Détection et prédiction de la satisfaction des usagers dans les dialogues Personne-Machine

Narjès Boufaden, Truong Le Hoang, Pierre Dumouchel
École de Technologie Supérieure et Centre de Recherche Informatique de Montréal
{Narjes.Boufaden,LeHoang.Truong,Pierre.Dumouchel}@crim.ca

Résumé. Nous étudions le rôle des entités nommées et marques discursives de rétroaction pour la tâche de classification et prédiction de la satisfaction usager à partir de dialogues. Les expériences menées sur 1027 dialogues Personne-Machine dans le domaine des agences de voyage montrent que les entités nommées et les marques discursives n'améliorent pas de manière significative le taux de classification des dialogues. Par contre, elles permettent une meilleure prédiction de la satisfaction usager à partir des premiers tours de parole usager.

Abstract. We study the usefulness of named entities and acknowldgment words for user satisfaction classification and prediction from Human-Computer dialogs. We show that named entities and acknowledgment words do not enhance baseline classification performance. However, they allow a better prediction of user satisfaction in the beginning of the dialogue.

Mots-clés : prédiction de la satisfaction usager, classification des dialogues Personne-Machine.

Keywords: prediction of user satisfaction, Human-Computer dialog classification.

1 Introduction

La progression des systèmes de dialogue Personne-Machine dans le marché du service à la clientèle crée des attentes grandissantes tant sur le plan de la gestion des données générées par ces systèmes que sur leur exploitation à des fins d'évaluation.

La classification, l'indexation et l'extraction d'information sont autant d'exemples d'applications peu ou pas encore explorées en gestion des dialogues Personne-Machine. La majeure partie de la recherche dans ce domaine est encore consacrée à l'évaluation de ces systèmes.

Dans cet article, nous explorons la classification des dialogues Personne-Machine dans le but de détecter les dialogues problématiques dans lesquels l'usager montre une insatisfaction par rapport au système. Nous explorons l'utilisation des entités nommés et des marques discursives de rétroaction pour la tâche de prédiction de la satisfaction usager durant et après le déroulement du dialogue.

Ces travaux s'insèrent dans le cadre d'un projet en cours avec une compagnie de télécommunication dans le but d'évaluer leur système de dialogue et analyser les dialogues qu'il génère. Comme première étape de ce projet, nous étudions la problématique de détection et prédiction de la satisfaction usager sur un corpus public : DARPA Communicator.

Dans la section 2, nous présentons l'état de l'art en évaluation des systèmes de dialogues Personne-Machine, en détection des dialogues problématiques et présentons le cadre théorique PARADISE. Dans la section 3, nous présentons le corpus DARPA Communicator avec lequel nous effectuons nos expériences. La section 4 décrit notre approche basée du cadre théorique PARADISE. La section 5 présente trois expériences de classification dans lesquelles nous expérimentons différentes combinaisons des attributs proposés pour détecter la satisfaction usager à partir du dialogue. Dans la section 6, nous étudions le rôle des entités nommées et des marques discursives dans la prédiction de la satisfaction usager en début de dialogue. Enfin, la section 7 présente une synthèse de nos résultats ainsi que les prochaines étapes de ce projet.

2 État de l'art

L'évaluation des systèmes de dialogue occupe depuis la fin des années 90 une place de plus en plus importante en recherche. Par exemple, (Eckert *et al.*, 1998) ont utilisé un modèle stochastique pour modéliser le comportement de différentes classes d'usager dans le but de collecter des statistiques sur les différents scénarii de dialogues. En particulier, ils ont montré que la longueur du dialogue en termes du nombre de tours de parole est un bon prédicteur des performances d'un système.

Dans une perspective plus générale, (Walker *et al.*, 1997) proposaient le cadre théorique PARADISE actuellement le plus utilisé pour l'évaluation des systèmes de dialogues. Ce cadre théorique s'inspire de la théorie de décision et pose comme hypothèse que la performance d'un système de dialogues est corrélée avec le degré de convivialité. Dans le cadre de notre application (le service à la clientèle), la convivialité se mesure en termes de satisfaction de l'usager.

