Prototypage rapide et évaluation de modèles de dialogue finalisés

Martin Rajman, Andréa Rajman, Florian Seydoux, Alex Trutnev

Laboratoire d'Intelligence Artificielle, E.P.F. Lausanne (Suisse)
{martin.rajman, florian.seydoux, alex.trutnev}@epfl.ch

Résumé

L'objectif de cette contribution est de présenter l'intégration de la notion d'évaluation dans la méthodologie de prototypage rapide de modèles de dialogue développée et mise en œuvre dans le cadre du projet *InfoVox*. L'idée centrale de cette méthodologie est de dériver un modèle de dialogue opérationnel directement à partir du modèle de la tâche à laquelle il est associé. L'intégration systématique de différents aspects de l'évaluation dans le processus de prototypage est alors utile afin d'identifier, dès la phase de conception, les qualités et défauts de l'interface. Toutes les conclusions présentées seront illustrées par des résultats concrets obtenus au cours d'expériences réalisées dans le cadre du projet *InfoVox*.

Mots Clés

Evaluation, Dialogue Homme-Machine, Prototypage Rapide, Wizard-of-Oz

1 Introduction

Le projet *InfoVox* (Van Kommer et al., 2000) est une réalisation conjointe de l'EPFL, de l'IDIAP, et des sociétés Swisscom et Omedia. L'objectif de ce projet était le développement d'un prototype de serveur vocal permettant un accès par téléphone à des informations sur les restaurants de la ville de Martigny, Suisse (par exemple, (Jurafsky, 1994)). Le but principal du projet était le développement du prototype et son évaluation au cours de 2 *field-tests*.
Dans la section 2 sont décrites les principales étapes de la méthodologie de prototypage rapide proposée. La section 3 concerne l'évaluation, et plus précisément la description de l'intégration de ses différents aspects dans les étapes de conception correspondant aux *field-tests* internes et externes, ainsi que les différents types d'analyses employés pour évaluer l' «utilisabilité» du système.

2 Prototypage rapide de modèles de dialogue finalisés

L'un des buts principaux du projet *InfoVox* était la spécification et la validation d'une méthodologie de prototypage de modèles de dialogue (MD) pouvant être déployée dans un contexte applicatif pour la conception rapide de MD à états finis finalisés (Cole et al., 1993). Par «conception rapide», on entend qu'une première version utilisable de MD devrait pouvoir être produite en quelques heures ; par «modèles de dialogue finalisés», nous entendons des modèles ciblés, strictement spécifiques à une tâche donnée. L'idée sous-jacente à la

méthodologie proposée est que le MD visé est un modèle à états finis pouvant être dérivé de manière relativement systématique à partir de la modélisation de la tâche à effectuer.

2.1 Production du modèle de tâche

L'approche proposée consiste à modéliser la tâche sous la forme d'une table relationnelle (formulaire) dont les champs représentent les différents attributs devant être informés pour que la tâche puisse être accomplie (Denecke, 1997). En d'autres termes, la tâche est modélisée comme une fonction dont les arguments correspondent aux attributs et dont l'appel résulte en l'accomplissement de la tâche. Par exemple, dans le projet *InfoVox*, le modèle de tâche (MT) est réduit à la fonction *choisirRestaurant(typeCuisine, localisation, ...),* dont les attributs identifient les différents critères de sélection disponibles pour la recherche de restaurant.
La méthodologie actuelle identifie deux classes de tâches élémentaires pouvant être traitées : les tâches de *recherche d'information* et les tâches de *commande*. Dans le premier cas, l'objectif principal est de sélectionner les éléments correspondant à des critères de recherche, sur la base d'attributs prédéfinis pour un ensemble de candidats stockés dans une base de données. Pour les tâches de commande, l'objectif est le contrôle d'une fonction spécifique.
Il est évident que ces deux types de tâches élémentaires ne permettent pas, à eux seuls, d'apporter une puissance descriptive suffisante pour modéliser la plupart des applications réelles. Deux mécanismes de composition permettant de construire des modèles plus complexes sont proposés : la *disjonction* et l'*imbrication*. La *disjonction* correspond simplement au regroupement (disjonctif) de plusieurs tâches élémentaires dans un MT unique. *L'imbrication* de tâches survient lorsque les valeurs de certains attributs, identifiés dans le modèle produit pour une tâche donnée, correspondent eux-mêmes au résultat de l'appel à une fonction nécessitant un dialogue.
Une fois le MT produit, il servira de base pour la génération automatique d'un MD initial, correspondant à une interface vocale permettant à l'utilisateur de spécifier les valeurs nécessaires pour accomplir la tâche.

