Lowe *et al.* (2015), nous choisissons de construire notre corpus à partir de la communauté Ubuntu. Si ces derniers proposent d'utiliser la plateforme pour construire un large corpus de chats en anglais, nous voyons de nombreux avantages à choisir Ubuntu. Premièrement, la plate-forme est libre, et distribuée sous une licence non restrictive. Deuxièmement, les données disponibles croissent continuellement, et sont donc représentatives de communications modernes à la fois en termes de fond et de forme. Troisièmement, la plate-forme propose à la fois des forums, des canaux IRC et une liste de diffusion, ce qui nous permet d'obtenir des données provenant de différentes modalités mais portant sur les mêmes thématiques. Enfin, il s'agit avant tout d'une plate-forme d'aide à la résolution de problèmes, ce qui correspond parfaitement à notre domaine de recherche.

Nous avons annoté les énoncés du corpus en termes d'actes de dialogue selon notre adaptation de la taxonomie DIT++. Les énoncés sont des groupes de mots, correspondant généralement à des phrases mais parfois à un niveau inférieur, une même phrase pouvant contenir plusieurs énoncés. Conformément à DIT++, chaque énoncé a été annoté avec au plus un acte de dialogue par dimension, chaque énoncé du corpus contenant au moins un acte du dialogue. En règle général, la très grande majorité des énoncés du corpus contiennent un seul acte. La principale annotatrice est une post-doctorante disposant de solides connaissances en linguistique et en TAL. 29 fils de discussions tirés des forums, 45 conversations issues de la liste de diffusion et 6 jours d'activité du canal IRC ont été annotés, ce qui représente plus de 1 200 énoncés pour chaque modalité. Une petite partie de ces conversations - 110 énoncés - a été annotée par un second annotateur expérimenté pour calculer un accord inter-annotateur. Le kappa de Cohen obtenu est de 0,69 pour les dimensions et de 0,70 pour les fonctions communicatives, ce qui correspond à un accord substantiel.

3. *L2-regularized L2-loss support vector classification (dual)*, $\epsilon = 0.1$, $C = 1$.

L'annotation couvre plus de 4 700 actes de dialogue. La table 3 montre la variation entre les comportements conversationnels des participants en fonction de la modalité. Les actes de gestion sociale du dialogue sont nettement plus présents dans les courriels. Dans les chats, la gestion du discours (*e.g.* « voici ma question ») est rare, contrairement aux forums et en particulier aux courriels. En revanche, les classes EVALUATION, ATTENTION PERCEPTION INTERPRETATION et PSYCHOLOGICAL STATE sont bien plus prévalentes, ce qui montre que la synchronisation informationnelle et émotionnelle des participants est plus importante dans les conversations synchrones que dans les conversations asynchrones. Les énoncés qui portent sur la gestion du temps, du contact et de la communication ne sont présents que dans les chats. La table 4 confirme que dans les courriels, les salutations, les adieux, les remerciements et autres actes sociaux semblent attendus et protocolaires pour encadrer un message. Nous observons également que les conversations chats sont plus directes et orientées vers l'action : les participants IRC sont deux fois plus sujets à accomplir un acte commissif et consacrent plus d'énoncés à réclamer des actions, des instructions ou de l'information. Globalement, il semble que les forums représentent une forme intermédiaire de conversation en ligne, à placer entre le style plus formel des courriels et les chats, plus directs et informels par nature.