PARADISE met en avant la maximisation de la satisfaction usager en maximisant les chances de réussir la tâche ou but du dialogue tout en minimisant les coûts associés à sa réalisation. Ces coûts sont définis en termes d'efficacité (i.e. nombre de tours de parole système et usager) et

qualité du système (i.e. temps de réponse du système). La Figure 1 illustre le cadre théorique PARADISE.

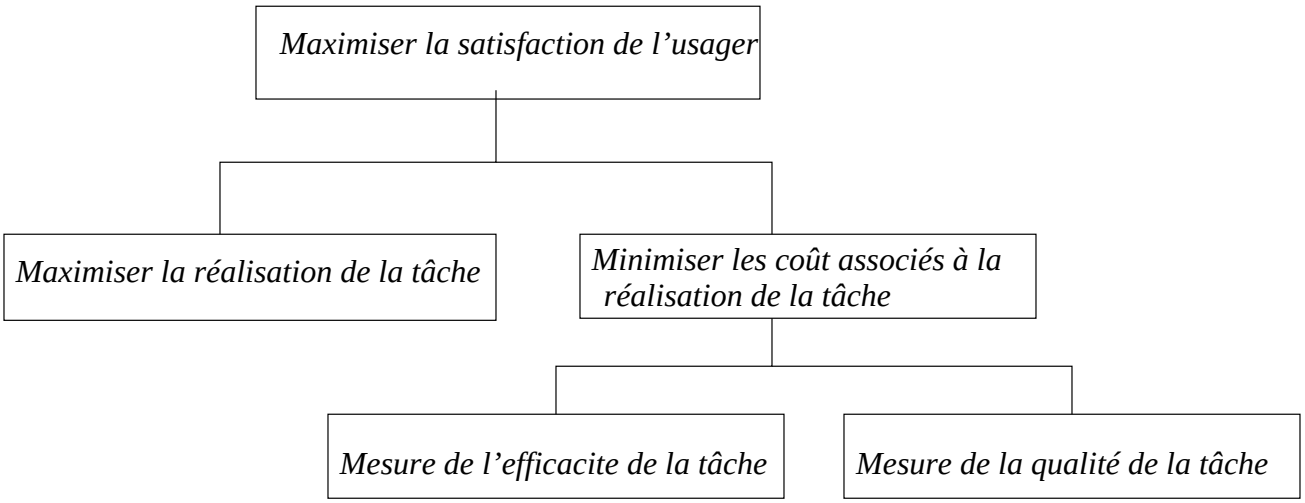

FIG. 1 – Cadre théorique PARADISE

Le cadre PARADISE a été utilisé dans plusieurs travaux (Walker *et al.*, 2000), (Lamel & Rosset, 2000), (Devillers & Rosset, 2000), notamment dans le but de choisir les stratégies de dialogues qui maximisent la satisfaction de l'usager (Walker & Passonneau, 2001), pour déterminer les composantes (reconnaissance de la parole, stratégie de dialogue) ayant le plus d'impact sur la performance d'un système ou encore pour détecter les dialogues problématiques (Hastie *et al.*, 2002).

En particulier, (Hastie *et al.*, 2002) ont proposé une approche pour la détection des dialogues problématiques utilisant 16 attributs représentant différentes mesures reliées aux trois dimensions définies dans PARADISE, à savoir :

Mesure du succès de la tâche Évaluation de l'usager indiquant la réalisation de la tâche.

Mesure de l'efficacité du système Traits extraits des traces du système de dialogue.

- Manuellement : Taux d'erreur de reconnaissance de la parole calculé à partir des transcriptions manuelles des dialogues et le taux d'erreur phrastique.
- Automatiquement : la durée de la tâche, le nombre de tours de parole durant la tâche, nombre de chevauchement de tours de parole système et usager, la moyenne des temps des tours de parole usager, la moyenne des mots par tour de parole usager, la moyenne des temps des tours de parole système, la moyenne des mots par tour de parole système et le type de téléphone (cellulaire ou fixe).

Mesure de la qualité du système Actes de dialogue associés aux tours de parole du système.

Dans une première expérience, les auteurs ont utilisé tous les traits : extraits automatiquement ainsi que ceux annotés manuellement. En utilisant un arbre de décision, ils ont obtenu un taux de classification de 54%. Toutefois, pour éviter l'utilisation des attributs annotés manuellement, ils ont entraîné un arbre de décision afin de prédire le **succès de la tâche** et les actes de dialogues des tours de parole système. Avec un taux de prédiction du **succès de la tâche** de 92% et un taux de prédiction des actes de dialogues de 98%, les auteurs ont obtenu un taux de classification similaire.