2.2 Dérivation du modèle initial de dialogue

Le MD est défini comme un ensemble de nœuds de dialogue génériques (GDN) interconnectés (Bilange, 1992), chacun de ces nœuds étant associé à l'un des attributs présents dans le MT. Pour un champ donné, le rôle du GDN associé est d'exécuter une interaction simple avec l'utilisateur, avec comme but l'acquisition d'une valeur acceptable pour l'attribut en question. D'une façon générale, un GDN est caractérisé (1) par un message/question à poser à l'utilisateur pour obtenir une valeur pour l'attribut associé ; (2) une grammaire permettant de passer des expressions utilisées par les utilisateurs aux valeurs canoniques utilisées dans le système ; et (3) une «action» conditionnant la gestion du flot du dialogue. Par ailleurs, dans chaque GDN, sont incorporés des mécanismes génériques permettant de gérer des situations problématiques pouvant survenir au cours du dialogue.
La gestion des GDNs se fait dans un module spécifique, l' «interpréteur de dialogues». D'autres éléments sont cependant nécessaires : en plus de la définition des GDNs et la spécification de l'interpréteur de dialogue, une logique de branchement, responsable du contrôle du dialogue, doit être définie. Dans notre approche, cette logique de branchement a tout d'abord été implémentée sous la forme d'un ensemble de conditions de transitions entre les GDNs, et est actuellement reformulée sous la forme d'un mécanisme générique opérant sur l'ensemble des configurations possibles (e.g., les solutions candidates dans le cas d'une

recherche d'information ou les actions possibles dans le cas d'un appel de commande) correspondant aux valeurs courantes des attributs du modèle.
Dans le cadre de l'architecture proposée, le MT est utilisé pour générer un MD d'initialisation, c'est-à-dire un MD dans lequel les GDNs sont instanciés avec des énoncés simples, générés automatiquement, sans grammaire d'interprétation, et sans actions spécifiques associées.
Le modèle minimaliste ainsi produit n'est donc pas directement utilisable. Son «instanciation» requiert une intervention humaine, qui peut être réalisée dans le cadre d'une expérience de type *Wizard-of-Oz*, telle que décrite dans la section suivante.

2.3 Instanciation du modèle initial : expérience *Wizard-of-Oz*

Dans le cadre spécifique du projet *InfoVox*, les objectifs de l'expérience *Wizard-of-Oz* (WoZ) étaient : (1) de permettre une première évaluation subjective du MD initial ; (2) d'acquérir des données utiles pour l'entraînement du reconnaisseur de parole (*STRUT*) utilisé dans le projet.
L'expérience WoZ du projet *InfoVox* a duré 38 jours ; 99 personnes y ont participé, permettant d'enregistrer 255 dialogues. L'expérience a mené à une première version totalement opérationnelle du MD, qui a ensuite subi les évaluations décrites dans les sections suivantes.