Dimensions	**Chats**	**Forums**	**Courriels**
Domain/Activity	82,35	80,1	67
Social Obligations Management	9,25	12,85	30,35
Discourse Management	0,85	4,8	2
Evaluation	3,95	1,65	0,15
Psychological State	1,45	0,45	0,35
Attention Perception Interpretation	0,6	0,15	0,05
Communication Management	1,35	0	0
Contact Management	0,9	0	0
Time Management	0,2	0	0

TABLE 3 – Distribution des dimensions des actes de dialogue

Fonctions	**Chats**	**Forums**	**Courriels**
Inform	26,95	31,3	33,35
Answer	17,65	20,05	11,2
Request for Information	16,35	9,2	6,95
Answer Positively	9,1	5,45	3,05
Request for Action	8,85	8,45	6,15
Greetings	5,65	5,1	6,8
Correct	4,75	3	1,3
Answer Negatively	3,4	2,85	1,9
Thanking	2,05	2,85	3,4
Commit	1,95	1,05	1
Request for Directive	0,85	0,9	0,35
Valediction	0,5	1,35	6,15
Apologizing	0,45	0,65	0,6
Anticipate Thanking	0,4	1,95	2,95
Final Self Introduction	0	0,9	9,2
Announce	0,35	0,8	1,25

TABLE 4 – Distribution des fonctions communicatives des actes de dialogue (les classes trop rares pour permettre une comparaison ne sont pas représentées)

6 Expériences et discussion

Dans cette section, nous décrivons les différentes expériences, rapportons leurs résultats et les discutons. Les deux tâches sont : la classification des énoncés en termes de dimensions sémantiques, et en termes de fonctions communicatives. Tous les scores sont calculés par validation croisée ($k = 10$). La partition s'est faite en prenant en compte les conversations, c'est à dire qu'une même conversation ne peut pas se retrouver à la fois dans le corps d'apprentissage et de test. Les métriques utilisées sont les suivantes : l'exactitude (nombre d'énoncés correctement classifiés sur l'ensemble des énoncés), la précision (nombre de vrais positifs sur l'ensemble des vrais et faux positifs) et le rappel (nombre de vrais positifs sur l'ensemble de vrais positifs et faux négatifs). La macro-moyenne (moyenne des moyennes) et la micro-moyenne (moyenne globale) sont rapportées pour la précision et le rappel. Pour évaluer nos résultats, nous avons choisi le maximum de vraisemblance comme approche de base. Nous commençons par rapporter les résultats de notre classifieur SVM sur l'ensemble du corpus multimodal avant de comparer les différentes modalités au travers de plusieurs expériences.

6.1 Comparaison d'un classifieur SVM à l'approche de base

Les premières expériences consistent à entraîner un classifieur SVM multiclasse sur l'ensemble des données (courriels, forums et chats) en utilisant les traits lexicaux. Les résultats de ces expériences sont rapportés en table 5. Pour la classification des dimensions sémantiques, l'exactitude est assez haute, avec 92 % des instances correctement classifiées. Les fonctions communicatives, quant à elles, sont correctement classifiées dans 63 % des cas. Les macro-moyennes, plus faibles, indiquent que le classifieur est moins performant sur certaines classes moins représentées. Le classifieur SVM multi-classes ne peut pas tenir compte de la possibilité qu'un énoncé puisse contenir plusieurs actes de dialogue, cependant dans les faits une infime minorité des énoncés contenant des annotations dans différentes dimensions, cette limitation n'a qu'un impact minime sur les résultats.

Tâche	Exactitude	Macro P.	Macro R.	Micro P.	Micro R.
dimensions : n-grammes	**0,92**	**0,56**	**0,37**	**0,81**	**0,91**
dimensions : max. vrais.	0,60	0,09	0,09	0,55	0,55
fonctions : n-grammes	**0,63**	**0,42**	**0,36**	**0,52**	**0,58**
fonctions : max. vrais.	0,23	0,03	0,03	0,14	0,14

TABLE 5 – Comparaison des performances d'un classifieur SVM entraîné sur les n-grammes et l'approche de base sur l'ensemble du corpus, toutes modalités confondues, pour la tâche de reconnaissance des dimensions sémantiques et des fonctions communicatives