(Walker *et al.*, 2002) ont testé la détection des dialogues problématiques en utilisant des attributs entièrement extraits de manière automatique. Ils ont utilisé aussi des mesures de l'efficacité

modélisé avec l'algorithme RIPPER (Cohen, 1995) : un algorithme de classification à base de règles. Sur le corpus généré par le système de AT&T *How May I Help You*, les auteurs ont obtenu un taux de classification de 70,1% à partir du premier tour de parole usager, 78,4% à partir des deux premiers tours de parole et de 83% sur tout le dialogue.

Les travaux que nous présentons se basent sur le cadre théorique PARADISE et s'inspirent des travaux de (Hastie *et al.*, 2002) réalisés sur le corpus DARPA Communicator.

3 Corpus DARPA Communicator

Le corpus DARPA Communicator (version 2001) est un corpus public distribué par le Linguistic Data Consortium (LDC). C'est le résultat d'une expérience menée sur plusieurs sites incluant huit systèmes de dialogue Personne-Machine dans le but de développer des approches robustes de reconnaissance de la parole et de gestion de dialogues pour l'accès interactif à l'information (Robust Recognition and Dialog Tracking for Interactive Information Access). Le domaine d'application choisi est celui des agences de voyages.

Un des défis de cette expérience était de collecter un corpus de dialogues Personne-Machine proche de la réalité. Pour ce faire, les scénarii des dialogues étaient en majorité prédéfinis (sept des neuf scénarii) puisque chaque participant savait l'origine et la destination de son voyage ainsi que les dates et la compagnie aérienne à choisir. Tandis que deux des scénarii étaient laissés au choix des participants. La version 2001 du corpus contient une sélection de plus de 1242 dialogues annotés avec la degré de satisfaction de l'usager. Les dialogues sont en moyenne composés de 51 tours de parole avec en moyenne 25,4 tours de parole usager et une longueur moyenne d'un tour de parole usager est de 64,6 mots.

Dans la version 2001 du corpus que nous utilisons dans nos expériences, chaque dialogue est accompagné de deux fichiers :
– Un fichier qui contient la transcription du dialogue (partie du haut de la Figure 2).
– Un fichier récapitulatif contenant les scores attribués par l'usager pour l'évaluation de la convivialité du système (partie du bas de la Figure 2). En allant de gauche à droite la ligne contient le nom du système (CMU pour Carnegie Melon University), l'index du site (le chiffre 27), le temps de début du dialogue (14 :25), la date du dialogue (2000/07/06), la complétion de la tâche et le reste des données sont respectivement les réponses à des questions portant sur la convivialité du système. Chaque question était notée sur une échelle de 1 à 5 avec la valeur 1 indiquant que l'usager est en parfait accord avec l'affirmation et le score 5 indiquant un total désaccord.

4 Approche

Nous proposons d'utiliser PARADISE pour détecter les dialogues problématiques et prédire la satisfaction usager durant et après le déroulement du dialogue.

Dans notre approche nous tenons compte de deux contraintes liées à notre projet :
– Minimiser la quantité d'information à annoter manuellement. Le but de ces expériences étant de concevoir une application réelle pour la détection des dialogues problématiques, nous voulons obtenir un système entièrement automatisé.

Thu Jul 6 2000 at 15 :15 :38.51 : Task-specific portion started. Thu Jul 6 2000 at 15 :17 :01.44 : Overall task started. Task completion status : not completed. Thu Jul 6 2000 at 15 :15 :24.85 to Thu Jul 6 2000 at 15 :15 :25.01 : New system turn began. Thu Jul 6 2000 at 15 :15 :24.94 : System started speaking. Thu Jul 6 2000 at 15 :15 :36.22 : System finished speaking. System said : . Hello. Welcome to the C M U Communicator. Please speak your 4-digit ID number using the phrase, . My ID number is Thu Jul 6 2000 at 15 :15 :38.51 : New user turn began. Thu Jul 6 2000 at 15 :15 :38.51 : User started speaking. Thu Jul 6 2000 at 15 :15 :43.53 : User finished speaking. Recognizer heard : MY I D NUMBER IS . ?THAT ?. . ?WONDER ?. ZERO EIGHT SEVEN User said : my i. d. number is one zero eight seven [h#] 1087_02_27_03_20000706
CMU 27 14 :25 EDT 2000/07/06 Alive No 4 1 4 3 4 (no comments provided)

FIG. 2 – Exemple de traces du système extrait du corpus DARPA Communicator.