3 Evaluation du système : *field-tests* interne et externe

L'un des objectifs importants du projet *InfoVox* était de proposer une méthodologie d'évaluation qui permette d'identifier et de mettre en évidence, dès l'étape de prototypage, les qualités et les défauts du système, ceci du point de vue des utilisateurs (Walker et al., 2000) .
Concrètement, il a été décidé d'effectuer l'évaluation sous la forme d'un questionnaire de satisfaction (QS), soumis à chacun des membres d'un groupe d'utilisateurs potentiels, après son interaction avec le prototype. Plus précisément, pour chacun des utilisateurs contactés, la procédure d'évaluation a été décomposée en les 4 étapes suivantes : (1) une courte description du projet et de la procédure d'évaluation est lue à l'utilisateur ; (2) l'utilisateur est mis dans un contexte d'application précis ; dans ce but, un scénario est choisi parmi quelque 20 scénarios, définis de façon à couvrir différentes combinaisons des attributs décrivant les restaurants, et lu à l'utilisateur ; (3) l'utilisateur est mis en communication avec le système par le biais d'une téléconférence ; et (4), une fois l'interaction entre l'utilisateur et le système terminée, l'utilisateur est prié de répondre à un QS. Ce QS contenait quelque 40 questions segmentées en 4 catégories : (1) les questions directement liées à la satisfaction globale des utilisateurs ; (2) les questions relatives à l' «utilisabilité» du système, i.e. le confort et la convivialité de l'interaction ; (3) les questions relatives à l'efficacité du système, i.e. la qualité et la pertinence des résultats produits ; et (4) les questions relatives à l'évaluation comparative du système.

3.1 *Field-test* interne

Le field-test interne a été effectué sur 2 jours et a impliqué un nombre limité de 20 utilisateurs «internes», typiquement les concepteurs du système, d'autres membres des laboratoires impliqués, etc. La plupart des utilisateurs connaissaient donc le prototype. En plus des réponses des utilisateurs au QS (données subjectives), le déroulement du dialogue ainsi que l'état interne du système ont été enregistrés tout au long de l'interaction, dans le but de produire des indicateurs objectifs. Enfin, les interactions de l'utilisateur avec le système ont été également enregistrées et manuellement retranscrites.
Concernant les résultats, il est important de noter que l'objectif principal du *field-test* interne était essentiellement la validation du prototype et de la procédure d'évaluation, plutôt que la

production de véritables résultats d'évaluation. De ce fait, le résultat du *field-test* interne a été la production d'un MD amélioré et la définition rigoureuse de la procédure d'évaluation. Ces deux éléments ont ensuite été figés et utilisés sans modifications dans le *field-test* externe.

3.2 *Field-test* externe

L'objectif principal du *field-test* externe était d'effecteur l'évaluation de la version du système résultant du *field-test* interne, au moyen de la procédure d'évaluation exposée précédemment. Dans ce but, une population d'utilisateurs «externes», i.e. n'ayant aucune connaissance du système, a été sélectionnée de façon aléatoire, sur la base de l'annuaire téléphonique électronique TelInfo de *SwissCom*. Le *field-test* a duré 6 semaines, impliquant 50 utilisateurs (seuls les particuliers possédant le téléphone fixe ont été contactés).
Les données subjectives et objectives brutes produites pendant le *field-test* externe ont servi de base pour l'évaluation et l'analyse du système. Comme indiqué précédemment, les données brutes disponibles pour l'analyse à la fin du *field-test* consistaient en : (1) les transcriptions des interactions entre les utilisateurs et le prototype ; (2) les *logs* produits automatiquement pendant les interactions ; et (3) les réponses aux questions soumises aux utilisateurs.

3.3 Exploitation des données du *field-test* externe

Plusieurs indicateurs subjectifs et objectifs ont été dérivés à partir des données brutes. Les indicateurs subjectifs correspondent essentiellement aux scores moyens obtenus pour les diverses questions fermées présentes dans le QS, tandis que les indicateurs objectifs ont été dérivés à partir des *logs* et correspondaient donc aux valeurs moyennes de différentes caractéristiques du système décrivant l'interaction avec chacun des utilisateurs.
En ce qui concerne l'exploitation des indicateurs produits, 3 types d'analyses ont été effectués : (1) analyse rétrospective des tendances prédominantes correspondant à l'identification, pour les questions fermées, des modalités significativement prédominantes ; les modalités prédominantes ainsi identifiées peuvent ensuite être utilisées pour produire, de façon rétrospective, une vue synthétique des opinions des utilisateurs sur leur interaction avec le système ; (2) analyse rétrospective des corrélations correspondant à l'identification de corrélations significatives entre les réponses à des paires de questions fermées ; (3) analyse prospective des corrélations correspondant à l'identification de corrélations significatives entre les réponses des utilisateurs à certaines questions fermées et certains indicateurs objectifs ; ce type d'analyse permet d'identifier les dépendances entre les caractéristiques objectives du système et les perceptions subjectives qu'en ont les utilisateurs. Les dépendances ainsi identifiées peuvent être utilisées pour aider à l'identification de modifications «prometteuses» du prototype pouvant conduire à une meilleure satisfaction des utilisateurs.