6.2 Comparaison d'approches et de jeux de traits sur différentes modalités

Dans cette section nous présentons les résultats obtenus avec des modèles propres aux différentes modalités (courriels, forums et chats). Les multiples ensembles de traits (racines, lemmes, formes morpho-syntaxiques et traits contextuels) ont été appliqués pour cette tâche. Les tables 8 et 9 rapportent les résultats pour la tâche de classification des dimensions sémantiques et des fonctions

Classe	Courriels			Forums			Chats		
	Nb.	P.	R.	Nb.	P.	R.	Nb.	P.	R.
Domain Activities	894	0,86	0,95	1040	0,70	0,96	1638	0,90	0,99
Social Obligations Management	405	0,70	0,97	161	0,86	0,83	184	0,99	0,74
Discourse Management	27	0,67	0,07	64	0,28	0,17	17	0,00	0,00
Evaluation	2	0,00	0,00	22	0,79	0,50	79	0,65	0,41
Extra Discourse	48	1,00	0,85	2	0,00	0,00	0	-	-
Psychological State	5	0,00	0,00	6	0,00	0,00	29	0,22	0,07
Communication Management	0	-	-	0	-	-	27	0,93	0,48
Contact Management	0	-	-	0	-	-	18	0,50	0,06

TABLE 6 – Dimensions sémantiques : nombre d'exemples dans la référence et résultats d'un classifieur SVM entraîné avec les n-grammes sur les différentes modalités du corpus (les deux classes non représentées sont absentes ou ont une précision et un rappel de 0,00 sur toutes les modalités)

Classe	Courriels			Forums			Chats		
	Nb.	P.	R.	Nb.	P.	R.	Nb.	P.	R.
Inform	461	0,56	0,81	415	0,33	0,66	541	0,46	0,70
Answer	155	0,29	0,22	255	0,30	0,38	354	0,37	0,34
Request for Information	96	0,74	0,62	118	0,67	0,59	328	0,85	0,80
Request for Action	85	0,41	0,27	113	0,42	0,32	178	0,52	0,36
Answer Positively	42	0,47	0,33	73	0,52	0,47	183	0,68	0,61
Greetings	94	0,86	0,95	62	0,82	0,97	113	0,98	0,82
Correct	18	0,00	0,00	34	0,00	0,00	95	0,52	0,18
Final Self Introduction	127	0,45	0,98	10	0,67	0,20	0	-	-
Answer Negatively	26	0,25	0,08	37	0,65	0,35	68	0,20	0,12
Thanking	47	0,64	0,62	38	0,76	0,66	41	0,83	0,73
Valediction	85	0,77	0,87	17	0,69	0,53	10	0,67	0,40
Anticipate Thanking	41	0,67	0,78	24	0,55	0,50	8	0,33	0,12
Commit	14	0,33	0,07	14	0,33	0,07	39	0,48	0,38
Report Speech	4	0,00	0,00	49	0,26	0,24	0	-	-
Boilerplate	48	0,98	0,88	0	-	-	0	-	-
Announce	17	0,50	0,12	11	0,00	0,00	7	0,00	0,00
Request for Directive	5	0,50	0,60	11	0,50	0,09	17	0,42	0,29
Apologizing	8	1,00	0,38	9	0,86	0,67	9	1,00	0,33

TABLE 7 – Fonctions communicatives : nombre d'exemples dans la référence et résultats d'un classifieur SVM entraîné avec les n-grammes sur les différentes modalités du corpus (les deux classes non représentées sont absentes ou ont une précision et un rappel de 0,00 sur toutes les modalités)

communicatives, respectivement. Il apparaît que le jeu de traits n'a que peu d'incidence sur les résultats, et que les approches basées sur les n-grammes sont quasiment toujours optimales, avec une légère amélioration lorsqu'ils sont couplés aux étiquettes morpho-syntaxiques. On peut en conclure qu'un système basé sur la classification multi-classe des énoncés atteint rapidement ses limites : certaines classes sont caractérisées par des mots très discriminants (*e.g.* « OK » pour ANSWER POSITIVELY, « bonjour » pour GREETINGS), d'autres ne le sont pas et nécessitent une approche différente pour être correctement reconnues. La table 6 présente les scores pour les différentes dimensions sémantiques et montre une forte variation entre les classes. On constate qu'elles sont très déséquilibrées : la plus importante, DOMAIN ACTIVITIES, regroupe 75 % des énoncés. La seconde, SOCIAL OBLIGATIONS MANAGEMENT, en regroupe plus de 15 %. Les huit autres se partagent donc seulement 10 % des énoncés restants, certaines ne contenant qu'une poignée d'énoncés. La