– Minimiser la quantité d'information dépendante du domaine. Notre système étant voué à être appliqué sur un corpus différent du DARPA Communicator, nous voulons minimiser l'information dépendante du domaine des agences de voyages.

Nous proposons de partir des **mesures d'efficacité** proposées par (Hastie *et al.*, 2002). Plus précisément, nous partons des mesures extraites automatiquement et étudions deux nouveaux traits indépendants du domaine : les **entités nommées** contenue dans un dialogue et les **marques discursives de rétroaction**. Nous proposons d'utiliser les **entités nommées** comme indicateur de la densité d'information que nous corrélons avec la qualité du dialogue. Nous pensons qu'une densité d'information faible (interruption du dialogue) ou trop importante (plusieurs répétitions) pourrait indiquer un problème de communication entre l'usager et le système. Les **marques discursives de rétroaction** sont utilisées dans la détection des actes de dialogues notamment d'acquiescement ou de confirmation, tous les deux correlés avec la satisfaction de l'usager (Colineau & Caelen, 1996).

4.1 Les entités nommées

Les entités nommées sont des noms propres ou chiffres référant à des noms de personnes, de lieux ou des noms de compagnies. L'apport des entités nommées pour les applications de compréhension du langage naturel a été largement démontré aussi bien en extraction d'information qu'en résumé automatique ou en traduction automatique. Leur intérêt réside dans la valeur sémantique de l'information qu'ils véhiculent. Dans le cadre de notre application, nous nous intéressons à la correlation existant entre les entités nommées et la densité d'information.

Nous proposons d'utiliser le compte des entités nommées comme indicateur de la densité d'information contenue dans un dialogue. Par ailleurs et dans la mesure où notre corpus n'est pas de taille importante, nous regroupons le compte de toutes les catégories d'entités nommées pour

réduire la dimension des vecteurs des données fournies comme données d'entraînement.

4.2 les marques discursives de rétroaction

Les marques discursives de rétroaction telles que `ok`, `yes` et `no` sont des marques lexicales utilisées pour détecter les actes de dialogues d'acquiescement ou d'opposition (Colineau & Caelen, 1996), (Jurafsky *et al.*, 1998). Ce sont des unités lexicales qui, dans le contexte des dialogues Personne-Machine, donnent une mesure qualitative sur le déroulement du dialogue.

Par ailleurs, les stratégies dialogiques des systèmes de dialogues Personne-Machine étant en grande partie composées de questions à base de choix binaires (Yes/No questions) ou de choix multiples, nous sommes assurés d'observer ces marques de manière significative dans les dialogues. Dans notre approche, nous ne faisons aucune distinction entre les marques de rétroactions indiquant un acquiescement `yes`, `ok`, `correct` ou `yep` de celles indiquant une opposition `no`, `wrong` ou `erase` puisque nous ne tenons pas compte du contexte précédent le tour de parole. Aussi, ce choix nous permet de réduire la dimension des vecteurs des données d'entraînement.

Dans ce qui suit, nous testons différentes combinaisons des attributs décrits dans notre approche pour la tâche de classification des dialogues et la prédiction de la satisfaction usager durant le déroulement du dialogue.

5 Expérience 1 : Détection des dialogues problématiques

Notre première expérience a pour but de classer des dialogues dans une des classes suivantes :
– Positive : L'usager a attribué un score de satisfaction < 12. Ce score est obtenu en cumulant tous les scores des cinq questions évaluant la convivialité du système. Le seuil de 12 est aligné sur celui proposé dans les travaux de (Hastie *et al.*, 2002) afin de permettre la comparaison de nos résultats.
– Négative : L'usager a attribué un score ≥ 12.
Afin d'évaluer notre approche, nous comparons nos résultats avec ceux de (Hastie *et al.*, 2002) en utilisant que les mesures d'efficacité (indépendantes du domaine) extraites de manière automatique et les actes de dialogues (dépendants du domaine) pour classer les dialogues.