3.3.1 Indicateurs subjectifs et analyse rétrospective des tendances prédominantes

Étant dérivés des questions fermées, les indicateurs sont organisés de façon naturelle selon la même structure que celle définie pour le questionnaire.
Ainsi, parmi les indicateurs subjectifs produits, certains concernaient l'évaluation de la satisfaction globale des utilisateurs, soit directement quantifiée par les utilisateurs eux-mêmes, soit indirectement par le biais de réponses à des questions comme «*Si le système était disponible, seriez-vous prêt à l'utiliser ?*». D'autres indicateurs concernaient l'«utilisabilité» et portaient donc sur la caractérisation de la qualité de différentes caractéristiques du système (questions fermées comme «*Diriez-vous que le système est difficile/facile à utiliser ?*») ou de l'interaction elle-même. L'efficacité (i.e. la qualité/cohérence des résultats produits) était

couverte par des indicateurs (scores moyens) dérivés de questions oui/non comme «*Est-ce que le système vous a renseigné correctement ?*». Enfin, les indicateurs relatifs à la comparaison du système avec des sources d'information alternatives ont également été dérivés.
La méthode adoptée dans le projet *InfoVox* pour l'analyse rétrospective des tendances était la suivante : pour toutes les questions fermées, on identifiait les modalités pour lesquelles la borne inférieure de l'intervalle de confiance à 95% de certitude était supérieure à la borne supérieure du même intervalle de confiance pour toutes les autres modalités. Les modalités prédominantes identifiées étaient ensuite fournies aux concepteurs pour interprétation.
Concrètement, l'exploitation des résultats obtenus lors du *field-test* externe a produit l'évaluation synthétique suivante : le taux global de satisfaction était de 63.75% ; 85.4% des utilisateurs étaient prêts à utiliser le système s'il était disponible et 76% serait même prêt à le conseiller à ses proches, amis ou collègues. Le système s'est avéré facile à utiliser (89.8%) et assez bien adapté pour une recherche de restaurants (79.2%). Cependant, ses capacités d'interaction restent essentiellement comparables à celles d'une machine (74.4%). Le système fournissait une information plutôt correcte (63.3%), mais les utilisateurs l'auraient voulue plus riche. 84.0% des utilisateurs considéraient le système au moins aussi efficace que les divers systèmes à base de touches de fréquences qu'ils connaissaient. Les messages du système étaient faciles à comprendre (94.4%), la durée des interactions était adéquate (72.9%) et le système a rarement (30%) fait quelque chose considéré par les utilisateurs comme étrange ou inutile. Cependant, la stratégie adoptée dans le modèle de dialogue aboutissait souvent à la répétition des messages (82%), même si la séquence des questions posées était considérée comme naturelle (93.9%). Les utilisateurs étaient le plus souvent détendus pendant l'interaction (84.8%), ils étaient rarement perdus (14%), mais assez souvent (48%) ils ont dû s'adapter au système et souvent (74%) ils avaient l'impression que le système ne comprenait pas certaines de leurs réponses. 79.2% des utilisateurs étaient sensibles aux messages de confirmation et considéraient ces confirmations très utiles (96.8%). Enfin, les opinions sont restées très partagées en ce qui concerne l'initiative durant le dialogue (l'utilisateur ou le système).