table 7 présente les scores des fonctions communicatives. On s'aperçoit que les classes représentant les fonctions communicatives sont un peu mieux équilibrées que celles représentant les dimensions, cependant on note tout de même que plusieurs d'entre elles ne sont pratiquement pas représentées dans le corpus.

Les tables 8 et 9 incluent également les résultats d'une expérience effectuée en remplaçant le classifieur SVM par un CRF. Nous avons utilisé la librairie Mallet (McCallum, 2002) pour notre implémentation. Le corpus de chats a été exclu de l'expérience puisqu'il ne peut pas être analysé séquentiellement avant que les conversations qu'il contient ne soient « démêlées ». Les modèles sont entraînés sur les n-grammes, auxquels sont ajoutés un trait indiquant l'auteur du message. On constate que l'approche basée sur les CRF obtient de très bon résultats sur les forums, où elle bat le classifieur SVM. Elle obtient de moins bons résultats sur les courriels, en particulier dans le cas des fonctions communicatives pour lesquelles les classes sous représentées sont nettement moins bien annotées, comme l'indique les macro-moyennes presque deux fois moins élevées qu'avec un SVM. Ce résultat est contre intuitif, puisque les courriels étant plus formellement construits et tendant plus à respecter un schéma standard, on pourrait s'attendre à ce qu'il soit pertinent de les étiqueter séquentiellement. En réalité, cet apprentissage structurel opéré par le CRF cause beaucoup d'erreurs dès qu'un courriel sort du schéma typique.

Corpus	Traits	Exact.	Macro-P	Macro-R	Micro-P	Micro-R
Courriels	maximum de vraisemblance	0,40	0,13	0,13	0,40	0,40
	n-grammes	**0,94**	0,42	0,40	0,81	0,93
	n-grammes + morpho-syntaxe	**0,94**	0,44	**0,41**	0,81	**0,94**
	lemmes	**0,94**	0,46	**0,41**	0,81	**0,94**
	n-grammes + lemmes	**0,94**	0,46	**0,41**	0,81	0,93
	racines	0,92	0,38	0,36	0,71	0,89
	n-grammes + racines	**0,94**	0,46	**0,41**	0,81	0,93
	n-grammes + contexte	0,93	0,46	0,40	0,79	0,91
	n-grammes (CRF)	0,91	**0,56**	0,37	**0,90**	0,91
Forums	maximum de vraisemblance	0,47	0,12	0,12	0,46	0,46
	n-grammes	**0,92**	0,38	0,36	**0,71**	0,89
	n-grammes + morpho-syntaxe	**0,92**	0,37	0,37	**0,71**	**0,90**
	lemmes	**0,92**	0,37	0,36	**0,71**	**0,90**
	n-grammes + lemmes	0,91	0,37	0,36	**0,71**	0,89
	racines	**0,92**	0,38	0,36	**0,71**	0,89
	n-grammes + racines	0,91	0,37	0,36	**0,71**	0,89
	n-grammes + contexte	0,89	0,29	0,38	0,68	0,86
	n-grammes (CRF)	0,91	**0,69**	**0,46**	**0,71**	0,89
Chats	maximum de vraisemblance	0,68	0,10	0,10	0,68	0,68
	n-grammes	0,89	0,42	0,30	0,89	0,89
	n-grammes + morpho-syntaxe	**0,90**	**0,54**	**0,32**	**0,90**	**0,90**
	lemmes	0,89	0,42	0,29	0,89	0,89
	n-grammes + lemmes	0,89	0,46	0,31	0,89	0,89
	racines	**0,90**	0,47	0,30	**0,90**	**0,90**
	n-grammes + racines	**0,90**	0,48	0,31	**0,90**	**0,90**
	n-grammes + contexte	**0,90**	0,47	0,30	**0,90**	**0,90**