Nous présentons quatre expériences combinant nos différents attributs :

Eff. composé uniquement des mesures de l'efficacité extraites automatiquent à partir des traces du système de dialogue.

Eff.+NE composé des mesures de l'efficacité et du compte des entités nommées toutes catégories confondues.

Eff.+ACK composé des mesures de l'efficacité et du compte des marques discursives de rétroaction.

Eff.+EN+ACK composé de tous les attributs.

Nous testons ces combinaisons avec trois algorithmes de classification : Support Vector Machine (SVM), k-Nearest-Neighbour (kNN) et un arbre de décision (DT). Le corpus utilisé est constitué de 1027 dialogues tirés du corpus DARPA Communicator 2001. À la base le corpus contenait une distribution de quatre dialogues de la classe positive pour un dialogue de la

classe négative. Les premiers résultats sur ce corpus avec une distribution très biaisée en faveur des dialogues positifs se rapprochaient sensiblement du baseline de 80%. Afin de remédier à ce biais, nous avons constitué un nouveau corpus à partir du corpus original en dupliquant de manière aléatoire les statistiques de dialogues de la classe négative et en retirant de manière aléatoire les statistiques de dialogues de la classe positive jusqu'à obtention d'une distribution uniforme (re-échantillonage avec duplication) avec 50% des données par classes.

Les résultats que nous présentons sont obtenus sur ce nouveau corpus. Ce sont les moyennes des résultats de 10 validations croisées.

Modèle	Baseline	Eff.	Eff.+NE	Eff.+ACK	Eff.+NE+ACK	(Hastie *et al.*, 2002)
SVM	50%	61,26%	62,9%	62,83%	63,51%	-
kNN	50%	91,41%	91,12%	91,6%	91,8%	-
DT	50%	85,75%	84,68%	87,41%	87,26%	54%

TAB. 1 – Performance des différents classificateurs en termes de taux de classification pour les différentes combinaisons d'attributs.

Contrairement à notre intuition quant à l'apport des entités nommées et des marques discursives, aucune amélioration significative est observée pour les différents modèles. Les résultats pour le modèles kNN est sensiblement le même pour toutes les combinaisons d'attributs. Tandis qu'une petite amélioration est observée pour l'arbre de décision (DT) et le SVM. Cependant, le résultat obtenu pour le kNN avec la combinaison de tous les attributs est supérieur aux résultats obtenus par (Hastie *et al.*, 2002). Toutefois, rappelons que la distribution des classes n'étant pas la même ont ne peut établir une comparaison directe entre nos résultats et les leurs.

Enfin, bien que nous n'ayons pas amélioré le résultat de la classification obtenu avec les mesures d'efficacité, nous avons obtenu un meilleur taux de classification de 91,8% avec le modèle kNN.

6 Expérience 2 : Prédiction de la satisfaction de l'usager

Malgré le peu d'amélioration des résultats obtenus sur la classification des dialogues en combinant toutes les marques, nous voulions tester l'apport des entités nommés et des marques discursives pour la prédiction de la satisfaction usager durant le déroulement du dialogue. Nous avons conduit neuf expériences dans lesquelles nous testons successivement les différentes combinaisons des attributs du Tableau 1 avec des données issues uniquement d'un nombre variable de tours de parole usager. Chacune des combinaisons est testée sur des parties d'un dialogue comprenant respectivement 1 tour de parole de l'usager, 2 tours de parole et ce jusqu'à 8 tours de parole et enfin sur tout le dialogue (soit 25,4 tours de parole).

Nous avons dressé des courbes illustrant la progression du taux de classification durnat le dialogue en augmentant le nombre des tours de parole de l'usager à chaque nouvelle expérience. Nous avons utilisé le même corpus que celui utilisé dans la première expérience et les résultats obtenus sont la moyenne de 10 validations croisées.

Les courbes obtenues sont montrées dans les Figures 3, 4 et 5. La première figure représente l'évolution du taux de classification en fonction du nombre de tours de parole usager et ce pour différentes combinaisons des attributs modélisés par un SVM. La deuxième figure montre la progression modélisée avec l'algorithme kNN et la dernière figure montre la progression modélisée avec un arbre de décision.