3.3.2 Analyse rétrospective des corrélations

Plusieurs résultats intéressants ont été produits par l'analyse rétrospective des corrélations. Tout d'abord, l'analyse a montré que le seul identificateur subjectif corrélé avec la satisfaction des utilisateurs était celui relatif à la qualité des résultats produits par le système ; plus précisément, les utilisateurs ayant considéré que le système ne fournissait pas des résultats corrects avaient une tendance significative à considérer le système comme non satisfaisant.
Le deuxième résultat intéressant a été que, contrairement à ce qui a été postulé dans la section précédente, il n'y a pas eu de corrélation significative entre la satisfaction des utilisateurs et les indicateurs subjectifs relatifs au fait que les utilisateurs étaient ou non prêts à utiliser le système ou à le recommander à des tiers. En d'autres termes, cela signifie que, dans le cas du projet *InfoVox*, le fait que l'utilisateur soit prêt à utiliser le système ne garantie pas une bonne satisfaction de sa part. Ce résultat ne nous semblait pas du tout évident à prédire !

3.3.3 Analyse prospective des corrélations

L'objectif de cette analyse était d'identifier les dépendances significatives entre les indicateurs objectifs et certains indicateurs subjectifs. L'avantage principal de l'identification de telles dépendances est qu'elles permettent de choisir de façon plus précise les caractéristiques du système qui doivent être améliorées en priorité afin d'augmenter les chances d'atteindre une meilleure satisfaction des utilisateurs.

Par exemple, si l'objectif est effectivement d'améliorer la satisfaction des utilisateurs, les conclusions sont d'essayer d'améliorer la perception par les utilisateurs de la qualité des résultats produits par le système. En conséquence, il peut être intéressant de savoir quels sont les indicateurs objectifs en corrélation avec ce type de perception.
Pour quantifier la force de la dépendance entre un indicateur subjectif et un indicateur objectif, la technique statistique utilisée consistait à tester la différence entre les moyennes du critère objectif obtenues pour chacune des modalités du critère subjectif considéré.
Appliquée à l'indicateur subjectif «qualité des résultats», cette analyse a fait ressortir une corrélation entre les opinions positives pour cet indicateur et la présence d'une fraction plus élevée (35% v.s. 14% dans le cas d'une perception négative de la qualité des résultats) d'énoncés du système durant la phase du dialogue où le système cherche à acquérir l'information nécessaire pour fournir une suggestion de restaurant (la phase d'acquisition des contraintes). Ce résultat peut être interprété de la manière suivante : du fait que les messages du système qui ne sont pas de simples questions correspondent à des réactions à des situations problématiques lors du dialogue (i.e. des réponses à des demandes d'aide, etc.), un ratio élevé de «simples questions» pendant la phase d'acquisition de contraintes signifie que l'interaction se passe bien.

4 Conclusions

Dans la première partie de cette contribution, nous avons présenté une méthodologie de prototypage rapide de modèles de dialogue permettant la production efficace de modèles de dialogues à états finis, simples et finalisés. La seconde partie de la contribution a été consacrée à la description des mécanismes d'instanciation et d'affinement des modèles de dialogue initiaux, en particulier par le biais d'expérience WoZ, ainsi qu'à la description de la procédure d'évaluation associée retenue. La méthodologie a été illustrée dans le cadre concret de la production du modèle de dialogue du serveur vocal d'information réalisé dans cadre du projet *InfoVox* et proposant des informations sur les restaurants de la ville de Martigny Suisse.

Références

Bilange E. (1992), Dialogue personne-machine, modélisation et réalisation informatique, Langue, Raisonnement, Calcul, Hermès, Paris.

Cole R., Novick D. G., Fanty M., Sutton S., Hansen B., Burnett D. (1993), Rapid prototyping of spoken language systems: The year 2000 census project, *Proceedings of the International Symposium on Spoken Dialogue*, Tokyo, Japan.

Denecke, M. (1997), An Information-based Approach for Guiding Multi-Modal Human-Computer-Interaction, in *Proc. of IJCAI (2),* 1036-1041.

Jurafsky D. (1994), The Berkeley Restaurant Project, in *Intl. Conference on Spoken Language Processing*.

Van Kommer R., Rajman M., Bourlard H., (2000), Heading Towards Virtual-Commerce Portals, C*omtec*, 9:10-13.

Walker M. A., Kamm C., Litman D., Towards Developing General Models of Usability with PARADISE, *Natural Language Engineering Special Issue on Best Practice in Spoken Dialogue Systems*.