TABLE 8 – Résultats pour la tâche de reconnaissance des dimensions sémantiques sur différentes modalités, différents traits et différentes approches

Corpus	Traits	Exact.	Macro-P	Macro-R	Micro-P	Micro-R
Courriels	maximum de vraisemblance	0,13	0,05	0,05	0,13	0,13
	n-grammes	**0,72**	**0,45**	**0,41**	0,59	0,68
	n-grammes + morpho-syntaxe	0,71	0,40	0,40	0,58	0,66
	lemmes	0,71	0,43	0,40	0,58	0,66
	n-grammes + lemmes	0,71	**0,45**	0,40	0,58	0,67
	racines	0,70	0,43	0,40	0,57	0,66
	n-grammes + racines	0,70	0,43	0,40	0,57	0,65
	n-grammes + contexte	0,70	0,43	0,40	0,57	0,65
	n-grammes (CRF)	0,69	0,23	0,22	**0,68**	**0,69**
Forums	maximum de vraisemblance	0,15	0,04	0,04	0,15	0,15
	n-grammes	0,61	0,40	0,28	0,41	0,51
	n-grammes + morpho-syntaxe	0,60	0,33	0,29	0,39	0,50
	lemmes	0,61	0,37	0,29	0,41	0,51
	n-grammes + lemmes	0,62	0,35	0,29	**0,47**	0,44
	racines	0,61	0,36	0,28	0,40	0,50
	n-grammes + racines	0,61	0,33	0,28	0,40	0,51
	n-grammes + contexte	0,57	0,29	0,28	0,37	0,46
	n-grammes (CRF)	**0,68**	**0,66**	**0,46**	**0,47**	**0,59**
Chats	maximum de vraisemblance	0,15	0,04	0,04	0,15	0,15
	n-grammes	**0,56**	0,35	**0,27**	**0,56**	**0,56**
	n-grammes + morpho-syntaxe	0,54	0,33	0,26	0,54	0,54
	lemmes	0,55	0,35	0,26	0,55	0,55
	n-grammes + lemmes	**0,56**	0,35	**0,27**	**0,56**	**0,56**
	racines	0,55	0,35	**0,27**	0,55	0,55
	n-grammes + racines	**0,56**	**0,36**	**0,27**	**0,56**	**0,56**
	n-grammes + contexte	**0,56**	0,35	**0,27**	**0,56**	**0,56**

TABLE 9 – Expériences pour la tâche de reconnaissance des fonctions communicatives sur différentes modalités, différents traits et différentes approches

7 Conclusion et travaux futurs

Nous avons présenté nos travaux en classification automatique des énoncés en termes d'actes de dialogue, dans un corpus de conversations écrites en ligne à modalités multiples portant sur la résolution collaborative de problèmes. Nous avons rapporté les résultats de nombreuses expériences visant à confronter les approches traditionnelles de classification des actes de dialogue à des données extraites de différentes modalités CMR. Il s'agit à notre connaissance du premier travail qui examine des données tirées de différentes modalités avec la même taxonomie et les mêmes approches. Nous avons rapporté les variations observées entre les modalités, et nous avons montré que des résultats intéressants peuvent être atteints même en se limitant à l'utilisation de traits purement lexicaux. Nous prévoyons de poursuivre ces travaux dans deux directions. D'abord, nous pensons qu'il est important de développer une taxonomie multi-dimensionnelle véritablement propre aux CMR et capable de traiter différentes modalités écrites de manière générique. Puis, nous souhaitons construire un système efficace de classification de ces actes de dialogue.