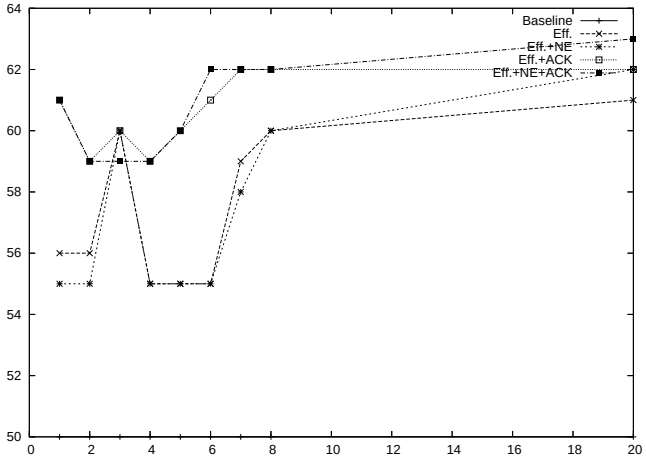

FIG. 3 – Courbes illustrant la progression du taux de classification durant le dialogue pour le modèle SVM.

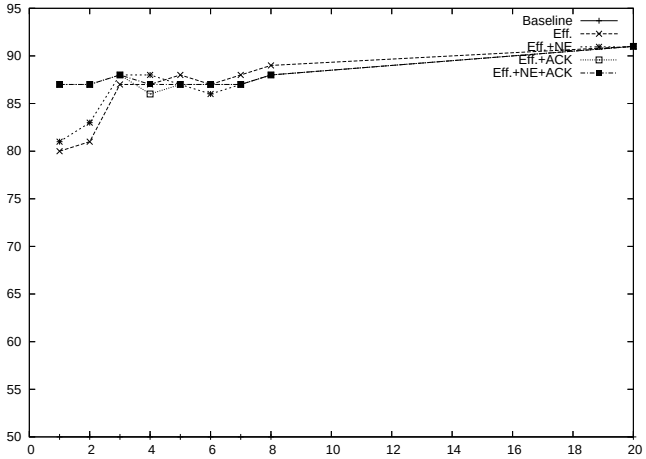

FIG. 4 – Courbes illustrant la progression du taux de classification durant le dialogue pour le modèle kNN.

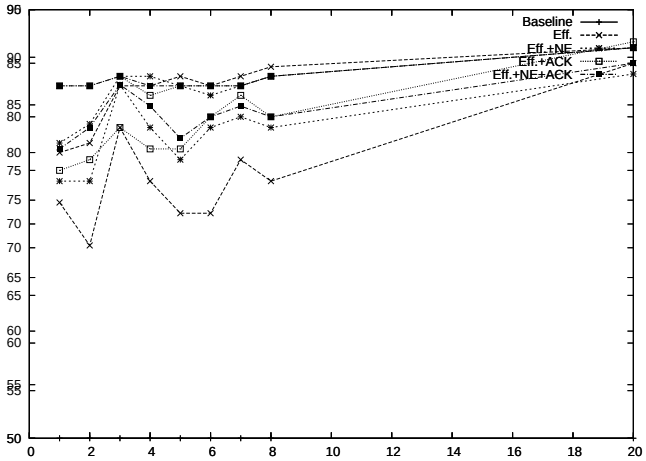

FIG. 5 – Courbes illustrant la progression du taux de classification durant le dialogue pour le modèle DT.

Nous remarquons sur les différents graphiques que pendant les trois premiers tours de parole usager, la combinaison de tous les attributs donne une meilleure prédiction de la satisfaction usager.

Aussi, nous remarquons que le meilleur résultat est en majorité obtenu pour la combinaison Eff.+ACK. Les marques discursives combinées aux mesures d'efficacité améliorent grandement le taux de classification pendant les premiers tours de parole et ce pour tous les modèles. Il semble que l'ajout des entités nommées introduit du bruit car le résultat obtenu pour la combinaison Eff.+EN+ACK est moins bon.

Par ailleurs, la meilleure performance est obtenue avec l'algorithme kNN qui se base sur la distance euclidienne pour classer les dialogues. Cela s'explique par le fait que les données d'entraînement sont des valeurs réelles qui représentent des fréquences et que la distance euclidienne est une fonction qui permet de représenter efficacement la similarité en termes de proximité.

Le Tableau 2 montre les performances du modèle kNN qui a donné le meilleur résultat de classification avec les différentes combinaisons d'attributs.

Nb tours de parole	Eff.	Eff.+EN	Eff.+ACK	Eff.+EN+ACK
1	80,49%	81,78%	87,07%	87,26%
2	81,47%	83,44%	87,65%	87,26%
3	87,80%	88,29%	88,48%	88,39%
tous	91,41%	91,12%	91,6%	91,8%

TAB. 2 – Progression du taux de classification sur les trois premiers tours de parole usager pour le modèle kNN.

7 Conclusion

Nous avons testé la combinaison des mesures de l'efficacité proposées par (Hastie *et al.*, 2002) avec deux nouveaux attributs : les entités nommés et les marques discursives de rétroaction pour la classification des dialogues et la prédiction de la satisfaction usager. Bien que ces deux attributs n'aient pas amélioré de manière significative les résultats de la classification des dialogues, ils ont permis une meilleure prédiction de la satisfaction usager durant les premiers tours de parole de l'usager.

En particulier, le compte des marques discursives a permis d'obtenir les meilleurs taux de prédiction en début de dialogue. Ce résultat a un intérêt particulier puisque moyennant cet attribut facilement calculable et non dépendant du domaine d'application, nous avons amélioré le taux de prédiction de la satisfaction usager de 6,6% pour le premier et second tour de parole par rapport à celui obtenu avec le modèle utilisant un arbre de décision.

Dans les prochaines étapes, nous exploiterons d'avantage les marques discursives de rétroaction en tenant compte de l'information prosodique extraites de l'audio de ces unités lexicales. L'ajout de la prosodie permettra de distinguer ces marques en leur associant un contenu émotionnel pour une meilleur prédiction de la satisfaction usager.

Remerciement

Ce projet est rendu possible grâce au support de Patrimoine Canada.

Références

COHEN W. (1995). Fast Effective rule induction. In *Proceedings of the 12th Conference on Machine Learning*.

COLINEAU N. & CAELEN J. (1996). Une approche lexicale pour la reconnaissance d'actes de dialogue. In *Séminaire lexique en traitement automatique de la parole*, p. 137–145, Toulouse, France.

DEVILLERS B.-M. H. L. & ROSSET S. (2000). Predictive performance of dialog systems. In *Int. Conf. on Language ResourcesandEvaluation,LREC2000*.

ECKERT W., LEVIN E. & PIERACCINI R. (1998). Automatic evaluation of spoken dialogue systems. *TWLT13 : Formal semantics and pragmatics of dialogue*.

HASTIE H., PRASAD R. & WALKER M. (2002). What's the Trouble : Automatically Identifying Problematic Dialogues in DARPA Communicator Dialogue Systems. In *Proceedings of the ACL 2002*.

JURAFSKY D., E. S., B. F. & CURL T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING 98 Workshop on Discourse Relations and Discourse Markers*.

LAMEL L. & ROSSET S. (2000). Considerations in the design and evaluation of spoken language dialog systems. In *ICSLP, 2000*.

WALKER M., LITMAN D. J., KAMM C. A. & ABELLA A. (1997). *Interactive Spoken Dialog Systems : Bridging Speech and NLP Together in Real Applications*, chapter Evaluating Interactive Dialogue Systems : Extending Component Evaluation to Integrated System Evaluation. Association for Computational Linguistics : New Brunswick, New Jersey.

WALKER M. A., LANGKILDE I., WRIGHT J., GORIN A. & LITMAN D. (2000). Learning to predict problematic situations in a spoken dialogue system : Experiments with how may i help you ? In *North American Meeting of the Association of Computational Linguistics*.

WALKER M. A., LANGKILDE-GEARY I., HASTIE H. W., WRIGHT J. & GORIN A. (2002). Automatically Training A Problematic Dialog Predictor for the HMOHY Spoken Dialog System. *Journal of Artificial Intelligence Research*, **16**, 293–319.

WALKER M. A. & PASSONNEAU R. (2001). Date : A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *In Proceedings of Human Language Technology Conference*, San Diego.