Remerciements

Ce travail s'inscrit dans le cadre du projet ODISAE (www.odisae.com) et a bénéficié du soutien du fond unique interministériel 17. Nous remercions nos relecteurs pour leurs commentaires constructifs.

Références

ANG J., LIU Y. & SHRIBERG E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *ICASSP (1)*, p. 1061–1064.

AUSTIN J. L. (1975). *How to Do Things With Words*. Oxford University Press, second edition.

BUNT H. (2009). The DIT++ taxonomy for functional dialogue markup. In *Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009)*, p. 13–24, Budapest, Hungary.

CARVALHO V. R. & COHEN W. W. (2005). On the collective classification of email speech acts. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 345–352: ACM.

COHEN W. W., CARVALHO V. R. & MITCHELL T. M. (2004). Learning to Classify Email into "Speech Acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, p. 309–316, Barcelona, Spain.

CORE M. & ALLEN J. (1997). Coding Dialogs with the DAMSL Annotation Scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, p. 28–35, Boston, MA, USA.

FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R. & LIN C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, **9**, 1871–1874.

FERNANDEZ R. & PICARD R. W. (2002). Dialog act classification from prosodic features using support vector machines. In *Speech Prosody 2002, International Conference*.

FERSCHKE O., GUREVYCH I. & CHEBOTAR Y. (2012). Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 777–786: Association for Computational Linguistics.

HA E. Y., MITCHELL C. M., BOYER K. E. & LESTER J. C. (2013). Learning Dialogue Management Models for Task-Oriented Dialogue with Parallel Dialogue and Task Streams. In *Proceedings of the 14th SIGdial Workshop on Discourse and Dialogue (SIGdial 2013)*, p. 204–2013, Metz, France.

IVANOVIC E. (2005). Dialogue act tagging for instant messaging chat sessions. In *Proceedings of the ACL Student Research Workshop (SRW 2005)*, p. 79–84, Ann Arbor, MI, USA.

KIM N. S., CAVEDON L. & BALDWIN T. (2010). Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, p. 862–871, Cambridge, MA, USA.

KIM N. S., CAVEDON L. & BALDWIN T. (2012). Classifying dialogue acts in multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 2012)*, p. 463–472, Bali, Indonesia.

LAMPERT A., DALE R. & PARIS C. (2006). Classifying Speech Acts Using Verbal Response Modes. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW 2006)*, p. 34–41, Sydney, Australia.

LOWE R., POW N., SERBAN I. & PINEAU J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, **abs/1506.08909**.

MCCALLUM A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

POESIO M. & TRAUM D. R. (1997). Conversational Actions and Discourse Situations. *Computational Intelligence*, **13**(3), 309–347.

QADIR A. & RILOFF E. (2011). Classifying Sentences As Speech Acts in Message Board Posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, p. 748–758, Edinburgh, UK.

RAVI S. & KIM J. (2007). Profiling student interactions in threaded discussions with speech act classifiers. *Frontiers in Artificial Intelligence and Applications*, **158**, 357.

SEARLE J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

SEARLE J. R. (1976). *A Taxonomy of Illocutionary Acts*. Linguistic Agency University of Trier.

SUN C. & MORENCY L.-P. (2012). Dialogue act recognition using reweighted speaker adaptation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 118–125: Association for Computational Linguistics.

TAVAFI M., MEHDAD Y., JOTY S., CARENINI G. & NG R. (2013). Dialogue Act Recognition in Synchronous and Asynchronous Conversations. Master's thesis, University of British Columbia.

TRAUM D. R. & HINKELMAN E. A. (1992). Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence*, **8**(3), 575–599.

UTHUS D. C. & AHA D. W. (2013). The ubuntu chat corpus for multiparticipant chat analysis. In *AAAI Spring Symposium: Analyzing Microtext*, volume SS-13-01 of *AAAI Technical Report*: AAAI.